Novel Evidence and Severe Tests

Author(s): Deborah G. Mayo

Source: *Philosophy of Science*, Dec., 1991, Vol. 58, No. 4 (Dec., 1991), pp. 523-552

Published by: The University of Chicago Press on behalf of the Philosophy of Science Association

Stable URL: https://www.jstor.org/stable/188479

Philosophy of Science

December, 1991

NOVEL EVIDENCE AND SEVERE TESTS*

DEBORAH G. MAYO†‡

Department of Philosophy
Virginia Polytechnic Institute and State University

While many philosophers of science have accorded special evidential significance to tests whose results are "novel facts", there continues to be disagreement over both the definition of novelty and why it should matter. The view of novelty favored by Giere, Lakatos, Worrall and many others is that of *use-novelty:* An accordance between evidence e and hypothesis h provides a genuine test of h only if e is not used in h's construction. I argue that what lies behind the intuition that novelty matters is the deeper intuition that *severe* tests matter. I set out a criterion of severity akin to the notion of a test's power in Neyman-Pearson statistics. I argue that tests which are use-novel may fail to be severe, and tests that are severe may fail to be use-novel. I discuss the 1919 eclipse data as a severe test of Einstein's law of gravity.

1. Introduction. Since the seventeenth century or earlier,[1] a number of philosophies of theory appraisal have accorded particular weight to "novel

*Received January 1989; revised March 1990.

†The research for this paper took place during tenure of a National Endowment for the Humanities Summer Stipend Fellowship and a National Endowment for the Humanities Fellowship for College Teachers. I gratefully acknowledge that support. Numerous stimulating discussions with Alan Musgrave were integral to formulating the ideas in this paper. I owe special thanks to George Barnard for insights into the statistical analysis of the 1919 eclipse data. For their valuable comments and criticism, I thank Dick Burian, Stephen Brush, Ronald Giere, Marjorie Grene, Larry Laudan, Ronald Laymon, Isaac Levi, Harlan Miller, Karl Popper, J. D. Trout, and John Worrall. Revisions were carried out while visiting the Center for Philosophy of Science at the University of Pittsburgh. I am grateful for the Center's stimulating environment, and for immensely helpful discussions on this paper with John Earman, Clark Glymour, Adolf Grünbaum, and Wesley Salmon. Versions of this paper have been given at the University of Hawaii, University of Maryland, and University of Pittsburgh.

‡Send reprint requests to the author, Department of Philosophy, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

[1]An excellent discussion of historical references to this idea occurs in Musgrave (1974).

Philosophy of Science, 58 (1991) pp. 523–552.

tests" of hypotheses, that is, tests whose results are "novel" in some sense. The basis of this intuition is that before a theory is supported it should have passed a genuine test, and that even if a hypothesis h perfectly fits evidence e something more is required to deem e a genuine test of h. Even those who share this intuition, however, continue to quarrel over the very definition of "novel test" or "novel fact".

At first "novel test" meant what is said: A novel test was a new test and a novel fact a *newly discovered* fact—one not known before used in testing (e.g., Popper, Lakatos). But many novelists became convinced that this strictly *temporal view of novelty* would not do. It denied special evidential significance to tests that intuitively seemed to possess it. For example, the motion of the perihelion of Mercury was known long before Einstein's theory was proposed, yet was considered to have provided significant evidence for it. In response to such objections to temporal novelty, novel accounts of novelty were proposed which turned on a fact's *heuristic* role: on whether the hypothesis it helped construct avoided being ad hoc in some sense. Zahar (1973) suggested that a fact is novel for a hypothesis if it "did not belong to the problem–situation which governed the construction of the hypothesis" (p. 103), pointing out that old facts (i.e., facts not temporally novel) could be novel facts in this new sense. Musgrave (1974) and others criticized this view as being too subjective and psychologistic; 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. In response to this criticism Worrall reformulated 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. Worrall's conception of *use-novelty* 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 must itself *not have been used* in h's construction. 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 novelty program. For this it suffices to restrict my discussion to Worrall's use-novelty criterion, which I will abbreviate as **UN**, namely, the condition that

(**UN**) (i) h entails e,
(ii) e is not used in the construction of h

since it, or something very much like it, is endorsed—at least as a necessary condition—by others who endorse use-novelty as well as by those who endorse temporal novelty.[2]

[2]In specifying **UN** I am deliberately keeping h and e vague: I see no other way to pick up the discussion as found in the novelty literature without prejudging the very issues involved. Here I follow Worrall in requiring empirical results to be entailed by the relevant hypotheses passed, and in allowing both h and e to range from highly specific to highly

Focusing on Worrall is useful because his account seems appropriately sensitive to the three main desiderata by which accounts of novelty are judged in the literature. First, the criterion of novelty supplied should be reasonably unambiguous: It should not make novelty turn on the subjective aims or intentions of scientists. Second, the criterion of novelty should yield an appraisal of tests that accords with actual scientific appraisals. Third, the criterion should have an underlying epistemological rationale.

As to the first concern, Worrall (1985) admits that "[a]llowing that heuristics play a role does indeed threaten to make confirmation a dangerously unclear and subjectivist notion" (p. 309). After all, "what exactly does it mean for an empirical result to be used in the construction of a theory?" (ibid., 310–311). In order to eradicate this difficulty, Worrall attempts to spell out some clear-cut cases of what might be termed "use-constructions", allegedly identifiable by objective historical means. Throughout, I keep to these relatively clear-cut violations of **UN**: "parameter adjustment" and "exception incorporation". The second concern, "fitting the facts" (of actual science) does not by itself provide a very severe test (of theories of severe testing). Theories of novelty agree that merely fitting the facts is too weak for a good test, so too should it be deemed too weak to test these theories themselves (on the metalevel). For example, most scientific cases are equally accommodated by (and hence fail to discriminate between) temporal and use-novelty, unsurprising since temporal novelty is sufficient, though not necessary, for use-novelty. Still the temporal novelist, claims Worrall, has merely found a correlation between scientific appraisals and the temporal novelty of the data. So, although Worrall offers historical cases in support of his account, he appears to agree that the third consideration—the account's epistemological rationale—is more hard-hitting (see, for example, Worrall 1978a,b; esp. 326). Even if intuitively plausible scientific theory appraisals were to accord with a rule, "require such-and-such criterion of novelty", it would not follow that scientists were in fact using such a rule, nor that the rule reflected any epistemological objectives. The focus of my critique will be this third concern. I will be asking: What is the epistemological rationale of use-novelty as given in criterion **UN**?

One important advantage Worrall claims use-novelty has over temporal novelty

---

general claims although in general I find both troublesome. In this paper I also keep to Worrall's framework where an account of tests, or, as he prefers, of support is to provide a necessary and/or sufficient condition for a good or severe test. My own view would extend this to degrees of severity, and in so doing avoid the overly stringent condition (i), that h entails e. Instead, the accordance between h and e would be a statistical one.

