CHAPTER EIGHT

Severe Tests and Novel Evidence

> I think that people emphasize prediction in validating scientific theories because the classic attitude of commentators on science is not to trust the theorist. The fear is that the theorist adjusts his or her theory to fit whatever experimental facts are already known, so that for the theory to fit these facts is not a reliable test of the theory.
>
> In the case of a true prediction, like Einstein's prediction of the bending of light by the sun, it is true that the theorist does not know the experimental result when she develops the theory, but on the other hand the experimentalist does know about the theoretical result when he does the experiment.
>
> —Steven Weinberg, *Dreams of a Final Theory*, pp. 96–97

WITHIN TWENTY-FOUR HOURS of the bomb explosion at the World Trade Center in New York City in February 1993 there were nineteen telephone calls from individuals or organizations claiming responsibility. The fact that the calls came after the first news reports of the explosion, it was generally agreed, greatly weakened the credibility of the claims. Our intuition here reflects the common principle that evidence predicted by a hypothesis counts more in its support than evidence that accords with a hypothesis constructed after the fact. Many may say that merely explaining known evidence provides little or no support for a hypothesis altogether. Nevertheless, it is equally clear that less than one week after the bombing the deliberate use of known pieces of evidence, in particular, "three feet of mangled, soot-encrusted steel" (*Newsweek*, 15 March 1993, 28), warranted the investigators to finger the man who rented the van that carried the explosive device.[1]

[1] Under the grime of this piece of the truck chassis was an identification number. This led authorities to New Jersey and to the Islamic fundamentalist who rented the van.

The seeming conflict in these two intuitions is at the heart of a long-standing dispute in philosophy of science. The dispute may be seen to center on the following methodological principle or rule:

> *Rule of novelty* (RN): for evidence to warrant a hypothesis *H, H* should not only agree with the evidence, but the evidence should be *novel* (in some sense).

On some accounts the novelty of the evidence is required, whereas for others the claim is comparative: novel evidence accords the hypothesis greater weight than if it were nonnovel. Laudan refers to this rule as "one of the most contested principles in recent philosophy of science" (Laudan 1990b, 57). In fact, the controversy over some version of RN has a very long history, marked by vehement debates between several eminent philosophers of science: between Mill and Peirce, Mill and Whewell, and Keynes and Popper.[2] The current dispute about the relevance of novelty is commingled with a family quarrel among those who endorse RN over the very definition of "novel test" or "novel fact." The disputants here tend to take an all or nothing approach. Proponents of novelty find the novelty of the evidence always relevant in assessing its bearing on hypotheses, holding that evidence is accorded extra weight (of some sort) simply by dint of being novel. Opponents deny that novelty ever matters. There are further disagreements even among proponents who share the basic definition of novelty over what counts as violating novelty and, most importantly, over *why* novelty should matter to the import of evidence.[3]

I will not attempt to survey the literature to which this quarrel about novelty has continued to give rise but will argue for a change of focus in the entire battle—on both of the fronts on which it has been fought. Novelty, I claim, was not the real issue in the first place. What lay behind the intuition that novelty mattered is that severe tests matter. What underlies the basic intuition that if the data are not novel, then they fail to test or support a hypothesis are the various impediments to severity that correlate with violating novelty of one sort or

[2] Early proponents often cited are Descartes and Leibnitz. An indication in Descartes is his claim in *Principles of Philosophy* that "we shall know that we have correctly determined these causes when we observe that we can explain, by their means, not only those phenomena which we have considered up to now, but also everything else about which we have not previously thought" (Descartes [1644] 1984, pt. 3, p. 104).

[3] For some associated readings on the novelty debate beyond those cited in this chapter, see Campbell and Vinci 1983; Howson 1984; Gardner 1982; Musgrave 1978, 1989; and Redhead 1986.

another. But this correlation is imperfect. Novelty and severity do not always go hand in hand: there are novel tests that are not severe and severe tests that are not novel.

As such, criteria for good tests that are couched in terms of novelty wind up being either too weak or too strong, countenancing poor tests and condemning excellent ones. I believe that our notion of severe tests captures pronovelty intuitions *just where those intuitions are correct.*

Understanding and resolving the dispute about RN is important for several reasons. For one thing, an important task for an adequate epistemology of experiment is to unearth the principles underlying familiar methodological rules. The controversy in this case has had particularly serious consequences. Finding fault with those who argue that novelty in some sense always matters has been taken by some as supporting the opposite view, that it never matters. This discounts the important kernel of rightness underlying those who think novelty matters: aspects of the hypotheses and data generation procedures need to be taken into account in assessing the goodness of tests. They may be relevant to the error probabilities and so to the severity of the overall experimental test.

In addition, the controversy has had disturbing consequences for historically minded philosophers of science. Discovering that several scientific episodes appear to fail so apparently plausible a rule has been taken as grounds for questioning the rationality and objectivity of the episode or, alternatively,. as grounds for questioning the viability of the methodological enterprise.

There is a further reason that resolving the novelty controversy is of particular importance for our program. Any philosophy of experimental testing adequate to real experiments must come to grips with the fact that the relationship between theory and experiment is not direct but is mediated along the lines of the hierarchy of models and theories as sketched in chapter 5. At various stages of filling in the links, it is standard to utilize the same data to arrive at as well as warrant hypotheses. It is commonly the case, for example, that raw data are used to construct as well as to test a hypothesis about experimental data, and experimental data are used to construct as well as to support an experimental hypothesis—the basis for a comparison with a theoretical prediction. As a matter of course, then, the inferences involved violate even the best construals of the novelty requirement. This is commonly so for the central task of experimental inference—estimating the effects of backgrounds. Indeed, if one goes down the list in chapter 2 of the standard problems of "normal" science, one finds again and again that they are tasks where hypothesized solutions are routinely affirmed by nonnovel evidence. So RN must be abandoned or qualified.

A realistic picture of the relationship between evidence and hypotheses reveals not only that nonnovel results often figure in altogether reliable inferences, but also that there is as much opportunity for unreliability to arise in reporting or interpreting (novel) results given knowledge of theoretical predictions as there is for it to arise in arriving at hypotheses given knowledge of (nonnovel) results. This is Weinberg's point in the epigraph that opens this chapter. The historian Stephen Brush finds cases where scientists are as concerned about the former as the latter, leading him to suggest that "the preference for forecasting implies a double standard for theorists and observers, based on a discredited empiricist conception of science" (Brush 1989, 1127). Why not be as suspicious of novel results claimed to be in accord with a known theoretical prediction? Lest we be driven to suspect all experimental inference, some distinctions clearly need to be made. These distinctions, I propose, appeal to reliability considerations that scientists standardly employ in devising and interpreting their experiments.

8.1 Logical and Evidential-Relationship Views versus Historical and Testing Accounts

If I am correct that the goal of novelty is severity, then the dispute between those who do and those who do not accept some version of the novelty principle emerges as a dispute about whether severity—or, more generally, error characteristics of a testing process—matters in evaluating the import of evidence. By and large, thinking novelty matters goes hand in hand with thinking severity matters. (Where there are exceptions, there is some question about the consistency of the view. I return to this in chapter 10.) This correlation is borne out in the historical disputes as well as in the current debate between Bayesian and non-Bayesian philosophies of hypothesis appraisal.

Alan Musgrave on Logical versus Historical Theories of Confirmation

Several philosophers have enlightened us on the history of the novelty dispute (e.g., Giere 1983; Lakatos 1978; Musgrave 1974). Musgrave puts his finger on what is common to historical opponents of novelty—they hold what he calls a logical theory of confirmation.

> According to modern logical empiricist orthodoxy, in deciding whether hypothesis *h* is confirmed by evidence *e*, and how well it is

> confirmed, we must consider only the statements *h* and *e*, and the logical relations between them. It is quite irrelevant whether *e* was known first and *h* proposed to explain it, or whether *e* resulted from testing predictions drawn from *h*. (Musgrave 1974, 2)

One finds this stated plainly in Hempel 1965.

Some variant of the logical (or logicist) approach was implicitly held by historical opponents to RN. "We find it in Mill, who was amazed at Whewell's view" that successfully predicting novel facts gives a hypothesis special weight (Musgrave 1974, 2). Whereas Whewell held that

> men cannot help believing that the laws laid down by discoverers must be in a great measure identical with the real laws of nature, when the discoverers thus determine effects beforehand, (Whewell [1847] 1967, vol. 2, p. 64)

in Mill's view, "such predictions and their fulfillment are . . . well calculated to impress the uninformed. . . . But it is strange that any considerable stress should be laid upon such a coincidence by persons of scientific attainments" (Mill 1888, bk. 3, chap. 14, sec. 6, p. 356). Keynes, another logicist, similarly held that the "question as to whether a particular hypothesis happens to be propounded before or after examination of [its instances] is quite irrelevant" (Keynes [1921] 1952, 305).[4] Clearly, if confirmation is strictly a logical function between evidence (or statements of evidence) and hypotheses, when or how hypotheses are constructed *will* be irrelevant.

The logical approaches to confirmation ran into problems, however, precisely because they insisted on purely formal or syntactical criteria of confirmation that, like deductive logic, "should contain no reference to the specific subject matter of the hypothesis or of the evidence in question" (Hempel 1965, 10). Enter what Musgrave calls the "historical (or logico-historical) approach" to confirmation.

In Musgrave's neat analysis of the situation, the contemporary accounts of novelty in the Popper-Lakatos school arose out of attempts to avoid the paradoxes of the traditional logical approaches to confirmation by requiring various background considerations in the form of novelty requirements. Musgrave calls such accounts historical because, he believes, "it will presumably be a historical task to determine" what the background knowledge is. In particular, "all variants of the historical approach will make the confirmation of a scientific theory some-

[4] This is as quoted in Musgrave 1974, 2.

how depend upon the historical setting in which that theory was proposed" (Musgrave 1974, 7).

E-R versus Testing Accounts of Inductive Inference

The situation twenty years after Musgrave's article is rather similar—with one key difference. Few philosophers of science still harbor the view that a theory of confirmation is merely a formal relationship between evidence and hypothesis. What distinguishes their accounts is the kind of background thought to be needed. Bayesians let the background enter via prior probabilities. In the Bayesian analogy between induction and deduction from chapter 3 we hear echoes of the position staked out by logical approaches to confirmation.

In contrast, most who follow what I termed "testing accounts" see the background as coming in via methodological principles such as those based on novelty.[5] Members of the Popper-Lakatos school (e.g., Watkins, Musgrave, Zahar, Worrall) proposed notions of novelty where the background knowledge was intended to reflect Popperian goals. Musgrave tells us that "Watkins begins with Popper's thesis that genuine confirmations can result only from the failure of genuine attempts to refute a hypothesis, from severe tests" (Musgrave 1974, 5). The central aim in each case was to use background knowledge to distinguish genuine or severe tests from spurious ones.