> . . . is that it comes equipped with a rationale. 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? (1989, 148)[3]

The rationale with which use-novelty "comes equipped" according to Worrall is this: Satisfying the criterion of use-novelty **UN** (avoiding what I call use-constructed hypotheses) furthers the aim of guaranteeing genuine and avoiding spurious tests. Worrall likens his intuitions about genuine tests to Popper's; and he intends his use-novelty criterion **UN** to capture Popper's requirement that before a theory is supported it should have passed a genuine or *severe test.*

I begin in section 2 to explicate a criterion of severity that I believe underlies the Popperian intuition which Worrall views his account as embodying—although it is not equivalent to any of Popper's formal measures of severity. It is akin to the notion of a test's *power* in Neyman-Pearson statistics. I go on to argue that tests which are use-novel may fail to be severe, and tests that are severe may fail to be use-novel, in sections 3 and 4, respectively. What lies behind the intuition that novelty matters, I suggest, is the deeper intuition that *severe* tests matter; novelty is thought to be important because or to the extent that it corresponds to severity. (The aim of severity, I would also argue, underlies the general intuition to prefer non-ad hoc hypotheses—but here I limit my discussion to use-novelty.) This critique yields an account of severe tests which I believe captures the intuitions underlying the novelty requirement *where those intuitions are correct.* In section 5 I consider the severity of the 1919 eclipse tests of Einstein's law of gravity. In section 6, I reply to two anticipated questions and finally, I suggest some future work.

2. Severity as the Rationale for Use-Novelty: A Severity Criterion. For Worrall, perhaps the clearest case of using facts to construct hypotheses is in *parameter adjustment.* So let us consider the reasoning underlying Worrall's intuition that such cases fail to provide genuine tests.

While he does not give a precise definition for parameter adjustment the general idea is this: ". . . a theory is first proposed which has some free parameters—these may be parameters in the usual specific sense

[3]Puzzlingly, Worrall (1978a) not only claim that "Those theories which receive genuine empirical support on this [temporal novelty] view have contributed to one of the aims of science" (p. 327), but took this to show that temporal novelty was *better* equipped with an epistemological rationale (while less good at fitting the facts of science) than the heuristic view he himself favored. For the most he saw to favor use-novel over use-constructed theories is that "the former could, while the latter could not, have contributed to the achievment of one of the aims of science—the extension of our factual knowledge" (ibid., 328).

(constants in some mathematical equation) or in a more general vaguer sense" (1985, 312). Experimental results are then accommodated by suitably fixing the values of these parameters. Worrall asks:

> But what of the theory with all the free parameters filled in? It certainly is testable in the logical sense—it has consequences which are directly comparable with, and indeed which compare favorably with, experiment. But this is no wonder since these consequences were all 'written into' the theory. . . . (Ibid., 313)

And again in Worrall (1989) he reasons that

> [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. (P. 149)

From this he concludes that "the intuitions behind the notion of a genuine test cannot be captured in purely logical terms but must involve consideration of how the theory concerned was constructed" (ibid.), that is, with whether it was use-novel. But what are these test intuitions?

Worrall (1985) claims that "[m]any of Popper's most perspicacious remarks are . . . based on an *intuitive* notion of testability" (p. 313) embodied in the heuristic account (i.e., in Worrall's account), although this, according to Worrall, "Popper has never, I think, fully and clearly realised . . ." (ibid.). I want now to examine some of Popper's "perspicacious remarks" in order to identify what Popper's intuition about genuine tests shares with Worrall's. To criticize Worrall's criterion on its own terms it must be clear what he intends it to accomplish and to avoid—at least at the very minimum.

It is worth going back briefly to Popper's familiar story of how his ideas about severe tests arose with respect to the theories evoking excitement in 1919: in particular, those of Freudian and Adlerian psychology, and Einstein's relativity theory:

> Once, in 1919, I reported to him [Adler] a case which to me did not seem particularly Adlerian, but which he found no difficulty in analysing in terms of his theory of inferiority feelings. . . . What, I asked myself, did it confirm? No more than that a *case could be interpreted* in the light of the theory. (1962, 35; emphasis added)

Popper describes how two opposed examples of human behavior (deliberately drowning a child versus sacrificing one's life to save a child) can be explained equally well by both theories (Adler's and Freud's), say, by feelings of inferiority and repression, respectively. Such flexibility in

interpreting behavior allowed Adler's theory to pass the test easily, even if it was wrong about the behavior's cause.

Popper contrasts this case with the planned test of Einstein's theory by checking 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. This is quite different from the situation I have previously described, when it turned out that the theories in question were compatible with the most divergent human behaviour. . . . (Ibid., 36)

The contrast Popper successfully draws is not between the scientific status of whole domains of inquiry—psychology being less scientific than physics[4]—but, rather, between genuine and spurious tests, or between genuine and spurious support. This distinction is identified in such passages as the following:

> [M]ere supporting instances are as a rule too cheap to be worth having; they can always be had for the asking; thus they cannot carry any weight; and any support capable of carrying weight can only rest upon ingenious tests, undertaken with the aim of refuting our hypothesis, *if it can be refuted*. (Popper 1983, 130; emphasis added)

> The theoretician will therefore try his best to detect any false theory . . . he will try to 'catch' it. That is, he will . . . try to think of cases or situations in which it is likely to fail, if it is false. Thus he will try to construct *severe* tests, and *crucial* test situations. (Popper 1979, 14)

> [We] try to select for our tests those *crucial cases* in which we should expect the theory to fail *if it is not true*. (Popper 1962, 112; emphasis added)

To formulate the pivotal requirement of severe tests, it is sufficient to consider the test outputs—"h passes" or "h fails"—a test T with outcome e. If one prefers, assertions that a hypothesis h is true [false] may be interpreted as assertions that h is [is not] reliable for an application, or rationally acceptable, or otherwise "merited". The severity requirement is this:

[4]As Grünbaum (e.g., 1979, 1989) shows, Popper's falsifiability criterion does not demarcate physics as *intrinsically* more scientific than psychology. Hypotheses in physics could be protected from falsification, and psychological hypotheses are open to being falsified.

> *Severity Requirement:* Passing a test T (with e) counts as a good test of (or good evidence) for h just in case T is a *severe test* of h[5]

and the criterion of severity **SC** I suggest is this:

> (**SC**): *Severity Criterion*: There is a very high probability that test T would not yield such a passing result, if h is false[6]

where "such a passing result" means one that accords at least as well with h as does e. If results that accord less well with h than does e lead to *failing* h, then our severity requirement asserts: A passing result is a good test of h just in case h is very likely to have failed the test T, if h were false (i.e., if h "merits" failing). What **SC** is requiring is that there be a high probability that the test does not pass hypotheses erroneously. Probability is understood as relative frequency in a (real or hypothetical) series of test applications.

Some may wonder how the probability in **SC** is to be obtained. How to apply **SC** will become clearer as we continue, and will be addressed directly in section 6 (question 1). For now this can be left aside, as my critique of use-novelty is based on extreme cases of violating or satisfying severity where the probabilities of h not passing when false are 0 and 1 respectively. I begin with the first extreme case, that of a *minimally severe* or a *0-severity* test.

Popper's criticism of Adlerian and Freudian tests—whether or not it can be sustained—is relevant here. Popper's charge, in effect, is that these tests are guilty of an extreme violation of **SC** because "it was *practically impossible* to describe any human behavior that might *not* be claimed to be a verification of these theories" (1962, 36; emphasis added). This identifies a test where h has minimal (0 or close to 0) probability of *not* passing even if h is false, that is, a test which h must pass, even if false. Thus, I define:

> *Passing a Minimally Severe (0-severity) Test*: h passes a 0-severity test with e iff there is no chance for the test *not* to yield such a passing result, even if h is false.

The rationale for the severity requirement in terms of **SC** seems clear: If h had little or no chance of not passing test T, even if false, then h's passing *does not provide a reason* for accepting or supporting h; it fails utterly to discriminate h being true from h being false. This fits perfectly

[5]Although Worrall's framework ignores degrees, they can be introduced by replacing "just in case" with "just to the extent that". These degrees must be carefully interpreted, however. (I attempt this elsewhere.)