Novelty in Testing Accounts of Inference. Three main types of novelty emerge. A novel fact for a hypothesis H may be (1) one not already known, (2) one not already used in arriving at or constructing H, or (3) one not already predicted (or one counterpredicted) by available hypotheses. The first corresponds to *temporal* novelty, the second, to *heuristic* or *use* novelty, the third to *theoretical* novelty. Each has itself been construed in a variety of ways, multiplying the number of different definitions. The third view, theoretical novelty, was the Popperian severity requirement discussed in chapter 6. The debate I will now consider revolves around temporal and heuristic or use novelty.[6]

[5] The more fundamental difference between these two approaches to the background is that testing accounts give a central role to what in statistical language are the error characteristics of a procedure of testing. This will be explored further.

[6] Calling in to take credit for the Trade Center bombing violates both types of novelty. Not only had the blast already occurred, it had already been reported on the news before the first phone call came in. It is a violation of use-novelty (or heuristic novelty) because that information was used to construct the particular hypothesis by filling in the blank: for example,___is responsible. So eschewing violations of either form of novelty explains our giving little credence to the calls. However, the mangled piece of steel (and a list of other clues) were also known and used to arrive at the identity of the driver of the suspected van.

A second school of testers—the "error statisticians"—has also upheld novelty principles in the form of rules of predesignation or rules against double counting of data. An important variant of the argument from error statistics warrants separate attention and will be considered in chapter 9. The argument most familiar to philosophers, represented by Ronald Giere, can be considered now with arguments from historical schools.

Recall that for a testing theorist the task of characterizing a good test is not distinct from saying when a hypothesis is warranted or well supported by evidence. To avoid confusion I have couched my account in terms of tests. However, Worrall and Giere often couch their remarks in terms of when data *support* hypotheses. They must not be mistaken as attempting to provide an evidential-relation measure. For them, to say when data support *H* is to say when data provide a good test of *H*—although degrees of goodness are still possible. The task for novelty criteria is to tell us what more beyond entailing or fitting evidence is required for a test to be genuine. The debate centers on how best to accomplish this.

The central problems are how background knowledge should be brought in so that: the account of testing is not turned into a subjective or relativistic affair, the resulting account accords with important cases of actual scientific appraisal, and there is a clear epistemological rationale for doing so. The ongoing debate has not made much progress in satisfying these three desiderata.

Temporal Novelty

On the view of temporal novelty, "novel fact" meant what it said: a novel fact was a newly discovered fact—one not known before its use in testing.[7] Known by whom? For some, it counts as known only if the general scientific community knows it, for others it is enough that a given investigator putting forth the hypothesis knows it (e.g., Gardner 1982).

The temporal view of novelty has been criticized on all three desiderata for judging novelty criteria. First, there is the problem of how the temporal novelty of the data can be characterized nonsubjectively. How, it is asked, can temporal novelty be determined without having to look into the psyches of individual scientists to determine what they knew and when they knew it? Second, the temporal novelty requirement denies special evidential significance to tests that intuitively seem

[7] Although a temporally novel fact is sometimes equated with a *predicted* fact, the term "predicted" in science generally does not require this temporal element—a point Stephen Brush (1989) makes.

to possess it. Take our example from the last chapter. Brownian motion was known long before the Einstein-Smoluchowski theory was proposed, yet was considered to have provided significant evidence for it. Likewise for the orbit of Mercury and Einstein's theory. Third, there is the question of its epistemological rationale. Why *should* time matter?

> If the time-order of theory and evidence *was* in itself significant for scientists then we should, I think, be reduced merely to recording this as a brute fact. For why on earth *should* it matter whether some evidence was discovered before or after the articulation of some theory? (Worrall 1989, 148)

In response to such objections to temporal novelty, novel accounts have been proposed in which novelty turns instead on the *heuristic* role of facts: on whether the theory it helped construct was in a certain sense ad hoc. Zahar (1973) suggested that a fact is *"novel with respect to a given hypothesis if it did not belong to the problem-situation which governed the construction of the hypothesis"* (p. 103). Old facts (i.e., facts not temporally novel) could be novel facts in this new sense so long as the theory was not devised to explain them. Musgrave and others criticized this view as being too subjective and psychologistic—even more so than temporal novelty. It seemed to make the answer to the question of whether a test was good relative to the specific aims of the designer of the theory. "To assess the evidential support of a theory 'one has to take into account the way [it] is built and the problems it was designed to solve'" (Musgrave 1974, 12). Furthermore, it seems that in Zahar's view the same evidence might accord a given theory as proposed by one scientist a different amount of support than as proposed by another, according to the heuristic route each scientist takes (p. 14).

Worrall reformulates Zahar's heuristic view: the question is not whether a theory was "devised to explain" a fact but whether the fact was "used to construct" the theory. With this new formulation Worrall intends to signal that although support is to be heuristic-relative, it is *not* to be person-relative (e.g., Worrall 1978b, 51). That is, while granting that the heuristic view allows the same theory, because differently arrived at, to be differently supported by the same evidence, Worrall believes that "the heuristic considerations which led to the construction of a theory can be objectively specified" (p. 51).

The Zahar-Worrall view of heuristic novelty may be called *use-novelty.* It requires that for evidence *e* to support hypothesis *H* (or for *e* to be a good test of *H*), in addition to *H* entailing *e, e* itself *must not have been used* in *H*'s construction. Worrall states the position as follows:

> The relation [of empirical support] holds if and only if the factual statement is implied by the theory but is not a member of the set of

> factual statements used in the construction of the theory. (Worrall 1978b, 50)

Since strict entailment is generally too strong,[8] I will allow that the use-novelty requirement, UN, for a good test is satisfied when

> i. *H* entails or is a good fit with *e*

and

> ii. *Use-novelty* UN: *e* is not used in the construction of *H*.

(Worrall also holds UN sufficient for a good test, but I leave that to one side. See Mayo 1991a.)

Use-novelty, or something very much like it, is endorsed—at least as a necessary condition—by other use-novelists as well as by temporal novelists. Its violation is commonly termed double-use or double-counting of data. If evidence is used in arriving at *H*, then it cannot be used again in *H*'s support. As a shorthand, let us call a hypothesis constructed to fit evidence *e* (however the construction is done) a *use-constructed hypothesis.* The use-novelty requirement for tests is this:

> *UN requirement:* Data *e* that was used to arrive at a use-constructed hypothesis *H* cannot also count as a good test of *H*.

The UN requirement does seem to reflect our ordinary intuitions in cases such as the after-trial claims of responsibility for the bombing of the World Trade Center and astrological retrodictions. This fact should also make it reasonable to suppose that the rationale for these intuitions, where correct, applies uniformly to day-to-day and scientific hypotheses.[9] The account I recommend does just this. Violating UN is correctly eschewed—whether in science or day-to-day reasoning—only if it results in violating reliability or severity.

Against this, proponents of the UN requirement recoil from ever crediting a use-constructed hypothesis for passing a test it was constructed to pass. Their basic intuition is this:

> If a hypothesis *H* has been arrived at to accord with data *e*, then that same data cannot also provide a good test of (or good support for) hypothesis *H*, *since H could not have failed this test.*

It is not that the constructed hypothesis is considered unbelievable or false but rather that the UN proponent denies that finding the accor-

[8] Worrall's own discussion bears this out. In any case, my arguments here will not turn on whether (i) requires strict entailment or allows a statistical type of fit.

[9] This counts against proponents of novelty who, when faced with counterexamples to UN, maintain that science is different.

dance with data should be credited to *H* because the accordance was assured—*no matter what.* This reasoning, while sounding plausible, is wrong.

Finding It Wrong for the Wrong Reasons

In finding the intuition underlying the UN requirement wrong, I am apparently in an odd sort of agreement with the Bayesian Way. Indeed, recent criticisms of the UN requirement, with few exceptions, have been leveled by those wearing Bayesian glasses. What has scarcely been noted, however, is that the Bayesian critiques are deeply flawed. From our discussion of key differences in aims or even from Musgrave's logical versus historical lesson, it is easy to guess at the flaw.

For Bayesian philosophers of science, just as with the earlier "logicist" approaches, there is no slot in which to take into account the novelty of the data. Thus when the rationale for the UN requirement is judged from this Bayesian standpoint, it is, unsurprisingly, found wanting. Finding UN unnecessary for earning high marks *according to the Bayesian account* of hypothesis appraisal, the Bayesian declares the arguments in favor of UN wrong (e.g., Howson and Urbach 1989, 549; Howson 1990). The problem is that even if the rule of novelty (RN) were good for its intended aims, running it through the Bayesian machinery has no chance of "passing" the rule. On Bayesian principles, if two hypotheses that entail evidence *e* are to receive different amounts of support from *e*, then the difference must lie in the prior probabilities. I differ from the Bayesian and concur with the UN proponent in holding that when a difference in appraisal is warranted, the fault lies in the testing process and not in our priors.

Adding to the confusion, there have been Bayesian attempts to support the non-Bayesian arguments for UN. Shooting holes in these Bayesian defenses have wrongly been taken to vitiate the non-Bayesian arguments. All of this warrants more careful scrutiny; I shall return to this discussion in chapter 10.

8.2 CHARACTERIZING UN VIOLATIONS (NONSUBJECTIVELY)

Worrall (1985) concedes that "allowing that heuristics play a role does indeed threaten to make confirmation a dangerously unclear and subjectivist notion" (p. 309). The viability of his position, he grants, rests on being able to find out if a theory is use-constructed by careful historical study, for example, by combing historical documents, notes, and letters without having to explore the psyches of individual scientists.

Three types of UN violations emerge in the accounts of Worrall and Giere:

Parameter fixing. There is a hypothesis with a free parameter *x* (a quantity or constant in some mathematical equation or other blank not yet fixed), for example, the number of molecules per unit is *x*. We can write the hypothesis with the blank as $H(x)$. The data *e* are used to work out the value of this or these parameters so that the resulting hypothesis, $H(e)$, yields, entails, accommodates, renders expected, or otherwise fits data *e*. This need not be a quantitative parameter. It includes any case where data are used to pin down a hypothesis sufficiently so as to account for or accommodate that same data. An example Worrall gives is the use-construction of a Newtonian explanation of Mercury's perihelion (Worrall 1978b, 48).

Exception barring or incorporation. Exception incorporation may arise when *H* fails to accord with result *e*, so *e* is anomalous for *H*. The constructed hypothesis, H', is the result of revising or qualifying hypothesis *H* so that H' accords with *e*. That is, H' is *H* plus some qualification or modification. An example would be where *H* is an alleged psychic's claim to be able to "see" via ESP a drawing in a sealed envelope. When he fails, he qualifies *H* so that it excludes cases where skeptical scientists are watching him. Thus revised, result *e* now allows *H* to pass.

Evidence as constraint. Giere alludes to another way in which UN may be violated, although in a sense it subsumes all the preceding ones. Here the violation occurs whenever the evidence acts as a "constraint" on any hypothesis that is going to be considered. An agent who follows this procedure will only put forward hypotheses that can accommodate a known result or phenomena *e*. Such a procedure, assuming it ends, must somehow accommodate *e*.