[6]Equivalently, "There is a very *low* probability that test T yields such a passing result if h is false". Outcome e is always understood by me as a generic, and not as a specific, occurrence.

with our intuitions about whether passing marks on an exam warrants merit of some sort. Consider a test to determine if a student can recite all the U.S. state capitals; say the hypothesis h is that the subject can correctly recite (aloud) all 50. Suppose the test passes a student so long as she can correctly assert the capital of any one state. That a person passes this test is not much of a reason to accept h because it is not a very severe test in my sense. Suppose now that the test passes a student so long as she can recite *anything* aloud. Granted being able to recite all 50 capitals entails being able to speak aloud (h entails e), but this test is even less severe than the first. It is even easier (more probable) for a pass to occur, even if the student is *not* able to recite them all (h is false), (see Popper 1979, 354 for an exam analogy used to make the same point).

Alternatively, if a student passes a test where passing requires reciting all 50 capitals correctly, certainly that is excellent support for hypothesis h, that the student can correctly recite them all. This identifies the other extreme, that of a maximally severe test:

> *Passing a Maximally Severe (100%-severity) Test*: h passes a maximally severe test with e iff there is no chance that the test yields such a passing result, if h is false.

While I take my severity requirement to capture the Popperian intuition in the passages cited above, Popper articulated differing views of severity. Most importantly, my probabilistic severity measure is not equivalent to any of Popper's formal measures. Rather, it is akin to the measure of a test's *power* in Neyman and Pearson statistics. The power of a statistical test of a hypothesis h is the probability that the test rejects (or does not pass) h when h is false (and so *should* not pass). Requiring severity is similar to requiring that a test have high power to detect the falsity of h, a point Ronald Giere has made.[7] (A main difference is that, unlike power, I take the specific outcome e as the cutoff for getting the probabilities in **SC**.[8] We need not pursue these technical points here.) Severity must not be equated with the probability of not getting a particular passing result e, nor does it equal $P(\sim e \mid h \text{ is false})$. *A test's severity is one minus the*

[7]Giere (1983) also stresses that what matters is not novelty, but severity in this sense. However, Giere takes violating use-novelty to preclude severity: Because Fresnel "was unwilling to consider any model that did not yield the right pattern for straight edges", Giere reasons, "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" (ibid., 282). As I argue in section 4, violating use-novelty does not entail the low or near 0-severity that Giere alleges here and elsewhere (e.g., Giere 1984).

[8]Power is a measure over the entire set of outcomes that would be taken to reject h; severity is a function of outcomes as or more discordant from h than is outcome e. See Mayo 1985 and 1988.

probability it yields some such passing result (or other), given the hypothesis passed is false. (See sections 4 and 6, question 2.)

It is straightforward that satisfying criterion **SC** accords with the aim of avoiding 0-severity tests and securing maximally severe tests. Satisfying use-novelty criterion **UN**, I will argue, does not: Tests can be use-novel and yet minimally severe, and they can violate **UN** and yet be maximally severe.

3. Use-Novelty is not Sufficient for Severity. In a 0-severity test of h, passing is ensured whether or not h is true. Hence, a 0-severity test of h is a spurious test, at least according to the intuitions of one who endorses use-novelty. But why suppose that satisfying **UN** *precludes* 0-severity tests? For use-novelty to be sufficient for a good test, it would have to ensure at least this much. Worrall does deem **UN** sufficient, claiming:

> In all cases where scientists have *not* recognized evidence e as fully supporting theory T, despite e's following from T, T had been modified or tinkered with or otherwise developed precisely so as to yield e. (1989, 148)

Thus, showing that satisfying **UN** does not preclude a 0-severity test shows that **UN** fails to be sufficient. I will sketch some key ways of generating tests with 0-severity, all of which are nevertheless able to satisfy **UN**.

(a) Biased Interpretation of Evidence: Use-novelty alone fails to rule out the very thing that first opened Popper's eyes as to why finding any and all facts in accord with a hypothesis, far from being desirable, "cannot carry any weight". By suitably *interpreting* the evidence one may make it accord with a hypothesis h so as to yield a "test" that h must pass even if h is false (0-severity). But this has nothing to do with use-novelty. One is not guilty of using data to construct the hypothesis, but of using the hypothesis to *interpret* the data so as to be favorable to h. This was the allegation in the Adler case. For a statistical case, consider observing the incidence of cancer in a sample exposed to some substance to test h: The substance does not increase the risk of cancer. By suitably raising the burden of proof required to reject h, any result may be interpreted as passing h (even if h is false).

(b) Insensitive Tests: A second way a test may have very low or even 0-severity while still satisfying requirement **UN** is this: Use a known or highly probable fact logically entailed by h as grounds for passing h (remember **UN** does not require e to be temporally novel). An example is the one mentioned in first defining low severity tests. Here, one passes the hypothesis h: The subject can recite all the capitals aloud on the basis

of the evidence e that the subject can talk. Hypothesis h entails e, but the test hardly probes the subject's knowledge of capitals. For much the same reason, as Musgrave (1978) said of earlier views of heuristic novelty, "the heuristic view seems to land us squarely back into the Raven Paradox" (p. 197, n. 17). The same insensitivity arises in nonartificial cases.

Consider the evidence that planets revolve around the sun. Although relativity theory entails this known fact, and it was not used in relativity theory's construction, it does not afford a good test of that theory. It is a terribly ineffective way to proceed to uncover flaws in that theory. Were satisfying **UN** sufficient, there would have been no need to wait for the 1919 eclipse to test Einstein's gravitational hypothesis; it entailed measurements on light deflection on terrestrial bodies, but they would fail to discriminate between rival predictions, for example, Newton's. Insensitive tests are a common problem in statistically testing the hypothesis h of no-increased cancer risk, given in (a) above. Suppose the result e is no observed increase in cancer incidence, so h accords with e. However, sample sizes are often so small (relative to risk) that such a passing result is highly probable even when an increased cancer risk exists (h false). Severity is violated, but use-novelty is satisfied.

(c) Biased Selection of Data: Satisfying **UN** does not block another classical way of getting too-easy confirmations: deliberately selecting, as a basis for testing hypothesis h, only data that accord with h. An example would be to make a number of predictions about events to occur in 1991 and then using only those events that come to pass to "test" a hypothesis about my clairvoyant abilities. (Whether the hypothesis is that all of my predictions are correct, or just that some proportion is, this fallacy may be generated.) Yet the hypothesis is use-novel, indeed, it is temporally novel. The problem is that, with predictions judiciously chosen, the hypothesis is guaranteed an overwhelmingly high chance of passing even if false (even if I have no clairvoyant abilities).

A less obviously fallacious case occurs in statistically testing a hypothesis h asserting the existence of a genuine correlation in some population. One "hunts with a shotgun" through sample correlations for those that are statistically significant at a prespecified level (e.g., .05), and then uses any such statistically significant sample correlation as grounds for passing hypothesis h (h is temporally novel). One need not hunt for long to make it highly probable to find such a significant sample correlation—even if h is false (the correlation is spurious). Because this would make the severity of the test low, the inference to h would not be warranted according to criterion **SC**. However, I see no violation of **UN** in this use of evidence in support of (prespecified) h.

In a telling passage, Popper (1974) explains that "Carnap thought that [his requirement of] 'total evidence' . . . would rule out situations where favourable or nonfavourable evidence had been carefully selected in order to obtain just the desired value" (p. 1080), and what Popper "hoped to point out" (ibid.) by stressing severity is that Carnap's total evidence is insufficient:

> . . . that we need only shut our eyes at appropriate moments (when we fear the evidence may be unhelpful) to make this "total evidence" arbitrary and biased. (Ibid.)