In each case *e* is being used to construct a hypothesis *H* to satisfy the condition that *H* "fits" *e*. Alternatively, *e* is used in constructing *H* to assure that *H* passes the test. Whenever the evidence *e* is taken at the same time as supporting or providing a good test of the use-constructed *H*, we have a UN violation.

Using Historical Data to Test Novelty Accounts

It is far from clear that these attempts to characterize violations of use-novelty ameliorate the difficulty of determining objectively whether a case is use-novel. In the one historical case that both Giere and Worrall look at with express interest in the question of novelty—

the case of Fresnel's (wave) theory of diffraction—they arrive at opposite pronouncements. Following the traditional reading of this episode, Giere finds that the ability of Fresnel's theory to account for the diffraction effect known to occur with straightedges was given less weight (by a prize committee at the time) than its ability to predict the temporally novel (and unexpected) "white spot" effect. In contrast, Worrall argues that accounting for the known straightedge result was given *more* weight (Worrall 1989, 142). But their disagreement goes further. Whereas Worrall holds up the straightedge effect as a case where UN is *satisfied*,[10] Giere holds it up as a case of a UN *violation* (of the evidence as constraint type).

This disagreement raises the problem of when historical data should be regarded as genuinely testing a methodological claim such as the UN requirement. Worrall takes this episode as evidence in support of use-novelty. Yet, even granting his reading of the case, namely, that temporal novelty did not seem to matter, the episode seems at most to be *consistent* with his position about UN (it is not a very severe test). Going by Worrall's discussion of the case, it is far from clear that the appraisal turned on novelty altogether—whatever the type. Worrall remarks:

> The report recorded that Fresnel had made a series of 125 experimental measurements of the external fringes outside the shadow of a straightedge, and that in this whole series the difference between observation and the value provided by Fresnel's integral was only *once* as much as 5/100 mm, only *three* times 3/100 mm and *six* times 2/100 mm. In all the other 115 cases disagreement between theory and observation did not exceed 1/100 mm. (Worrall 1989, 144)

Members of the prize committee were impressed, it seems, by how *often* Fresnel's predictions came very close to the observed results with straightedges. Their argument clearly assumed, as it required such for it to have weight, something like the following: If Fresnel's hypothesis were wrong, we would expect larger differences more often than were observed. That is, such good accordance would be very improbable in a series of 125 experiments if Fresnel's account was not approximately correct in regard to diffraction. It is this argument—an argument from error—that mattered, and not how Fresnel's account was constructed.

We can agree with Worrall (1989) that to assess support, "we need know nothing about Fresnel's psyche and need attend only to the development of his theory of diffraction as set out in great detail and

[10] Worrall's purpose in citing the Fresnel example is to argue that scientific judgments reflect a concern with use-novelty and not with temporal novelty.

clarity in his prize memoir" (p. 154). Psyches have nothing to do with it, and scientific reports (if adequate) are enough. But in my view what we really need to learn from these reports is not the route by which the theory was developed, but *the reliability of its experimental test.* I propose that the whole matter turns on how well the evidence—used or not—*genuinely indicates* that hypothesis *H* is correct. What matters is how well the data, together with background knowledge, rule out ways in which *H* can be in error. While this calls for attending to characteristics of the entire testing process, which may be influenced by aspects of the generation of test hypotheses and data, it does not call for reconstructing how a scientist came to develop a hypothesis. In planning, reporting, and evaluating tests, it is to the relevant characteristics of the testing process that scientists need to attend.

But I am ahead of myself. My point now is that it is difficult to argue for what historical cases show about some methodological rule without understanding how the rule functions in experimental learning. Worrall does not stop with testing UN via historical data but goes on, in several papers, to wrestle with its epistemological rationale, to which we will now turn.

8.3 The (Intended) Rationale for Use-Novelty is Severity

Despite the fairly widespread endorsement of something like the UN requirement (especially among testing accounts), enormous confusion about what might be its epistemological rationale persists. In this chapter I will focus on two variants of a single argument that has received considerable attention by philosophers, as found in discussions by Worrall and Giere. I shall argue that the (implicit or explicit) *intended* rationale for use-novelty is severity.

Violating Use-Novelty and the Ease of Erroneous Passing

An important advantage, Worrall (1989) claims, that use-novelty has over temporal novelty "is that it comes equipped with a rationale" (p. 148). Nevertheless, Worrall fails to come right out and say what that rationale is. The most telling hint is that he intends his use-novelty criterion UN to capture Popper's requirement for a genuine or severe test: "Many of Popper's most perspicacious remarks are . . . based on an intuitive notion of testability" (Worrall 1985, p. 313) embodied in the Zahar-Worrall use-novelty account, which, Worrall says, "Popper has never, I think, fully and clearly realized" (ibid.). Paying attention to the manner of theory construction can fully capture the spirit of Popper's intuition about tests, whereas Popper's purely logical account cannot. Popper's intuition, noted in chapter 6, is that we are to "try to think of cases or situations in which [a hypothesis] is likely to fail, if it is false" (Popper 1979, 14).

At the heart of the matter is insisting on *riskiness* of some sort. Popper's favorite example is the test of Einstein's theory by checking the predicted light deflection during the eclipse of 1919:

> Now the impressive thing about this case is the *risk* involved in a prediction of this kind. . . . The theory is *incompatible with certain possible results of observation*—in fact with results which everybody before Einstein would have expected. (Popper 1962, 36)

Popper contrasts this with the way popular psychological theories of his day seemed able to accommodate any evidence and with how unwilling or unable the latter were at putting their hypotheses to genuine test. Yet Popper's logical requirement for a genuine test, I concur with Worrall, does not capture his own informal remarks about what a good test requires (as seen in section 6.7).

Heuristic novelty, Worrall proposes, does a better job than both temporal and theoretical (Popperian) novelty at capturing the needed risk requirement. Worrall reasons that

> if some particular feature of *T* was in fact tied down on the basis of *e* . . . then checking *e* clearly constitutes no real test of *T*. . . . In such a case even though *e* follows from *T* and hence not-*e* is, in Popper's terminology, a potential falsifier of *T*—it wasn't *really* a potential falsifier of *T*, since *T* was, because of its method of construction, never at any risk from the facts described by *e*. (Worrall 1989, 148–49)

Ronald Giere (1984a) makes a parallel assertion. Whether known (nontemporally novel) facts may provide good evidence for a hypothesis, Giere claims,

> depends on whether the known facts were used in constructing the model and were thus built into the resulting hypothesis. If so, then the fit between these facts and the hypothesis provides no evidence that the hypothesis is true. These facts had no chance of refuting the hypothesis even if it were wildly mistaken. (P. 161)

The final sentences of Worrall and Giere's passages can and have been misinterpreted (see chapter 10). They should not be taken to mean that some particular data *e* could not have but compared favorably with *H*. For that would be so whenever a hypothesis fits or accords favorably with data—even in the best of tests. After all, if *H* is in accordance with *e*, then there is no chance that it is not in accordance with *e*. What Worrall and Giere must intend to be pointing out about use-constructed cases is that evidence *e*—whatever it is—is guaranteed to accord with hypothesis *H* whenever *H* is deliberately constructed to be in accordance with *e*. Any facts resulting from such a process, Giere is saying, had no chance of refuting the hypothesis (constructed to fit them) "even if it were wildly mistaken." That is, Giere is saying, the test fails to be severe in the sense of error-severity.

But why do they suppose violating UN leads to violating severity? I will consider their arguments in turn, for each represents a well-entrenched position.

Worrall and a False Dilemma

Worrall's (1989) position goes like this: Consider the kind of reasoning we seem to use when we *do* take a theory's empirical success as showing it to be true, empirically adequate, or in some way reflecting "the blueprint of the Universe"—"whether or not it can be given some further rationale" (p. 155).

> The reasoning appears to be that it is unlikely that the theory would have got this phenomenon precisely right just "by chance." . . . The choice between the "chance" explanation and the "reflecting the blueprint" explanation of the theory's success is, however, exhaustive only if a third possibility has been ruled out—namely that the theory was engineered or [use-constructed]. (Worrall 1989, 155)

For, in the use-constructed case, Worrall says,

> the "success" of the theory clearly tells us nothing about the theory's likely fit with Nature, but only about its adaptability *and* the ingenuity of its proponents. (Ibid.)

The presumption seems to be that in use-constructed cases the proponent's ingenuity and/or the theory's adaptability suffice to explain the success, and that such success is likely even if the theory does not fit well with Nature. This is tantamount to asserting that use-novelty is necessary for severity. Let us mark this premise (which UN proponents state in various ways) as (*). Here Worrall states it as follows:

> (*) If *H* is use-constructed, then it cannot be argued that its successfully fitting the evidence is unlikely if *H* is incorrect.

His argument seems to be that since the success of a use-constructed hypothesis can be explained by its having been deliberately constructed to accord with the evidence, there is no need or no grounds for seeking its explanation in the correctness of *H*. We have "used up" the import of the evidence, so to speak.

Is there not some confusion here between two senses of explaining a success? Consider an imaginary trial of a suspect in the World Trade Center bombing:

Prosecutor: There is no other way to explain how well this evidence fits with X as the culprit (the twisted metal matching the rented van, the matching fingerprints, etc.) save that X was (at least part of the group) responsible.
Defense: Yes there is. The investigators built up their hypothesis about the guilty party so as to account for all of the evidence collected.

It matters not, for the sake of making out the equivocation, that the hypothesis here is different from most scientific ones.

In other words the problem is to find a way of accounting for some evidence or some observed effect. Let us say that the problem is declared solved only when a hypothesis is reached that satisfactorily accounts for it. Now suppose the problem is solved, but that particular features of the effect to be accounted for have been used in reaching the solution. One can ask: Why does the hypothesized solution accord so successfully with the evidence? In one way of reading this question (Worrall's), a perfectly appropriate answer is that *any* solution put forward would accord with the evidence: the solution was use-constructed. Quite a different reading of this question—the one to which the prosecutor is answering in the affirmative—has it asking whether the accordance with the evidence indicates the correctness of the hypothesis. The question, on this second reading, is whether the successful accordance with the evidence satisfactorily rules out the ways in which it would be an error to declare *H*. That *H* was constructed to account for the evidence does not force a "no" answer to this second question. *H* might be use-constructed, but use-constructed reliably.

Conflating the two renders mysterious ordinary scientific discernments. Finding a correlation between a certain gene and the onset of Alzheimer's disease led Dr. Allen Roses (from Duke University) to hypothesize a genetic cause for certain types of Alzheimer's. What accounts for his hypothesis successfully explaining this correlation? The answer, interpreting the question one way, might be that Dr. Roses found this correlation and used it, as well as known properties of the gene, to develop a hypothesis to account for it. A separate question is why his genetic explanation fits the facts so well. One major Alzheimer's researcher declared (despite Roses's hypothesis going against his own work) that after only ten minutes he could see that the data pointed to Roses's hypothesis. Yes, Roses used the data to construct his hypothesis. The particular way in which he did so, however, showed that it provided at least a reasonably severe test in favor of his hypothesis (that the gene ApoE has a genuine connection with Alzheimer's). Had Roses's evidence been less good—as was the case a few years earlier—the scientists would have (and did) largely dismiss the agreement.[11]

To summarize this subsection, we have good grounds for the correctness of *H* to the extent that the circumstances by which *H*'s assertion would be in error have been well ruled out. Evidence may be used to construct *H* and still do a good job of ruling out *H*'s errors. To suppose that these are mutually exclusive is a false dilemma.

Proponents of Zahar and Worrall's argument for use-novelty might agree with all this. Nevertheless they may insist that except for when the use-construction method is clearly reliable, they are right to require, or at least to prefer, use-novel to use-constructed hypotheses. Let us grant them this and declare that there is no disagreement between us here. But let us push them a little further. What is the worry in those cases where the use-construction method is *not* clearly reliable? The worry is that it is one of those dreadful methods of cooking up hypotheses—the kind of method that is always everywhere available. And what is wrong with the kind that is always everywhere available? It is available whether or not the hypothesis reached is true! (Remember gellerization.)

That is precisely my point. The reason for eschewing use-construction methods is the condemnation of unreliable procedures of data accommodation. The underlying rationale for requiring or preferring use-novelty is the desire to avoid unreliable use-constructing procedures. The best spin I can glean from the Zahar-Worrall rationale for requiring UN shows it to be the one I claim. If there is a different rationale, perhaps its proponents will tell us what it is.

An Argument from Giere

Giere does not beat around the bush but plainly declares (*) from the start. Where Worrall emphasizes the cleverness of proponents and

[11] When in 1991 Roses first reported having pinpointed the approximate location of a gene in families with late-onset Alzheimer's, it was given little credence by neuroscientists. It not only went against the generally accepted thinking of the time, but Roses clearly had not yet ruled out the numerous ways in which he might have been mistaken to infer a causal connection from such a correlation. Later, a biochemist at Duke's lab sought out natural substances that chemically bind to amyloid. Perhaps some substance was sticking to amyloid, causing the buildup of plaques in the brain. What he thought was an experimental contaminant was ApoE. The biochemist was able to take advantage of the fact that studies of heart disease had already located and isolated the gene for this cholesterol-carrying sub-

the adaptability of hypotheses in order to explain why it is no wonder that a success accrued, Giere notes how the tester himself may simply refuse to consider any theory or model that does not successfully fit the data.

Although Fresnel's wave model accounted for the known diffraction pattern of straightedges, Giere says this did not count as a good test of the model because the straight-edged pattern violated use-novelty in the sense that it "acted as a constraint on his theorizing":

> He [Fresnel] was unwilling to consider any model that did not yield the right pattern for straight edges. Thus we know that the probability of *any* model he put forward yielding the correct pattern for straight edges was near unity, independently of the general correctness of that model. (Giere 1983, 282)

What emerges once again, though much more directly put, is a version of premise (*), that use-novelty is necessary for severity. Because Giere gives us a separate account of testing, it is possible to extricate his full argument for the UN requirement.

Giere, at least in 1983, endorsed an error-statistical account of testing (along the lines of Neyman-Pearson).[12] He characterized "an *appropriate test* as a procedure that has *both* an appropriately high probability of leading us to accept true hypotheses as true and to reject false hypotheses as false" (Giere 1983, 278). A test should, in short, have appropriately low error probabilities. We have:

1. A successful fit does not count as a good test of a hypothesis if such a success is highly probable even if the hypothesis is incorrect. (That is, a test of H is poor if its severity is low.)

stance. It turned out that the gene for ApoE was located in the very place Roses had found the suspect gene in families with Alzheimer's.

[12] There are important differences between the error-statistical account I favor and Giere's most current decision-theoretic account of testing (Giere 1988). First, I reject the idea of modeling scientific inference as deciding to choose one model over another. (See chapter 11.) Second, Giere's decision strategy is to choose a model M when evidence is very probable were M correct while being very improbable were some alternative model correct. Yet evidence may be very improbable under a rival to model M and not count as passing M severely. (See, for example, chapter 6.) Finally, the models in Giere's "model-based" probabilities are allowed to be full-blown scientific models, such as Dirac's and Schrödinger's models (Giere 1988, chap. 7). The assessments of these probabilities rely more or less on the intuitive judgments of scientists, and as Giere's own discussions show, are subject to serious shifts (even for a given scientist). In my account, the probability assessments must be closely tied to experimental models about which the statistics are at least approximately known.

Violating use-novelty, Giere suggests, precludes (or at least gets in the way of) the requirement that there be a low probability that *H* is accepted if false. This gives an even stronger version of premise (*) than that found in Worrall:

> (*) If a hypothesis *H* is use-constructed, then its success is high ("near unity") even if it is false.

From premise (1) and (*) we get the conclusion that use-construction procedures fail to count as good tests (of the hypotheses they reach). I agree with premise (1)—it is premise (*) that I deny.

The Gierean Argument for (*) *the Necessity of UN.* Giere's argument for the necessity of UN seems to be that in a use-constructed case, a successful fit is obviously not unlikely—it is assured no matter what. That is, the basis for (*) is an additional premise (2):

2. *Basis for* (*): If *H* is use-constructed, then a successful fit is assured, no matter what.

Ah, but here is where we must be careful. This "no matter what" can be interpreted in two ways. It can mean

a. no matter what the data are

or it can mean

b. no matter if *H* is true or false.

Although the assertion in (2) is correct with the replacement in (*a*), thus construed it provides no basis for (*), that UN is necessary for severity. For (2) to provide a basis for (*), the replacement would have to be as in (*b*). However, (2) is false when replaced with the phrase in (*b*). Once this flaw in the pivotal intuition is uncovered it will be seen that UN fails to be a necessary condition for a severe test, and that (*) is false.

To clarify, consider two different probabilities in which one might be interested in appraising the test from which a passing result arises:

A. The probability that test *T* passes the hypothesis it tests.
B. The probability that test *T* passes the hypothesis it tests, *given that the hypothesis is false.*

Note that here two things may vary: the hypothesis tested as well as the value of *e*. Now consider a test procedure that violates UN in any of the ways this can come about. To abbreviate, let *H*(*e*) be a use-constructed hypothesis—one engineered or constrained to fit evidence *e* (it may be read "*H* fixed to fit *e*"). Then the following describes what may be called "a use-constructed test procedure":

> *A use-constructed test procedure T:* Use *e* to construct *H*(*e*), and let *H*(*e*) be the hypothesis *T* tests. Pass *H*(*e*) with *e*.

Since *H*(*e*), by definition, fits *e*, there is no chance of *H*(*e*) *not* passing a use-constructed test *T*. The relative frequency of passing in a series of applications of a use-constructed test equals 1. That is,

A. The probability that (use-constructed) test *T* passes the hypothesis it tests

equals 1.

But that is different from asserting that the test *T* is guaranteed to pass the hypothesis it leads to testing, *even if that hypothesis is false.* That is, asserting that (A) equals 1 is different from asserting that

B. The probability that (use-constructed) test *T* passes the hypothesis it tests, *even if it is false,*

equals 1.

Yes, the use-constructing procedure always leads to passing one hypothesis or another—provided it ends—but this is not incompatible with being able to say that it never, or almost never, leads to passing a false hypothesis.

Imagine that each experimental test rings a bell if the hypothesis it tests passes, and sounds a buzzer if it fails. (A) is a statement about how often the bell would ring in a series of experimental tests of a certain kind. A use-constructed test procedure would always culminate in a ring—if it ended at all.[13] So (A) equals 1 in the case of use-constructed tests. The probability in (B), on the other hand, asks about the incidence of erroneous bell ringing, where by erroneous bell ringing I mean that the test rings the bell when a buzzer should have been sounded. This need not equal 1, even in use-constructed tests. It can even be 0. Those who hold UN necessary do so because violating UN leads to a test that has to result in sounding the bell and never in sounding the buzzer. The assumption is that severity requires some of the test results to lead to buzzing, but this is a mistake.

The illustration with bells and buzzers is just to bring out, once again, the idea of an experimental (or sampling) distribution. A particular experimental test is viewed as a sample from a population of such experimental tests. The probabilities refer to relative frequencies with

[13] In many cases procedures can be guaranteed to end.

which a "sample test" has some characteristic, in this (actual or hypothetical) population of tests. The characteristic in (A) is "passing one hypothesis or another." The characteristic in (B) might be described as "giving an erroneous pronouncement on the hypothesis passed." How to calculate these probabilities is not always clear-cut, but for my point it is enough to see how, in cases where they *can* be calculated, the two are not identical.

This is a beautiful example of how the informal side of arguing from error can and should lead the way in disentangling the confusions into which one can easily wade in trying to consider the formal probabilistic arguments. To understand why (A) differs from (B), and, correspondingly, why UN is not necessary for severity, we need only think of this informal side. In particular, we need only think of how a procedure could be sure to arrive at some answer and yet be a procedure where that answer is rarely or never wrong. This is the basis for the counterexamples I will now consider.

8.4 USE-NOVELTY IS NOT NECESSARY FOR SEVERITY: SOME COUNTEREXAMPLES

To give a counterexample to the thesis that UN is necessary for severity, I have to describe a case that violates use-novelty yet provides a severe test of the use-constructed hypothesis.

Example 8.1: SAT Scores

For a trivial but instructive example consider a hypothesis H about the average SAT score of the students who have enrolled in my logic class:

$H(x)$: the average SAT score (of students in this class) $= x$,

where x, being unspecified, is its free parameter. Fixing x by summing up the scores of all n students and dividing by n qualifies as a case of parameter-fixing yielding a use-constructed hypothesis $H(e)$. Suppose that the result is a mean score of 1121. The use-constructed hypothesis is

$H(e)$: the average SAT score $= 1121$.

Surely the data on my students are excellent grounds for my hypothesis about their average SAT scores. It would be absurd to suppose that further tests would give better support. For hypothesis H follows deductively from e. Since there is no way such a result can lead to passing H erroneously, H passes a maximally severe test with e.

In much the same vein, Glymour, Scheines, Spirtes, and Kelly (1987) allude to the procedure of counting 12 people in a room and constructing the hypothesis that there are 12 people in the room. One may balk at these examples. Few interesting scientific hypotheses are entailed by experimental evidence. But allowing that such cases provide maximally severe tests while violating UN suffices to show that criterion UN is not necessary for severity. It may be asked: Is not UN required in all cases *other* than such maximally severe ones? The answer is no. Tests may be highly severe and still violate UN. The extreme represented by my SAT example was just intended to set the mood for generating counterexamples. Let us go to a less extreme and very common example based upon standard error statistical methods of estimation.