My point is that use-novelty does not safeguard against such bias either.

In each of these cases the trouble arises not because the data are used to construct the hypotheses tested, but because the tests allow maximal (or near maximal) violations of severity. Since criterion **UN** is claimed to be sufficient to avoid (at least) these most flagrant cases of spurious tests, these cases are counterexamples to that thesis. Nor are these types of counterexamples meant to be exhaustive. In defense of the sufficiency of **UN**, one could maintain that **UN** is also violated in these cases of 0- or low severity tests. This would require reconstructing my examples so that the hypotheses being affirmed were not the ones I give, but other ones that do violate **UN**. But why struggle to construe intuitively poor tests as violating use-novelty when the real issue underlying these intuitions is severity? (Doing so would seem to be contrary to the very spirit of use-novelty!)

4. Use-Novelty is not Necessary For Severity. Whatever problems criterion **UN** runs into as a sufficient condition, all heuristic or use-novelists, as well as many philosophers of science, hold **UN**, or close versions of it,[9] as (obviously) necessary for a good test. Use-novelty is also necessary for temporal novelty since violating use-novelty entails violating temporal novelty. I will now argue that **UN** fails to be a necessary condition as well. This argument will uncover, I believe, a flaw in the pivotal intuition shared by both use- and temporal novelists:

> And the basic intuition of the heuristic view becomes plain enough: if a theory has been deduced from some phenomena, then those phenomena cannot also support it. (Musgrave 1989, 28)

The basic intuition might be called the **UN** requirement:

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

[9]For example, Musgrave (1989) holds that a fact does not support a hypothesis if "it figured in the premises" (p. 28) from which the hypothesis was deduced.

To violate the **UN** requirement just seems like cheating. It is not that we think such use-constructed hypotheses are false or improbable. It simply does not seem that such a passing result should be credited to h. Worrall (1985) asserts the necessity of **UN** to be "almost self-evident":

> If the theory was adjusted so as to yield a certain result, then its yielding that result tells us something only about the ingenuity of man; it tells us nothing about the likelihood that the theory reflects some part of the blueprint of the universe, or even about its 'rational acceptability'. (P. 323)

Now, it is true that many of the tactics that yield low or 0-severity tests are often easier to implement if violations of **UN** are allowed. However, **UN** may be violated even where severity is most strongly and clearly satisfied, namely, with *maximally severe* tests. Accounts that hold **UN** to be always necessary open themselves to easy counterexamples.

For a trivial, but instructive example of "parameter adjustment",[10] consider a hypothesis 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 $h(e)$. Supposing the result is a mean score of 1121, it yields:

$h(e)$: The average SAT score = 1121.

Surely the data on my students are excellent evidence for my hypothesis about their average SAT scores. It seems ridiculous to suppose further tests would give better evidence. For $h(e)$ follows deductively from e. There is no way for this test to pass the hypothesis it constructs erroneously; it is a maximally severe test. So maximal severity, as I define it, fits these intuitions about maximal evidence.

Nevertheless, on the **UN** requirement, "the results which were used to fix the parameter values provide no such support" (1985, 313) for hy-

[10]Since setting out this example I have discovered, in Glymour, Scheines, Spirtes, and Kelly (1987), Howson (1984), Howson and Urbach (1989) and Nickles (1987), similar examples of counting and of standard (Neyman-Pearson) estimation being put forward as counterexamples to the necessity of use-novelty. However, my grounds for considering these counterexamples to the **UN** requirement are very different from those of each of these authors. In addition, my view of what counts as *sufficient* for a good test differs markedly from these philosophers, with the exception of Glymour et al. See my reply to question 2 in section 6 for contrasts between requiring severity and Bayesian support.

pothesis $h(e)$.[11] For $h(e)$ "was, because of its method of construction, never at any risk from the facts described by e" (Worrall 1989, 149). At first glance such a no-risk test might seem also to violate my severity criterion **SC**. But, it does not: It cannot be a problem (for severity) that a test could not have done other than pass hypothesis h, if such a passing result is impossible (or even highly improbable) were h not true. That using e to construct and appraise h appears to violate **SC** by allowing a 0-severity test, I believe, explains why the use-novelty intuition might seem obviously correct. That is, the apparent necessity of **UN**, I claim, is due to supposing that a test that violates **UN** is a test with 0-severity. This supposition is false. The aim of severity, as Popper (1979) stresses, is: "to think of cases or situations in which it is likely to fail, *if it is false*" (p. 14; emphasis added). My point is that the clause "if it is false" is crucial, but is missing in the injunction against violating **UN**.

To bring this out more clearly, 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 hypothesis is false*.

Note that here two things may vary: the hypothesis tested as well as the value of e. Now consider a rule for violating **UN** via parameter-fixing. Let a hypothesis constructed by using e to fix its parameter(s) so that h entails e be written $h(e)$ (it may be read "h fixed to entail e"). (This can be made more general by replacing "entails" with a statistical measure of fit.) The following gives a rule for constructing a test that violates **UN**:

To Construct a Test T that Violates **UN** (*by parameter-fixing*): Use e to construct $h(e)$. Let $h(e)$ pass the test iff $h(e)$ entails e.

By definition, $h(e)$ entails e. Thus, there is no chance for a test T that violates **UN** to fail the $h(e)$ it constructs. That is:

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

equals 1. But that is different from asserting that the test T is guaranteed to pass a hypothesis $h(e)$, *even if that hypothesis is false*—i.e., from asserting T ensures that the probability in (B) equals 1. And, since 1 −

[11]Or he may make the weaker claim that $h(e)$ is less well supported than a hypothesis that did not violate **UN**. But even this weaker claim will not hold up. Any hypothesis reached without using the data here would not be better supported than the one reached using the data on my students.

(B) equals severity, **SC** is violated only if the probability in (B) is high (maximally, 1).

Those who hold the **UN** requirement overlook the possibility that h might be constructed to entail e in such a way that the probability in (B) is kept low, and hence the test's severity is kept high. 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 e would not have resulted from the experiment unless $h(e)$ were true (or approximately true).

To calculate the probabilities in (A) and (B) when h is use-constructed requires taking into account the construction rule employed—in this case rule R-1. (A) becomes:

> (A:R-1): The probability that test T passes the hypothesis $h(e)$ constructed by rule R-1.

The value of the probability in (A:R-1), as with that in (A), is 1, because test T is guaranteed to pass *any* hypothesis fixed to entail e. Nevertheless, the probability in the corresponding (B) statement:

> (B:R-1): The probability that test T passes the hypothesis $h(e)$ constructed by rule R-1, *given that h(e) is false*

is equal to 0.

To see how this second probability is to be understood, consider the rule used in fixing the mean SAT score of students in my class. This is an example of a rule R-1. To evaluate (B:R-1) we ask: What is the probability that test T passes the hypothesis it constructs by rule R-1, for example $h(1121)$, given that the mean SAT score of students in my class is *not* equal to 1121? The answer is 0, equivalently, the severity of the test is 1. While there is always some rule by which to arrive at a use-constructed hypothesis (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. The common intuition to eschew using the same data both to construct and to test hypotheses (to eschew "double counting" of data), I claim, derives from the fact that a test that violates **UN** is guaranteed to pass the hypothesis it tests—the value of the probability in (A) is 1. But this does not entail that it is guaranteed to do so *whether or not the hypothesis it tests is false*. Indeed the test may have no chance of passing a false hypothesis: The value of the probability in

(B) may equal 0. Granted *if* (B) equals 1, then (A) equals 1; but the converse does not hold.[12]

Of course, most interesting hypotheses are not 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, I claim, is no. Tests may be highly severe, and still violate **UN.** This can be assured by a use-constructing rule that has built into it the requirement that the test have a degree of severity, call it β, sufficiently close to 1.