Example 8.2: Reliable Estimation Procedures

We are very familiar with the results of polls in this country. A random sample of the U.S. population is polled (the sample size being specified along the lines discussed in chapter 5), and the proportion who approve of the President is recorded. This is the evidence *e*. Say that 45 percent of the sample approve of the President. The poll report would go on to state its margin of error, say, of 3 percentage points. (The margin of error is generally around 2 standard deviations.) The resulting report says: Estimate that 45 percent of the U.S. population plus or minus 3 percentage points approve of the President. The report is a *hypothesis* about the full population (that *p*, the population proportion, is in the interval [.42, .48]). The estimate is constructed to accord with the proportion in the sample polled. The procedure may be characterized as follows: Hypothesize that *p*, the proportion who approve of the President, is an interval around the data *e*, the observed proportion who approve. The interval is given by the margin of error. Call it ∂. So the procedure is to infer or "pass" hypothesis $H(e)$ where

$H(e)$ asserts: p is equal to $e \pm \partial$.

This procedure, (confidence) interval estimation, will be discussed more fully in chapter 10. The margin of error corresponds to giving the overall reliability of the procedure. A 95 percent estimation procedure has a .95 probability of yielding correct estimates. What is known is that any particular estimate (hypothesis) this procedure yields came about by a method with a high reliability. Depending on the poll, the uncertainty attached might be .05 or .01 (reliability .95 or .99). The hypothesis thereby reached may be false, that is, the true value of *p* may be outside the estimated range, but, we argue, if *p* were outside the interval, it is very unlikely that we would not have detected this in the poll. With high probability we would have.

Now consider this question: Is it possible for the data e—the observed proportion who approve of the president in the sample—to fail to pass the hypothesis $H(e)$ that will be arrived at by this estimation procedure? Given that $H(e)$ is use-constructed (to be within ∂ units from e), the answer must be no. For the procedure, by definition, passes the hypothesis: p is in the interval $e \pm \partial$. Even before we have a specific outcome we know there is no chance that the result of this data generation process will fail whatever (use-constructed) hypothesis $H(e)$ the procedure winds up testing.

Compare this with the test's severity in passing $H(e)$. The fit with the resulting hypothesis (the interval estimate $H(e)$) is given by the specified margin of error ∂, say 2 standard deviation units. It is rare for so good a fit with $H(e)$ to occur unless the interval estimate $H(e)$ is true (i.e., unless the population proportion really is within 2 standard deviation units of e). So the severity in passing $H(e)$ is high. (The reasoning is this: To say that the interval estimate $H(e)$ is not true means that the true population proportion p is not in the hypothesized interval. But the p values excluded from this interval are those that differ from the observed proportion who approve, e, by more than 2 standard deviation units (in either direction), and the probability of e differing from p by more than 2 standard deviation units is small [.05].)[14]

Contrast this estimation procedure with one that, regardless of the observed outcome, winds up estimating that at least 50 percent of the population approve of the president. This "wishful thinking" estimation procedure does suffer from a lack of reliability or severity: it always infers a 50 percent or better approval rating in the full population even if the true (population) proportion is less than that. Regardless of whether use-constructing is involved or not, it is for this reason that we condemn it.

An Anticipated Objection. Against my counterexamples one might hear the following objection. The data in my examples provide *evidence* for the hypotheses reached, but they are not tests. I am confusing tests with evidence.

In my account, which is a testing account, there is no distinction.

[14] One must be careful to avoid misinterpreting this probability. The .05 does not refer to the probability that the particular estimate arrived at is true, i.e., includes the true value of parameter p. That probability is either 0 or 1. It is the procedure that formed the estimate of p that has a .05 probability of yielding false estimates.

To insist that any example I might give where evidence is used to construct a hypothesis cannot count as a test is to beg the question at issue. The question is whether there ever exists a use-constructed example where the evidence is, nevertheless, good support for or a good test of the hypothesis. Those who deem UN necessary—at least Giere and Worrall—do not say that non–use-novel data count as great evidence but no test. They say that such data fail to count as good evidence or good support for the hypotheses constructed, and they say this *because* the data fail to test these hypotheses. What is more, the method of standard confidence interval estimation is mathematically interchangeable with a corresponding statistical test. In a nutshell, the parameter values within the interval constructed consist of values that would pass a statistical test (with corresponding reliability). So I am in good company in regarding such procedures as tests.

A Curious Bayesian Aside. Curiously, confidence intervals have also been appealed to (e.g., by Howson) in Bayesian arguments against the necessity of use-novelty. Swept up in the task of showing the intuitive plausibility of use-constructed estimations, Howson allows himself the admission that

> there is no question but that confidence interval estimates of physical parameters, derived via some background theory involving assumptions about the form of the error distribution, are the empirical bedrock upon which practically all quantitative science is built. (Howson 1990, 232)

He is quick to add that the Bayesian can show how to assign a high degree of belief to the correctness of the estimate, as if to say that Bayesians can endorse the estimate as well. But the force of the intuition to which Howson is appealing is plainly the reliance science puts on the standard, Neyman-Pearson, estimates. (Hark back to our analogy of Leonardo da Vinci in chapter 3.) The irony of this Bayesian reliance on the intuitive plausibility of non-Bayesian procedures will not be fully appreciated until chapter 10.[15]

8.5 Summary and Some Consequences for Calculating Severity

Tests of use-constructed hypotheses are eschewed because a passing result is assured. But what matters is not whether passing is assured

[15] Other examples occur in Howson 1984 and Howson and Urbach 1989. Nickles (1987) makes a point similar to Howson's, informally. Like the use-novelist

but whether erroneous passing is. There is no problem with a test having a high or even a maximal probability of passing the hypothesis it tests; there is only a problem if it has a high probability of passing hypotheses erroneously. Hypotheses might be constructed to accord with evidence *e* in such a way that although a passing result is assured, the probability of an erroneous passing result is low; equivalently, the test's severity is kept high. The common intuition to eschew using the same data both to construct and to test hypotheses (to require UN), I claim, derives from the fact that a test that violates UN is guaranteed to pass the hypothesis it tests—no matter what the evidence. But this does not entail that it is guaranteed to pass some hypothesis *whether or not that hypothesis is false.* Indeed, a use-constructed test may have a low or even no probability of passing hypotheses erroneously. Granted, *if* a test is guaranteed to pass the hypothesis it tests, even if that hypothesis is false (i.e., (B) equals 1), then the test is guaranteed to pass the hypothesis it tests (i.e., (A) equals 1); but the converse does not hold.[16]

To hold UN as necessary is to overlook deliberate rules for use-constructing hypotheses with high or even maximal severity. Consider first a rule for using *e* to construct *H*(*e*) so as to ensure maximal severity of the test:

> *Rule R-1 for constructing maximally severe tests:* Construct *H*(*e*) such that a worse fit with *e* would have resulted (from the experiment) unless *H*(*e*) were true (or approximately true).

(That is, if *H*(*e*) were false, a worse fit would have occurred.) Such a rule is obviously available only in very special cases. But the point is that to calculate the probability in (B), the probability of erroneously passing the (use-constructed) hypothesis tested, requires taking into account the construction rule employed—in this case rule R-1. That is why the severity criterion is modified for cases of hypothesis construction in section 6.6.

Let us abbreviate a use-construction test that arrives at its hypothesis via rule R-1 as test *T*(R-1). Test *T*(R-1), by the stipulated definition, is guaranteed to pass *any* hypothesis fixed to fit *e*. (So (A) equals 1.) Nevertheless, the probability of (B)—the probability needed for calculating severity (by taking 1 minus it)—is this:

and unlike me, however, Nickles denies that data used to fix the parameter can count as giving what he calls generative as well as consequential support to a hypothesis so fixed.

[16] One might be led to the error of thinking that the converse does hold if one erroneously takes the "if" clause in (B) as a material conditional instead of as the appropriate conditional probability.

> (B) in test T(R-1): the probability that test T(R-1) passes the hypothesis it tests, given that hypothesis is false.

This equals 0. As such, the severity of a test of any hypothesis constructed by rule R-1 is 1—the severity is maximal.

The rule used in fixing the mean SAT score of students in my class is an example of rule R-1. While there is always some rule by which to arrive at a use-constructed H (the so-called "tacking" method of use-construction will do), the ability to apply the very special rule R-1 is hardly guaranteed. But if one does manage to apply R-1, the constructed hypothesis to which it leads cannot be false. Although one can rarely attain the security of rule R-1, the experimenter's tool kit contains several use-constructing rules that afford a high degree of reliability or severity, call it π. Such a use-construction rule may be written as rule R-π:

> *Rule R-π for constructing highly severe tests* (e.g., to degree π): Construct $H(e)$ such that the probability is very small (1-π) that a result from the experiment would accord as well with $H(e)$ as does e, unless $H(e)$ were true (or approximately true).

Examples of high severity construction rules are found in rules for the design and interpretation of experiments. They are the basis of standard estimation theory, as example 8.2 showed.

Let test T(R-π) be the use-construction test based on a construction rule (R-π). What is the severity of test T(R-π)?

The answer, of course, is π.

The Informal Calculation of Severity in Use-Constructed Tests

One need not be able to formally calculate the severity π. The identical rationale underlies informal rules for using evidence. In qualitative experimental tests one may only be able to say that the severity is very high or very low, yet that is generally enough to assess the inference. I might mention a (real) example that first convinced me that UN is not necessary for a good test. Here evidence was used to construct as well as to test a hypothesis of the form

> $H(x)$: x dented my 1976 Camaro.

The procedure was to hunt for a car whose tail fin perfectly matched the dent in my Camaro's fender to construct a hypothesis about the likely make of the car that dented it. It yielded a hypothesis that passed a high severity test. I was able to argue that it is practically impossible for the dent to have the features it has unless it was created by a specific type of car tail fin. Likewise for the rule the investigators followed in pinpointing the driver of the van carrying the explosive in the World Trade Center bombing. Such rules violate use-novelty; but they correctly indicate attributes of the cause of the explosion and the dent in my Camaro (and ultimately the identity of the drivers) because they are severe in the sense of rule R-π.

One may object, But science is not like that, there are too many possible alternative hypotheses. Having discussed the problem of alternative hypotheses in chapter 6, I am assuming those points here. If the objector persists that there is no way to obtain reliable knowledge or have a severe test of such and such a scientific theory, then, assuming he or she is correct, it must be agreed that no reliable use-construction procedure can accomplish this either. But this is irrelevant to my thesis that there are reliable experimental arguments that violate use-novelty.

We might connect the point of this chapter to our recasting of Kuhn in chapter 2. What is objectionable is not that practitioners are determined to find a way of accommodating data (to solve a given problem); what is objectionable is an accommodation that is not severely constrained (e.g., that it involves changing the problem), which results in unsolved problems often being declared solved when they are not. Alternatively, in a reliable use-construction one can argue that if $H(e)$ were incorrect, then with high probability the test would not have led to constructing and passing $H(e)$.