> *Rule R-β: For Constructing Highly Severe Tests (e.g., to degree β)*: Construct $h(e)$ such that the probability is extremely small, $(1 - \beta)$, for an experimental result to accord as well with $h(e)$ as does e, unless $h(e)$ were true (or approximately true).

Examples of rule R-β are found in rules for Neyman-Pearson confidence interval estimates and the design and interpretation of experiments. One need not be able to formally calculate β. The identical rationale underlies informal rules for using evidence. Consider how one might use detailed data on the shape of a dent in one's fender to construct a hypothesis about the likely make of the car that dented it. Such rules violate use-novelty; but they correctly indicate attributes of the dent's cause if they are severe in the sense of rule R-β, for example, if one can argue it is practically impossible for the dent to have the features it has unless it was created by a specific type of car tailfin.

My aim is not to find ways to avoid having to get new data. My aim is to avoid the tendency of some, having dismissed the **UN** requirement as too strong, to overlook the special constraints that tests that violate **UN** must satisfy in order to produce severe tests. Such violations of **UN** are typical, as would be expected, where they cannot be helped: where hypotheses are arrived at and affirmed by data, and it is impossible or impractical to obtain new data (e.g., evolutionary theory, epidemiology, anthropology, psychology and so on). However, as the next section shows, violations of **UN** are required even in cases lauded as models of severe and crucial tests, as the 1919 eclipse tests of Einstein's gravitational hypothesis.

5. The 1919 Eclipse Tests of Einstein's Law of Gravitation. Examining this case is intended to strengthen the above arguments, and to suggest how my severity criterion works in nonartificial testing contexts.

[12]This seems puzzling if one erroneously takes the "if" clause in (B) as a material conditional instead of as a conditional probability.

As noted earlier, terrestrial tests of Einstein's law of gravitation could not be severe since any light deflection would be undetectable with the instruments available in 1919. According to Einstein's theory, 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. 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 Eddington (1935) noted, "by strange good fortune an eclipse did happen on May 29, 1919" (p. 113) 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 Principe in the Gulf of Guinea, West Africa.

Eddington, together with Davidson, and Dyson, the Astronomer Royal, in Dyson et al. (1923), outline three hypotheses for which "it was especially desirable 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):

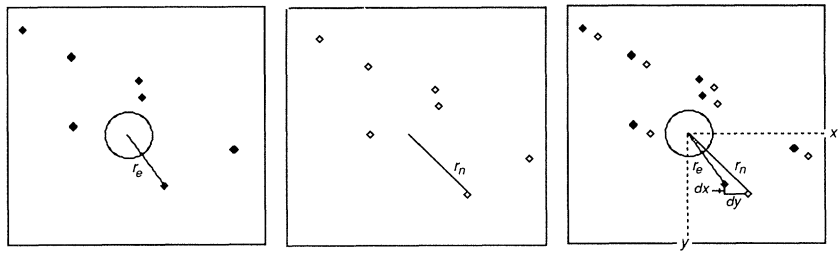
(1) Gravitation affects star light according to Einstein's law of gravitation: the deflection at the limb of the sun $\lambda = 1.75''$.
(2) Gravitation affects light according to the Newtonian law of gravitation: the deflection λ of a star at the limb of the sun $\lambda = 0.87''$.
(3) Gravitation does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

The "Newtonian" predicted deflection, (2), which stems from assuming 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 Principe, Eddington (1918) 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. (P. 36)

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

> . . . the results of the expeditions to Sobral and Principe can leave little doubt that a deflection of light takes place in the neighborhood of the sun and that it is of the amount demanded by Einstein's generalized theory of relativity, as attributable to the sun's gravitational field. (Dyson et al. 1923, 332)



(a) = "eclipse plate" with sun and surrounding stars

(b) = corresponding "night plate" taken of same star field when visible at night

(c) = both plates combined as they appear in the measuring machine

Figure 1. Comparing the "eclipse plate" and the "night plate" (adapted from von Klüber 1960, 52).

That the eclipse results support the Einstein hypothesis, then, states two things: first, that the test detected a deflection effect of the amount predicted by Einstein as against Newton and second, that the observed effect was "attributable to the sun's gravitational field" as described in Einstein's hypothesis. Correspondingly, the appraisal of the results involved two stages, which I label (i) and (ii). Each stage involved testing more local hypotheses, first to discriminate between values of parameter λ, and second to discriminate causes of the observed λ. Unfortunately, much of the fascinating data analysis must be omitted here. I trust that the following will suffice to make my case.

Stage (i): Estimating the Eclipse Deflection at the Limb of the Sun: The "observed" deflection (on May 19) actually has to be estimated by comparing the position of each star photographed at the eclipse (the eclipse plate) with its normal position as 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 (see Figure 1) allows the tiny distances to be measured in the x and y directions yielding dx and dy. These values, however, will depend on many factors: the way in which the two plates are accidentally clamped together, possible changes in the scale—mainly due to differences in the focus setting occurring between exposure of the eclipse and the night plate, on a set of other plate parameters, and finally, on the light deflection, λ, itself. (See a detailed discussion of this and several other eclipse tests of Einstein's deflection provided by H. von Klüber 1960. See also D. Moyer 1979.)

Consider the hypothesis about what degree of deflection has actually been observed in the eclipse experiments. In reaching *this* hypothesis (not the theoretical prediction of 1.75″), the data must be used to fix each of

these experimental parameters, so **UN** is apparently violated. But great pains were taken to ensure that 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 accuracy is required" (Eddington 1935, 115–116) than in the eclipse test.

This method allowed assigning probabilities to experimental outcomes under different hypotheses about λ (allowing severity to be calculated). The researchers arrived at hypotheses about the expected deflection (at the limb of the sun), λ, along with their probable errors (or, the measure now used, their standard errors). The two eclipse results, one from Sobral, one from Principe, taken as crucial support for Einstein were, with their standard errors:

(Sobral): The eclipse deflection = $1.98'' \pm 0.18''$

(Principe): The eclipse deflection = $1.61'' \pm 0.45''$.

Eddington (1935) 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. (P. 118)

The severity criterion **SC** explains the weight accorded each result. As stated in the passage, the hypothesis that "passes" here may be seen as $h: \lambda > 0.87''$. (So not-h asserts $\lambda \leq 0.87''$.) Consider passing h with the Sobral result of $1.98'' \pm 0.18''$. The probability of such a passing result given that λ is the Newtonian (half-deflection) $0.87''$ is practically 0. (The result is over 6 standard deviations in excess of 0.87.) So β, in rule R-β, is nearly 1. The Principe result, being around 1.6 standard deviations in excess of $0.87''$, is a reasonably severe passing result. That is, with reasonably high probability, around 0.95, there would *not* be such a passing result were λ equal to the Newtonian value ($0.87''$). (Here $\beta = 0.95$.) I say more about these severity assignments in section 6, question 1.

However, there was a third result also obtained from the Sobral expedition. In contrast to the other two this third result pointed not to Einstein's prediction, but, as Eddington declares, "with all too good agreement to the 'half-deflection,' that is to say, the Newtonian value . . ." (1935, 117). It also differed from the other two in being discounted as due to systematic errors. The instrument used, an astrographic telescope, was of

the same type as that used in the counted Principe 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.

Now, Popper had held this test up as a model of severity, unlike those of psychological theories of the day, we said, because the Einstein prediction dared to stick its neck out: A deflection far from the predicted value and near 0.87″, 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 a second key way such a violation can come about. In addition (but closely connected to) parameter adjustment, there is the violation due to *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 predicted λ. A holder of the **UN** requirement would deny the eclipse evidence supported (1*) since **UN** is violated. For Worrall, any hypothesis that explained e yet did not violate **UN**, *by dint of that alone,* is to be preferred (e.g., Worrall 1985, 313). There clearly was such a hypothesis, namely, the "Newtonian hypothesis", (2), to which the discounted (Sobral astrographic) result pointed. So on Worrall's view, it seems, the result used to construct (1*) would support Newton better than Einstein; it would not support the explanation of systematic distortion in (1*), as it was actually appraised.

Worrall may retort that exception incorporation is allowable in this case because there is a violation of "initial conditions", and his account assumes those are met (i.e., that all things are equal). But I do not think he can consistently do so, for Worrall is supposed to be giving us an account of when an experimental result counts as good evidence for a hypothesis—and thus it should be applicable to hypothesis (1*). And if we apply his account to hypothesis (1*), it would tell us the inference is *not* allowable—since it involves exception incorporation.

That an account of hypothesis appraisal should be applicable to a hypothesis like (1*) is underscored by the fact that it is *not* obvious that experimental error justified discounting one of the sets from Sobral, that is, that all things are *not* equal. Indeed, John Earman and Clark Glymour (1980) accuse Eddington of bias precisely because he "claimed the superiority of the qualitatively inferior Principe data, and suppressed ref-

erence to the negative Sobral results" (p. 84)—the Sobral astrographics. It was biased, because, on 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 another" (ibid., 75).

According to Earman and Glymour (1980), "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 locate the reason scientists went along with Eddington in his convincing them to see the eclipse results as "a rapprochement between German and British scientists" (ibid., 83), and conclude "This curious sequence of reasons might be cause enough for despair on the part of those who see in science a model of objectivity and rationality" (ibid., 85). The **UN** requirement does not distinguish problematic from unproblematic exception incorporations, so it cannot be looked to for help in defending the eclipse appraisal against the Earman-Glymour charge of bias.

But 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"—hypothesizing (1*)—is that it involved well-understood methods for constructing such a hypothesis. Results were deemed usable for estimating 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 0.03 mm.)—precisely what was *absent* from the suspect Sobral plates.[13] Consider the actual notes penned by Sobral researchers as reported in Dyson et al.:

> "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". (1923, 309; emphasis added)

Only after a goodly analysis of the distortion was it evident that the Sobral astrographic results pointed to only one hypothesis—(1*), systematic error.

[13]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 0.015 millimeters on the photographic plate. See von Klüber (1960, 50).

So it appears that the appraisal at stage (i), estimating the light deflection λ, violates use-novelty by both exception incorporation, and parameter-fixing: One and the same set of eclipse data was used both in constructing and testing hypotheses. 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-β).

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: the deflection effect $\lambda = 1.75''$, they did not blithely accept that Einstein's law of gravitation had thereby been given crucial support. Their challenge revolved around the second stage, stage (ii), determining the *cause* of the observed eclipse deflection. The problem was whether the test discriminated adequately the effect due to the sun's gravitational field from others that might explain the eclipse effect. A positive answer required 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 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 (1919b) wrote:

> [B]efore 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. (P. 138)

Were there *any* other cause 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 $1.75''$ is due to gravity).

The challenges at stage (ii) to the pro-Einstein interpretation of the observed deflection, then, 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, rather than gravitational effects of the sun

where N is a factor that at the same time saved the Newtonian law from refutation. Correspondingly, defenders of the Einstein hypothesis (1) responded both that the effect of the conjectured N-factor is simply too small to account for the eclipse effect; and if it were large enough to account for it, would have other false or contradictory implications.

Indeed stage (ii) seems to embody the very thing abjured by all novelty requirements (and by condemnations of ad hoc hypotheses): letting the road to hypothesis construction be constrained to explain e, while also counting e in that hypothesis's support. But, 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 in appraising these hypotheses. 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 used to save the Newtonian law, they also violated **UN** by exception-incorporation. What matters for my thesis is that although 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 the use of 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) just so long as one could argue that such a favorable result would be improbable if h were false, that is, so long as the test was reasonably severe. Examples abound in the literature; I briefly cite a few.

a. 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. These attempts were rejected, but only after being seriously considered by several scientists (e.g., by Arthur Schuster 1920). Their rejection did not turn on their violating **UN**. The problem, as is well put by Moyer (1979), was this:

> [T]he available adjustments are adjustments of parameters of trustworthy laws. . . . 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)

If, in order to get a hypothesis to pass, inconsistent parameter adjustments are permitted, then the test makes it very easy to pass hypotheses erroneously. The test has low severity.

b. 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. As was typical, there was a twofold response, here, by the scientist Lindemann and others. The required refraction 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.

c. Ether Modifications: Sir Oliver Lodge (e.g., Lodge 1919), who had promised ahead of time that if the Einstein effect was obtained he would save Newton by modifying conditions of the ether with special mechanical and electrical properties, proceeded, after the results were in, to do just that. (Lodge, a major proponent of spiritualism, held that the ether enabled contact with departed souls, in particular his son, Raymond.) Strictly speaking, since these hypotheses were constructed by Lodge in advance of 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. . . . [It] sounds quite promising at first since it explains . . . also the shift in the perihelion of Mercury, as well as the . . . shift of the spectral lenses, 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)

Lindemann is giving an apt description of what, in my view, is a pejorative case of parameter-fixing. In the pejorative case—far from trying to uncover the falsity of hypotheses—the test employs techniques that easily allow hypotheses to pass, whether or not they are true. This is not the case for tests of use-constructed hypotheses that ensure severity.

By prohibiting indiscriminantly 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 admitted only a construction of the Einsteinian gravitational hypothesis. This was argued by Harold Jeffreys (1919a, 1919b) (despite his initially assigning an extremely low Bayesian prior probability to Einstein's law). Jeffreys—one of the last holdouts—explains:

> It just 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. (1919a, 116; emphasis added)

In other words, in order to use the data to construct a hypothesis in a way that ensures high severity, one is led to Einstein's law of gravity! After reviewing the tests and all of the rival explanations (to which the February 1921 issue of *Nature* 1920–1921, 781–820, was entirely devoted), Dyson and Crommelin could conclude, "Hence we seem to be driven by exhaustion to the Einstein law as the only satisfactory explanation" (p. 788).

6. Anticipated Questions. *1. a.) How is the severity criterion* **SC** *to be applied; does the falsity of h refer to all of the (possibly infinite) alternatives to h? b.) Won't the existence of alternatives to h that imply the same experimental results (e.g., the same parameter value) as h preclude satisfying* **SC**?