In one passage—although it is almost only in passing—Popper seems to capture what I have in mind about warranting severity. He says (in replying to his critics) that

> supporting evidence consists solely of attempted refutations which were unsuccessful, *or of the "knowledge" (it does not matter here how it was obtained) that an attempted refutation would be unsuccessful.* (Popper 1974, 992; emphasis added)

In a reliable use-constructed case one can sustain this "would be" argument. This is just what is wanted to affirm that evidence indicates the "probable success" of the hypothesis (in the sense of chapters 4 and 5).

Proponents of use-novelty often view Whewell and Peirce as forerunners of their view. While to an extent they are right, both Whewell and Peirce also discuss the kinds of cases where use-constructions are allowable. (I shall save Peirce's remarks for our later discussion of him.) Whewell also considered "the nature of the artifices which may be used for the construction of formulae" when data of various types are in hand (Whewell [1847] 1967, 392). The artifices he has in mind correspond to cases where the construction may be seen to ensure high severity. The "special methods of obtaining laws from observations" (p. 395) that Whewell cites include the method of curves, the method of means, the method of least squares, and the method of residues (essentially the statistical method of regression).

Such rules are typical, as would be expected, where violations of UN cannot be helped: where hypotheses are arrived at and affirmed by data, and it is impossible or impractical to obtain additional evidence (e.g., theories about dinosaurs, evolutionary theory, epidemiology, anthropology). As the next section shows, however, violations of UN are required even in cases lauded as models of severe and crucial tests, such as the 1919 eclipse tests of Einstein's gravitational hypothesis. Indeed, once the piecemeal aspect of testing is uncovered, such use-construction rules are indispensable. The same data may be used both to construct and ground hypotheses—so long as it is improbable that reaching so good an agreement is erroneous.

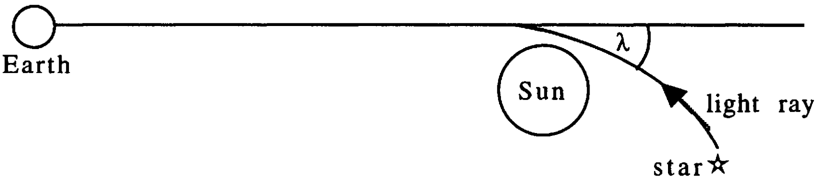

FIGURE 8.1. Deflection of starlight by the sun.

8.6 The 1919 Eclipse Tests of Einstein's Law of Gravitation

According to Einstein's theory of gravitation, to an observer on earth, light passing near the sun is deflected by an angle, λ, reaching its maximum of 1.75″ for light just grazing the sun. Terrestrial tests of Einstein's gravitation law could not be severe, since any light deflection would be undetectable with the instruments available in 1919. Although the light deflection of stars near the sun (approximately 1 second of arc) *would* be detectable, the sun's glare renders such stars invisible, save during a total eclipse. "But," as Arthur Eddington ([1920] 1987, 113) noted, "by strange good fortune an eclipse did happen on May 29, 1919," when the sun was in the midst of an exceptionally bright patch of stars, providing a highly severe test such as would not recur for many years. Two expeditions were organized: one to Sobral in northern Brazil, another (including Cottingham and Eddington) to the island of Príncipe in the Gulf of Guinea, West Africa.

Eddington, Davidson, and Dyson, the astronomer royal (henceforth Dyson et al. 1920), outline three hypotheses "which it was especially desired to discriminate between" (p. 291). Each is a statement about a parameter, the deflection of light at the limb of the sun, λ (in arc seconds):

1. Gravitation affects starlight according to Einstein's law of gravitation: the deflection at the limb of the sun $\lambda = 1.75''$.
2. Gravitation affects light according to the Newtonian law of gravitation: the deflection of a star at the limb of the sun $\lambda = 0.87''$.
3. Gravitation does not affect light, $\lambda = 0$.

The "Newtonian"-predicted deflection, (2), which stems from assuming that light has a certain mass and follows Newton's law of gravity, is exactly half that predicted by Einstein's law. Before setting out for Príncipe, Eddington suggests that

> apart from surprises, there seem to be three possible results: (1) A deflection amounting to 1.75" . . . which would confirm Einstein's theory; (2) a deflection of 0.87" . . . which would overthrow Einstein's theory, but establish that light was subject to gravity; (3) no deflection, which would show that light, though possessing mass, has no weight, and hence that Newton's law . . . has broken down in another unexpected direction. (Eddington 1918, 36)

A little over one year later, the results are in, and the conclusions given:

> The results of the expeditions to Sobral and Príncipe can leave little doubt that a deflection of light takes place in the neighbourhood of the sun and that it is of the amount demanded by Einstein's generalised theory of relativity, as attributable to the sun's gravitational field. (Dyson et al. 1920, 332)

This capsulizes the two key inferences from the eclipse inquiry: first, that there is a deflection effect of the amount predicted by Einstein as against Newton (i.e., the "Einstein effect"), and second, that the effect was "attributable to the sun's gravitational field" as described in Einstein's hypothesis.

The appraisal of the results by numerous scientists consisted of two corresponding parts or stages, which I label i and ii. Stage i involved inferences about the value of λ and critical discussions of these inferences, stage ii, inferences about the cause of λ and the associated (heated) discussions about these inferences. Each stage involved testing more local hypotheses, first to discriminate between the values of parameter λ, and second to discriminate between causes of the observed λ. Eddington, an adept data analyst, provides lavish and fascinating discussions of the nitty-gritty details of the data extraction and modeling. This, together with the intrinsic importance of the case, makes it an excellent subject for applying the full-blown hierarchy of models framework. Aspects of the data gathering were touched on in chapter 5. Lest I try my readers' patience, however, I will limit my discussion to aspects of the case most relevant to the present issue.

Stage i: Estimating the Eclipse Deflection at the Limb of the Sun

The "observed" deflection (on May 19), as with most experimental "results," is actually a hypothesis or estimate. Due to two major sources of error, arriving at the result is a matter of statistical inference: First, one does not observe a deflection, but at best observes (photographs of) the positions of certain stars at the time of the eclipse. To "see" the deflection, if any, requires learning what the positions of these same stars would have been were the sun's effect absent—a "control" as it were. Eddington remarks:

> The bugbear of possible systematic error affects all investigations of this kind. How do you know that there is not something in your apparatus responsible for this apparent deflection? . . . To meet this criticism, a different field of stars was photographed . . . at the same altitude as the eclipse field. If the deflection were really instrumental, stars on these plates should show relative displacements of a similar kind to those on the eclipse plates. But on measuring these check-plates no appreciable displacements were found. That seems to be satisfactory evidence that the displacement observed during the eclipse is really due to the sun being in the region, and is not due to differences in instrumental conditions. (Eddington [1920] 1987, 116)

Where the check plates could serve as this kind of a control, the researchers were able to estimate the deflection by comparing the position of each star photographed at the eclipse (the eclipse plate) with its normal position photographed at night (months before or after the eclipse), when the effect of the sun is absent (the night plate). Placing the eclipse and night plates together allows the tiny distances to be measured in the x and y directions, yielding ∂x and ∂y (see figure 8.2). These values, however, depend on many factors: the way in which the two plates are accidentally clamped together, possible changes in the scale—due mainly to the differences in the focus setting that occur between the exposure of the eclipse and the night plates—on a set of

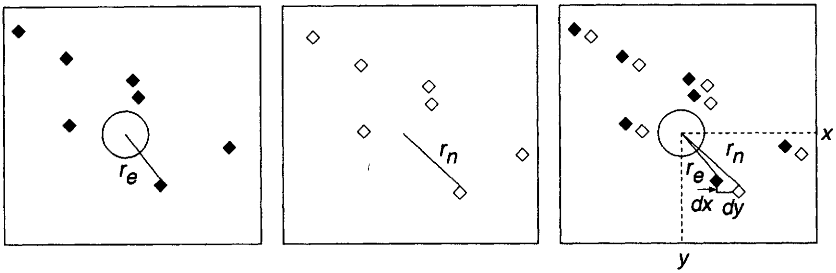

FIGURE 8.2. Comparing the "eclipse plate" and the "night plate" (adapted from von Klüber, 1960, 52). (a) "eclipse plate" with sun and surrounding stars. (b) corresponding "night plate" taken of the same star field when visible at night. (c) both plates combined as they appear in the measuring machine. (From Mayo, 1991a)

other plate parameters, and finally, on the light deflection, λ itself.[17] By what is quite literally a "subtraction" method, it was possible to estimate λ.

A second important source of error stems from the fact that the predicted deflection of 1.75″ refers to the deflection of light just grazing the sun; but the researchers only observed stars whose distance from the sun is at least two times the solar radius. Here the predicted deflection is only about 1″ of arc. To compare the evidence with the theoretical prediction it is necessary to estimate what the deflection would have been for starlight near the sun.

Thus, despite the novelty of the theoretical prediction of 1.75″, to reach the hypothesis about the estimated deflection, the eclipse data themselves must be used, both to fix each of the experimental parameters and to arrive at the extrapolation to the limb of the sun. Furthermore, checking the validity of these inferences requires using, once again, the eclipse data. So the UN requirement is apparently violated. But great pains were taken to ensure that reliability or severity was not. They used only those results for which there were measurements on enough stars (at least equal to the number of unknown parameters in the equations—6) to apply a reliable method of fixing: the statistical method of least squares (regression), a technique well known to astronomers from determining stellar parallax, "for which much greater

[17] A detailed discussion of this and several other eclipse tests of Einstein's deflection is provided by H. von Klüber (1960). See also D. Moyer 1979.

accuracy is required" (Eddington [1920] 1987, 115–16) than that for the eclipse test.

(Note also that it was impossible to adhere to the classic requirement to prespecify the sample size. Before obtaining and analyzing the data one did not know how many of the photographed stars would be usable.[18] I will return to the issue of prespecification in chapter 9.)

The Results

Subtracting out the variety of factors algebraically, one arrives at estimates of λ from the different sites, along with their probable errors (or, the measure now used, their standard errors).[19] The "observed results," in short, are actually hypotheses about the expected deflection (at the limb of the sun), λ. The two eclipse results, one from Sobral, one from Príncipe, taken as crucial support for Einstein were, with their standard errors,[20]

Sobral: the eclipse deflection = 1.98″ ± 0.18″.

Príncipe: the eclipse deflection = 1.61″ ± 0.45″.

Using either standardized measure of error allows assigning probabilities to experimental results under different hypotheses about λ. This permits severity to be calculated. Eddington reasons:

> It is usual to allow a margin of safety of about twice the probable error on either side of the mean. The evidence of the Principe plates is thus just about sufficient to rule out the possibility of the "half-deflection," and the Sobral plates exclude it with practical certainty. (Eddington [1920] 1987, 118)

The severity criterion (SC), the formal analog to our argument from error, explains the weight accorded to each result. The pattern of reasoning is one with which we are by now very familiar. An observed difference from a value predicted by a hypothesis H_0 genuinely indicates that H_0 is in error, if so large a difference is very improbable (just) if the error is absent. The appraisal at stage i had several parts. In the portion of the appraisal alluded to in the above passage, H_0, the hy-

[18] This was Barnard's point in discussing the eclipse results in Barnard 1971.