While it was not the purpose of this paper to give an account of how to obtain the probabilities in **SC**, a few brief remarks should help avoid anticipated objections.

a.) The approach being advocated in this paper is piecemeal, where one question is asked at a time. Accordingly, it is not too difficult to set out all possible answers to that one question—limited as it is to hypotheses about the particular experimental phenomena. The probabilities come from requiring the hypotheses to be about specific experimental processes from which the outcomes may have originated.[14] One has to delimit the possible hypotheses by saying specifically which question is being asked in the experimental test. (Strategic experimental planning

[14]The view I have in mind, but cannot develop here, is that these tests would serve to check if certain common types of errors were being committed—for example, spurious correlations, erroneous parameter values, mistaken causal factors.

plays an important role here.) As the eclipse test shows, even appraisals of global claims may be broken down so that local tests do the work. At stage (i) the possible hypotheses were values for λ. One accepted hypotheses of form h: $\lambda > \lambda'$ (as against $\lambda \leq \lambda'$). For a particular observed mean deflection L, the severity criterion warrants accepting the use-constructed hypothesis

$h(L)$: λ exceeds $L - 2$ s.e.

where s.e. is the standard error of L. That "$h(L)$ is false" means λ does not exceed $L - 2$ s.e. Although this includes infinitely many alternative values of λ, the high severity requirement is met for each. The probability of *not* observing a deflection as large as L, given that λ is any of the values under "$h(L)$ is false" is high (at least 0.97).[15] There is no use of prior probability assignments.

At stage (ii), each N-factor conjectured to be responsible for the deflection effect was subjected to a *separate* severe test. Again, each concerned the value of a parameter. Examples of the hypotheses accepted were of the following form: The extent of the deflection effect produced by factor N (e.g., Anderson's shadow effect) is (much) less than 1.75″ (as against its negation). Jeffreys used the eclipse effect together with Kepler's law and the Mercury perihelion to accept Einstein's form of the gravitation law. Even without sustaining such a strong result, the high severity that is obtainable in such local tests allows accepting at least aspects of a cause, and values of experimental quantities.

b.) The existence of alternatives to h that imply the same experimental results as h would be a problem for the severity account being proposed only if it pronounced all such hypotheses equally well tested. This it manifestly does not do. Stage (ii) of the Einstein tests is particularly relevant here. The Newtonian defenders constructed alternative hypotheses (N-factors) so that they would predict the same deflection as Einstein's hypothesis and still save Newton. However, the only tests that such alternative hypotheses could pass were so lacking in severity as to enable their rejection. I have not discussed rejection here. While one can treat rejecting h as accepting not-h, having a separate account of the rejection of h is more natural. This would tell us whether h's failing a test (because of a discordance between evidence and h) warrants *rejecting* h. Such a rejection would be warranted, on the present account, to the extent that such a discordance is improbable, were h true. The proposed alternative explanations of the eclipse effect were ruled out on such grounds.

[15]More precisely, this probability is high ($\geq$0.97) for each value of λ alternative to $h(L)$. I thank Clark Glymour for noticing this needed clarification. A further discussion of severity applied to risk assessment occurs in Mayo (1988).

What about all the other alternative hypotheses that may be dreamt up that may 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 with Einstein's law? It is readily admitted that the eclipse tests were not tests of these alternative conceptions, and the severity criterion explains why: The eclipse results would be unable to test them severely. The eclipse tests were not even considered tests of Einstein's full theory. As Eddington (1919) 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. (P. 391)

It is important to stress, however, that the existence (or logical possibility) of alternative hypotheses that are not themselves tested by a given experiment does not alter the assessment of hypotheses that are tested. The severity calculation is unchanged. 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 by accepting 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 1935, 126)

2. How does my critique of use-novelty relate to Bayesian appraisals of novelty?

My critique agrees with the Bayesian in claiming **UN** is not necessary for a good test.[16] But our grounds for doing so are very different, and unless one sees why, my account will be misunderstood. My appraisal deliberately kept to the testing intuition underlying the **UN** requirement—a Popperian intuition of severity which I take to be captured by my definition. (While Popper's view of severity varied, he is clearly opposed to a Bayesian notion of building up probabilistic support.) I argued that the **UN** requirement does not accord with its *intended* aim of securing severe tests. In contrast, Bayesians (e.g., Howson 1984, Howson and Urbach 1989, Rosenkrantz 1977) argue that satisfying **UN** is not necessary for

[16]On one way of assigning probabilities—one where known evidence e yields $P(e/h) = P(e) = 1$—temporal and use-novelty *always* matters to a Bayesian in that old evidence fails to provide support for hypothesis h. This is Glymour's (1980) criticism. However, most Bayesians avoid assigning probability 1 to known evidence.

good support *according to Bayesian measures*. For example, Howson and Urbach (1989) conclude

> . . . that attempts to show that data which hypotheses have been deliberately designed to entail, as opposed to independently predicting, do not support those hypotheses fail. On the contrary, the condition for support, that $P(e| \sim h)/P(e|h)$ be small, may be perfectly well satisfied in many such cases. (P. 279)

However this Bayesian condition for support—that the likelihood ratio (or Bayes factor) be small— is at odds with the underlying aim of the **UN** requirement. It is unaltered by when or how hypotheses are generated. This stems from the likelihood principle which follows from Bayes's theorem. Roughly, this principle asserts that if each of two pieces of evidence gives the same likelihood to hypothesis *h,* then they have the same evidential import for *h*. Thus, as Rosenkrantz (1977) stresses, "The likelihood principle implies . . . the irrelevance of predesignation, of whether an hypothesis was thought of beforehand or was introduced to explain known effects" (p. 122). Most importantly, the Bayesian condition for support may be satisfied even where violating use-novelty does lead to a low severity test.

As already noted, my definition of severity stems from a non-Bayesian account: Neyman-Pearson tests. The severity of a test is not a measure of probability or support to hypotheses, but a measure of the probability (relative frequency) with which the test would lead to correctly failing (or not passing) hypotheses in some sequence of applications. (It is an "operating characteristic".[17]) According to Bayesians, as D. V. Lindley (1971) stresses:

> [U]nbiased estimates, . . ., sampling distributions, significance levels, power, all depend on something more [than the likelihood function]—something that is irrelevant in Bayesian inference—namely the sample space. (P. 436)

Our measure of severity, in contrast, does depend on all of the results the entire test process might have yielded.

The conflict between Bayesian principles and the criterion of severity may be seen by reference to an example discussed in arguing against the sufficiency of **UN**: "hunting with a shotgun" to find statistically significant correlations. While the present account condemns such a strategy where it yields a test with low or even 0-severity, for a Bayesian this fact

[17]It is often, wrongly, supposed that Neyman-Pearson methods require temporal and use-novelty (e.g., Gardner 1982, 14). All Neyman-Pearson principles require is that one take into account the alteration of error probabilities that may result from certain data-dependent hypotheses constructions.

does not alter the data's evidential import. (Discussions of how this type of hypothesis construction leads to bias in Bayesian methods occur in Giere 1969 and Mayo 1981.) Granted, with given assumptions about one's prior probabilities (generally, degrees of belief) in hypotheses, certain violations of severity can be made to correspond to tests which are poor or comparatively poor on Bayesian grounds. (Various attempts by which, through just the right assumptions and prior probabilities, use-novel hypotheses receive higher Bayesian support than use-constructed ones include Campbell and Vinci 1983, Howson and Urbach 1989, Maher 1988, Redhead 1986 and Rosenkrantz 1977.) But, in my view, that would still not identify the actual rationale for taking certain violations of use-novelty as problematic. When a passing result is condemned because it stems from a test that often passes hypotheses even if they are false, one is finding fault with the reliability of the test process. Such properties of the test process hold independently of how strongly I believe in a hypothesis it passes, even if I were able to quantify belief.