[19] One probable error equals .68 standard errors. A standard error is the estimate of the standard deviation. The reason for the choice of the probable error as a standard is that a sample mean differs from a (Normal) population mean by one or more probable errors (in either direction) 50 percent of the time. (It differs from the population mean by one or more standard errors in either direction about 32 percent of the time.)

[20] The probable errors are, respectively, .12 and .30.

pothesis found to be in error, is the Newtonian "half-deflection," that $\lambda = .87''$. The hypothesis H that "passes" is

H: the half-deflection is in error, $\lambda > .87''$.

Consider passing H with the Sobral result of 1.98″ ± 0.18″. We ask: What is the probability of "such a passing result"—that is, one as far or farther from .87″ than the observed result—given that λ is the Newtonian (half-deflection) .87″? The answer is that this probability is practically 0. (The result is more than 6 standard deviations in excess of .87.) So π, in construction rule R-π, is nearly 1. The Príncipe result, being around 1.6 standard deviations in excess of .87″, is only a reasonably severe passing result. That is, with reasonably high probability, around .95, a result more in accordance with $\lambda = .87''$ would be expected, if λ were equal to the Newtonian value (.87″). ($\pi = .95$.)[21]

The probabilities, it must always be remembered, are not assigned to the hypotheses about λ. Universes are not as plenty as blackberries—to recall Peirce (from chapter 3). There is one universe, this one blackberry, within which a hypothesized value for λ either does or does not hold true. We know, however, that there are a variety of sources of error that produce differences between actual and estimated deflections. Making use of this knowledge of error we can argue as follows: were the experimental differences from the half-deflection due to the variety of known sources of error and not to a genuine discrepancy from .87″, they would practically never, or extremely rarely, be expected to occur in a series of (hypothetical) eclipse experiments at the two sites. This is our standard canonical argument for inferring that a discrepancy from a parameter value is real.

If one were filling out the hierarchy of models, one would explore how at stage i a single question is split off from the primary one. The possible hypotheses at stage i are values for λ. These are the possible answers to this one subquestion. One would describe the link between an observed mean deflection L (itself a model of the data) and hypotheses about λ within the experimental model. The severity criterion warrants accepting the use-constructed hypothesis

$H(L)$: λ exceeds $L - 2$ standard errors,

where the standard error is the estimated standard deviation of L. To see why $H(L)$ is warranted, notice that "$H(L)$ is false" asserts that λ

[21] I do not mean that this is the only work in weighing these two inferences. Detailed checks to affirm the assumptions of the experimental data models are also needed and would have to be incorporated in a full-blown discussion of the experimental inquiry.

does *not* exceed $L - 2$ standard errors. To calculate severity one calculates the probability of *not* observing a deflection as large as L, given that λ is any of the values included under "$H(L)$ is false." The value of this probability is high (at least .97).[22]

A Result in "All Too Good Agreement" with Newton

There was, however, a third result also obtained from the Sobral expedition. In contrast with the other two this third result pointed not to Einstein's prediction, but, as Eddington ([1920] 1987) declares, "with all too good agreement to the 'half-deflection,' that is to say, the Newtonian value" (p. 117). It also differed from the other two in being discounted due to systematic errors! The instrument used, an astrographic telescope, was of the same type as that used in the counted Príncipe result. Nevertheless, upon examining these Sobral astrographic plates the researchers constructed a hypothesis *not* among the three set down in advance. Because this new hypothesis incorporates the alleged exception into Einstein's hypothesis (1), we may denote it by 1*:

> 1*: The results of these (Sobral astrographic) plates are due to systematic distortion by the sun and not to the deflection of light.

Popper held up this test as a model of severity, unlike the tests of psychological theories of the day, because the Einstein prediction dared to stick its neck out: a deflection far from the predicted value and near .87", Eddington (1918) declared, "would overthrow Einstein's theory" (p. 36). So what is to be made of this discounting of one set of results from Sobral?

Certainly this violates UN. It exemplifies the second entry in our list of ways that such a violation can occur: *exception barring*, or what Worrall calls *exception incorporation*. Here, when confronted with an apparent piece of counterevidence, one constructs a new hypothesis to account for the exception while still saving the threatened hypothesis—in this case, Einstein's. Moreover, while the Einstein hypothesis can accommodate the Sobral astrographics with the help of 1*, Newton's hypothesis accommodates them without any such contrivance. According to the UN requirement, it seems, the result used to construct 1* would count more for Newton than Einstein—contrary to the actual appraisal.

Now, the proponent of UN may deny that this really counts as exception incorporation (because there is a violation of "initial condi-

[22] Note that although this includes infinitely many alternative values of λ, the high severity requirement is met for each. This instantiates my point in section 6.4.

tions"), but what cannot be denied is that constructing 1* violates UN. Still, it might be objected: the UN requirement never intended to condemn this kind of violation of UN. Here the data are being used to arrive at (and affirm) some low-level auxiliary hypothesis, one which, in this case, indicates that the data may be discounted in appraising the primary hypothesis. Are we to understand the UN theorist as allowing use-constructions in the case of low-level auxiliary hypotheses? Surely not. Otherwise the kind of UN violation (exception incorporation) that started all the fuss in the first place would pass muster. All of this underscores the main thesis of this chapter: the UN requirement fails to discriminate between problematic and unproblematic use-constructions (or double-countings).

Let us return to Eddington and the mirror hypothesis (1*). Consider the actual notes penned by Sobral researchers as reported in Dyson et al. 1920:

> May 30, 3 a.m., four of the astrographic plates were developed. . . . It was found that there had been a serious change of focus, so that, while the stars were shown, the definition was spoilt. *This change of focus can only be attributed to the unequal expansion of the mirror through the sun's heat.* . . . It seems doubtful whether much can be got from these plates. (P. 309; emphasis added)

This is not to say that it was an obvious explanation that could be seen to be warranted right off. It called for a fair amount of (initially unplanned) data analysis, and gave rise to some debate—all of which of course depended on using the suspect data themselves. However, the dispute surrounding this inference was soon settled, and because of that it sufficed for most official reports to announce that the astrographics, despite appearing to support the Newtonian value, were discounted due to distortions of the mirror. Such a report, not surprisingly, raises the antennae of philosophers looking into this historical episode.

John Earman and Clark Glymour (1980) point a finger at Eddington precisely because he "claimed the superiority of the qualitatively inferior Príncipe data, and suppressed reference to the negative Sobral results" (p. 84)—the Sobral astrographics. According to Earman and Glymour, "Dyson and Eddington, who presented the results to the scientific world, threw out a good part of the data and ignored the discrepancies" (p. 85). They question his suppression of these results because, in their view, "these sets of measurements seem of about equal weight, and it is hard to see decisive grounds for dismissing one set but not the other" (p. 75).

There were, however, good grounds for dismissing the Sobral astrographics—one could not obtain a usable standard error of the estimate. Moreover, as the journals of the period make plain, the numerous staunch Newtonian defenders would hardly have overlooked the discounting of an apparently pro-Newtonian result if they could have mustered any grounds for deeming it biased. And the reason they could not fault Eddington's "exception incorporation" in hypothesizing 1* is that it involved well-understood principles for using this type of data to test, and in this case reject, a key assumption of the experiment. Results were deemed usable for estimating the deflection λ, we said, only if the statistical method (least squares) was applicable; that is, when there was sufficiently precise knowledge of the change of focus (scale effect) between the eclipse and night plates (within .03 mm)—precisely what was *absent* from the suspect Sobral plates.[23]

The discussion of this mirror distortion hypothesis brings out an interesting feature of the analysis. A. C. D. Crommelin (*Nature,* 13 November 1919) remarked that "there is reason to suspect systematic error, owing to the very different character of the star images on the eclipse and check plates" (p. 281). This is contrasted with the Sobral results taken with the 4-inch lens, whose use was allowed:

> The small coelostat used with the 4-inch lens did not suffer from deformation, the images of stars during totality being of the same character as those on the check-plates; this increased the weight of the determination with that instrument. (Crommelin 1919, 392)

Even small systematic errors of focus are of crucial importance because the resulting scale effect (from this alteration of focus) quickly becomes as large as the Einsteinian-predicted deflection effect. After an analysis of the distortion it was evident that the Sobral astrographic results pointed to only one hypothesis: 1*, systematic error.[24]

Over and over again, the discussions reveal that what was of central importance in declaring results usable was the similarity of the pattern of error in the eclipse and check plates. This similarity, or the lack of it, was discerned statistically; and this discernment is a formal

[23] To see how important even small systematic errors of focus are, one need only look at how the resulting scale effect (from this alteration of focus) quickly becomes as large as the Einsteinian-predicted deflection effect of interest. The effect of 1.75" refers to the deflection of the light of a star just at the limb of the sun; but the researchers only observed stars whose distance from the sun is at least 2 times the solar radius. Here the predicted deflection is about 1" of arc or .015 millimeters on the photographic plate. See von Klüber 1960, 50.

[24] Eddington was a specialist in techniques of data analysis, and his notes offer, in effect, rules for legitimately using eclipse evidence to "use-construct" hypotheses about λ (that is, instances of rule R-π).

analog of the informal detection of error patterns noted in chapter 1.[25] The mirror problem of 1919 became what I have in mind by a canonical model of error, and it was used in subsequent eclipse experiments.

Whereas this first stage was relatively uncontroversial, the second stage was anything but.

Stage ii: Can Other Hypotheses Be Constructed to Explain the Observed Deflection?

While even staunch defenders of Newton felt compelled to accept that the eclipse evidence passed the hypothesis that the deflection effect $\lambda = 1.75''$, they did not blithely accept that Einstein's law of gravitation had thereby also passed a good test. As the scientist H. F. Newall put it, "I feel that the Einstein effect holds the day, but I do not yet feel that I can give up my freedom of mind in favour of another interpretation of the effects obtained" (Newall 1919, 395–96). Such skeptical challenges revolved around stage ii, determining the *cause* of the observed eclipse deflection. At issue was the possibility of a mistake about a causal factor. The question, in particular, was whether the test adequately discriminated between the effect due to the sun's gravitational field and others that might explain the eclipse effect. A "yes" answer boiled down to accepting the following hypothesis:

> (ii)(1): The observed deflection is due to gravitational effects (as given in Einstein's law), *not* to some other factor *N*.

The many Newtonian defenders adduced any number of factors to explain the eclipse effect so as to save Newton's law of gravity: Ross's lens effect, Newall's corona effect, Anderson's shadow effect, Lodge's ether effect, and several others. Their plausibility was not denied on the grounds that they were deliberately constructed to account for the evidence (while saving Newton)—as the UN requirement would suggest. On the contrary, as Harold Jeffreys wrote,

[25] Dyson, Eddington, and Davidson (1920) say this about the discounted Sobral results:

> The images were diffused and apparently out of focus. . . . Worse still, this change was temporary, for without any change in the adjustments, the instrument had returned to focus when the comparison plates were taken in July. (P. 309)

Interestingly, Crommelin explained that if we assume that the bad focus left the scale unaltered, then the value of the shift from these results is 1.54″, thereby no longer pointing to the Newtonian value.

> before the numerical agreements found are accepted as confirmations of the theory, it is necessary to consider whether there are any other causes that could produce effects of the same character and greater in magnitude than the admissible error. (Jeffreys 1919b, 138)

Were *any* other cause to exist that was capable of producing (a considerable fraction of) the deflection effect, Jeffreys stressed, that alone would be enough to invalidate the Einstein hypothesis (which asserts that *all* of the 1.74″ are due to gravity).

Everything New Is Made Old Again

The historian of science Stephen Brush (1989) found, in apparent violation of the rule to prefer (temporal) novel predictions, that the ability to explain the known fact about Mercury's orbit provided *stronger* support for Einstein's theory of gravitation than did the theory's ability to predict the new fact (in the 1920s) about the deflection of light. Getting Mercury's orbit correct counted more in favor of Einstein's theory than light bending did, not despite the fact that the former was known and the latter new, but because of that very fact. Severity considerations explain why. The known fact about Mercury—being an anomaly for Newton—was sufficiently important to have led many to propose and test Newtonian explanations. These proposed hypotheses, however, failed to pass reliable tests. In contrast, when light bending first became known to exist "one might expect that another equally or more satisfactory explanation would be found" (Brush 1989, 1126). It is as if before this novel effect could count as an impressive success for Einstein's theory, scientists had to render it old and unsatisfactorily explained by alternative accounts (much like Mercury). I think Brush is right on the money in declaring that

> the eclipse results . . . provoked other scientists to try to give plausible alternative explanations. But *light bending could not become reliable evidence for Einstein's theory until those alternatives failed, and then its weight was independent of the history of its discovery.* (Brush 1989, 1127; emphasis added)

Let us now consider how the new light-bending effect was made appropriately old.

Using the Eclipse Results at Stage ii

The challenges at stage ii to the pro-Einstein interpretation of the observed deflection were conjectures that the effect was due to some factor other than the Einstein one (gravity in the sun's field). They were hypotheses of the form

> (ii)(2): The observed deflection is due to factor *N*, other than gravitational effects of the sun,

where *N* is a factor that at the same time saved the Newtonian law from refutation. Each such hypothesis was criticized in a two-pronged attack: the effect of the conjectured *N*-factor is too small to account for the eclipse effect; and were it large enough to account for the eclipse effect, it would have other false or contradictory implications.

Stage ii exemplifies several UN violations: the road to hypothesis construction was constrained to account for evidence *e*, and *e* also counted in support of that hypothesis. Once the deflection effect was affirmed at stage i it *had* to be a constraint on hypothesizing its cause at stage ii; at the same time, the eclipse results had to be used a second time in appraising these hypotheses. (A similar eclipse would not occur for many years.) Typically they were used to fix a parameter, the extent to which a hypothesized factor *N* could have been responsible for the observed deflection effect. When explicitly used to save the Newtonian law, they also violated UN by exception incorporation. Note also that the alternative hypotheses were at the same level as the primary hypothesis here.

The arguments and counterarguments (scattered through the relevant journals from 1919 to around 1921) on both sides involved violating UN. What made the debate possible, and finally resolvable, was that all who entered the debate were held to shared standards for reliable experimental arguments. They were held to shared criteria for acceptable and unacceptable use-constructions. It was acceptable to use any evidence to construct and test a hypothesis *H* (about the deflection effect) so long as it could be shown that the argument procedure was reliable or severe—that it would very rarely yield so favorable a result erroneously. Examples abound in the literature. They supply a useful sampling of canonical arguments for ruling out hypothesized causes of an effect of this sort. I will briefly cite a few.

The shadow effect. Alexander Anderson (1919, 1920) argued that the light deflection could be the result of the cooling effect of the moon's shadow. Eddington responded that were the deflection due to this shadow effect there would have had to be a much larger drop in temperature than was actually observed. (It might have been responsible for the high value of the deflection found at Sobral.) Anderson did not give up, but attempted other hypotheses about how the moon's shadow could adjust conditions just enough to explain the effect and save the Newtonian law. These attempts were found wanting, but only after being seriously considered by several scientists (e.g., by Arthur Schuster [1920]). The problem, in each case, was not that Anderson repeatedly use-constructed his hypotheses, but that in so doing he was forced into classically unreliable arguments. The problem, well put by Donald Moyer (1979), was this:

> The available adjustments are adjustments of parameters of trustworthy laws and these adjustments are tightly constrained by the connections among these laws of phenomena. Temperatures, or air currents, or density gradients cannot be adjusted in one law without also adjusting all the other laws where these terms occur as well and this must not introduce consequences not observed. (P. 84)

A test procedure that relies on inconsistent parameter adjustments to get a hypothesis to pass would frequently pass hypotheses erroneously. The test is highly unreliable.

Newall's corona lens. Another *N*-factor seriously entertained was put forward by H. F. Newall (1919, 1920), that of the intervention of a corona lens. Again, there was a two-fold response, here by the scientist F. A. Lindemann and others. The refraction required to cause the eclipse result, Lindemann (1919) argued, would require an amount of matter many orders of magnitude higher than is consistent with the corona's brightness, and were there enough matter to have caused it, comets passing through the region should have burned up.

Ether modifications. Sir Oliver Lodge (e.g., Lodge 1919) promised that if the Einstein effect was obtained he would save Newton by modifying conditions of the ether with special mechanical and electrical properties; after the results were in, he did just that. (Lodge, a proponent of spiritualism, held that the ether effected contact with departed souls, in particular his son, Raymond.) Strictly speaking, since these hypotheses were constructed by Lodge before the results, it seems that the case satisfies temporal novelty, and so use-novelty. This hardly made Lodge's arguments more impressive. The problem was not *when* Lodge formulated his hypotheses, but that his procedure for passing them required inconsistent parameter adjustments. Consistent adjustments showed that each hypothesized factor *N* could not have caused the observed deflection. As Lindemann (1919) put it:

> Sir Oliver Lodge has suggested that the deflection of light might be explained by assuming a change in the effective dielectric constant near a gravitating body. This way of looking at it had occurred to me. . . . It sounds quite promising at first since it explains . . . the shift of the perihelion of Mercury as well as the . . . shift of the spectral

> lines, if this exists. *The difficulty is that one has in each case to adopt a different constant in the law,* giving the dielectric constant as a function of the gravitational field, *unless some other effect intervenes.* (P. 114; emphasis added)

The kinds of tactics Lodge employed lead many to insist on the UN requirement. Far from striving to steer clear of classic unreliable use-construction procedures, he employed (whether deliberately or not) precisely the kind of rigging that would allow hypotheses to pass, whether or not they were true.

Not that one can see immediately which use-constructions are kosher and which are not—even the constructors themselves cannot do this. This is because one cannot see immediately which ones have arguably passed severe tests. By indiscriminately prohibiting all tests that violate UN, the UN requirement cannot provide an epistemological ground for the reasoning in this dispute, nor for the way it was settled.

What finally settled the matter (around 1921) was not the prediction of novel evidence, but the extent to which known evidence warranted only a construction of the Einstein gravitational hypothesis. This was argued by Harold Jeffreys (1919a, 1919b) (despite his having initially assigned an extremely low Bayesian prior probability to Einstein's law). Jeffreys—one of the last holdouts—explains:

> It so happens that the three known facts, the truth of Kepler's third law, the motion of the perihelion of Mercury, and the displacement of star images, give different equations for the constants, and *the only solution that satisfies those three conditions happens to be Einstein's theory.* . . . It must be accepted as the only theory that will satisfactorily coordinate these facts. (Jeffreys 1919a, 116; emphasis added)

What he is saying is that in order to use the known results (the eclipse effect together with Kepler's law and the Mercury perihelion) to construct a hypothesis, and do so reliably, one is led to Einstein's law of gravity! After reviewing the tests and all the rival explanations, Dyson and Crommelin concluded in the February 1921 issue of *Nature,*[26] which was entirely devoted to the eclipse tests: "Hence we seem to be driven by exhaustion to the Einstein law as the only satisfactory explanation" (p. 788).

What about other alternative hypotheses that may be dreamt up that will not disagree with H on any experimental results (either of a given test or of any conceivable test)? What about, say, possible alternative conceptions of space and time that would agree experimentally

[26] Dyson and Crommelin, *Nature* 106 (1920–21): 781–820.

with Einstein's law? We already took up this issue in discussing underdetermination (chapter 6). It is readily admitted that the 1919 eclipse tests were not severe tests of these alternative conceptions. The eclipse tests were not even considered tests of Einstein's full theory. As Eddington remarked:

> When a result that has been forecasted is obtained, we naturally ask what part of the theory exactly does it confirm. In this case it is Einstein's *law* of gravitation. (Eddington 1919, 398).[27]

It is important to stress, however, that the existence (or logical possibility) of alternative hypotheses that are not themselves tested by a given experiment leaves unaltered the assessment of hypotheses that *are* severely tested. The severity calculation is unchanged. On the informal side this means that we can learn things one piece at a time and do not have to test everything at once. That is why Jeffreys (and others) could laud the eclipse results as finally putting the Einstein law on firm experimental footing, apart from any metaphysical concepts (e.g., about space and time). (See, for example, Jeffreys 1919b, 146.) However Einstein's full theory is modified, the knowledge gained in the severely tested experimental law remains:

> In this form the [Einstein] law appears to be firmly based on experiment, and the revision or even the complete abandonment of the general ideas of Einstein's theory would scarcely affect it. (Eddington [1920] 1987, 126)[28]

Summary and Next Step

Our critique of use-novelty showed it to be neither necessary nor sufficient for severity. This finding discredits the rule of novelty (RN) when viewed as a policy always to be followed to satisfy severity. It also teaches us how UN may be violated yet avoid a possible threat to severity, namely, with a reliable rule for use-constructing hypotheses such as rule R-π. Proponents of UN err by taking a handful of cases in which UN is violated and where the test lacks severity, and then generalizing to eschew all violations of UN.

There are, however, contexts of inquiry where the methodological rules that have been developed to ensure reliability are invalidated

[27] One reason for this is that the redshift prediction had thus far not passed a severe test.

[28] It is interesting to consider in this connection the recent progress in partitioning theories of gravity and determining which theories are consistent with given experimental results, as reported in Earman 1992. See section 6.3.

when UN is violated. Their reliability guarantees break down when use-constructing is allowed. These contexts comprise an important subset of standard Neyman-Pearson tests. In other contexts, however, Neyman-Pearson methods seem happy to violate rules against use-constructing. For a long time this has caused a good deal of confusion. Let us see if we cannot dispel it once and for all.