7. A Glimpse of Some Ways to Use These Results. Having the actual rationale of use-novelty in hand, one can quickly get to the bottom of the ongoing debate between proponents and opponents of the **UN** requirement: The debate is largely over whether, in appraising a particular test result, the severity of the testing process is relevant. For those who agree that it is severity that matters, it becomes possible to distinguish between legitimate and illegitimate tests that violate **UN.** The challenge, then, for (non-Bayesian) severity theorists (e.g., Worrall), is to articulate use-construction rules that ensure high severity—instances of my rule R-β. For Bayesians who agree that severity (in my sense) matters, the challenge is to square this with the likelihood principle.

REFERENCES

Anderson, A. (1919), "The Displacement of Light Rays Passing Near the Sun", *Nature 104*: 354.

———. (1920), "Deflection of Light During a Solar Eclipse", *Nature 104*: 436.

Campbell, R. and Vinci, T. (1983), "Novel Confirmation", *The British Journal for the Philosophy of Science 34*: 315–341.

Dyson, E. W.; Eddington, A. S.; and Davidson, C. (1923), "A Determination of the Deflection of Light by the Sun's Gravitational Field, from Observations made at the Total Eclipse of May 29, 1919", *Memoirs of the Royal Astronomical Society 62*: 291–333.

Earman, J. and Glymour, C. (1980), "Relativity and Eclipses: The British Eclipse Expeditions of 1919 and Their Predecessors", *Historical Studies in the Physical Sciences 11*: 49–85.

Eddington, A. (1918), "Gravitation and the Principle of Relativity", *Nature 101*: 34–36.

———. (1919), "Joint Eclipse Meeting of the Royal Astronomical Society", *Observatory 42*: 389–398.

———. (1935), *Space, Time and Gravitation, An Outline of the General Relativity Theory*. Reprint. Cambridge: Cambridge University Press.

Gardner, M. (1982), "Predicting Novel Facts", *The British Journal for the Philosophy of Science 33*: 1–15.

Giere, R. (1969), "Bayesian Statistics and Biased Procedures", *Synthese 20*: 371–387.

———. (1983), "Testing Theoretical Hypotheses", in J. Earman (ed.), *Minnesota Studies in the Philosophy of Science*. Vol. 10, *Testing Scientific Theories*. Minneapolis: University of Minnesota Press, pp. 269–298.

———. (1984), *Understanding Scientific Reasoning*. 2d ed. New York: Holt, Rinehart & Winston.

Glymour, C. (1980), *Theory and Evidence*. Princeton: Princeton University Press.

Glymour, C.; Scheines, R.; Spirtes, P. and Kelly, K. (1987), *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Orlando: Academic Press.

Grünbaum, A. (1979), "Is Freudian Psychoanalytic Theory Pseudo-Scientific by Karl Popper's Criterion of Demarcation?", *American Philosophical Quarterly 16*: 131–141.

———. (1989), "The Degeneration of Popper's Theory of Demarcation", in F. D'Agostino and I. C. Jarvie (eds.), *Freedom and Rationality: Essays in Honor of John Watkins*. Dordrecht: Kluwer Academic, pp. 141–161.

Howson, C. (1984), "Bayesianism and Support by Novel Facts", *The British Journal for the Philosophy of Science 35*: 245–251.

Howson, C. and Urbach, P. (1989), *Scientific Reasoning: The Bayesian Approach*. Lasalle, IL: Open Court.

Jeffreys, H. (1919a), Contribution to "Discussion on the Theory of Relativity", *Monthly Notices of the Royal Astronomical Society 80*: 116.

———. (1919b), "On the Crucial Test of Einstein's Theory of Gravitation", *Monthly Notices of the Royal Astronomical Society 80*: 138–154.

Lindemann, F. A. (1919), Contribution to "Discussion on the Theory of Relativity", *Monthly Notices of the Royal Astronomical Society 80*: 114.

Lindley, D. V. (1971), "The Estimation of Many Parameters", in V. P. Godambe and S. A. Sprott (eds.), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston, pp. 435–447.

Lodge, O. (1919), Contribution to "Discussion on the Theory of Relativity", *Monthly Notices of the Royal Astronomical Society, 80*: 106–109.

Maher, P. (1988), "Prediction, Accommodation, and the Logic of Discovery", in A. Fine and J. Leplin (eds.), *PSA 1988*, vol. 1. East Lansing: Philosophy of Science Association, pp. 273–285.

Mayo, D. (1981), "Testing Statistical Testing", in J. C. Pitt (ed.), *Philosophy in Economics: Papers Deriving from and Related to a Workshop on Testability and Explanation in Economics held at Virginia Polytechnic Institute and State University, 1979*. Dordrecht: Reidel, pp. 175–203.

———. (1985), "Behavioristic, Evidentialist, and Learning Models of Statistical Testing", *Philosophy of Science 52*: 493–516.

———. (1988), "Toward a More Objective Understanding of the Evidence of Carcinogenic Risk", in A. Fine and J. Leplin (eds.), *PSA 1988*, vol. 2. East Lansing: Philosophy of Science Association, pp. 489–503.

Moyer, D. (1979), "Revolution in Science: The 1919 Eclipse Test of General Relativity", in A. Perlmutter and L. Scott, (eds.), *On the Path of Albert Einstein*. New York: Plenum Press, pp. 55–102.

Musgrave, A. (1974), "Logical versus Historical Theories of Confirmation", *The British Journal for the Philosophy of Science 25*: 1–23.

———. (1978), "Evidential Support, Falsification, Heuristics, and Anarchism", in G. Radnitzky and G. Andersson (eds.), *Progress and Rationality in Science*. Dordrecht: Reidel, pp. 181–201.

———. (1989), "Deductive Heuristics", in K. Gavroglu, Y. Goudaroulis, and P. Nicolacopoulos (eds.), *Imre Lakatos and Theories of Scientific Change*. Dordrecht: Kluwer Academic, pp. 15–32.

Nature. (1920–1921). Volume 106.

Newall, H. F. (1919), Contribution to "Joint Eclipse Meeting of the Royal Society and the Royal Astronomical Society", *The Observatory 42*: 395–396.

———. (1920), "Note on the Physical Aspect of the Einstein Prediction", *Monthly Notices of the Royal Astronomical Society 80*: 22–25.

Nickles, T. (1987), "Lakatosian Heuristics and Epistemic Support", *The British Journal for the Philosophy of Science 38*: 181–205.

Popper, K. (1962), *Conjectures and Refutations*. New York: Basic Books.

———. (1974), "Replies to My Critics", in P. A. Schilpp (ed.), *The Philosophy of Karl Popper*. Book 2. Lasalle, IL: Open Court, pp. 961–1197.

———. (1979), *Objective Knowledge: An Evolutionary Approach*. Oxford: Oxford University Press.

———. (1983), *Realism and the Aim of Science*. Totowa, NJ: Rowman & Littlefield.

Redhead, M. (1986), "Novelty and Confirmation", *The British Journal for the Philosophy of Science 37*: 115–118.

Rosenkrantz, R. (1977), *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*. Boston: Reidel.

Schuster, A. (1920), "The Influence of Small Changes of Temperature on Atmospheric Refraction", *Proceedings, Physical Society of London 32*: 135–140.

von Klüber, H. (1960), "The Determination of Einstein's Light-Deflection in the Gravitational Field of the Sun", in A. Beer (ed.), *Vistas on Astronomy 3*, pp. 47–77.

Worrall, J. (1978a), "Research Programmes, Empirical Support, and the Duhem Problem: Replies to Criticism", in G. Radnitzky and G. Andersson (eds.), *Boston Studies in the Philosophy of Science*. Vol. 58, *Progress and Rationality in Science*. Dordrecht: Reidel, pp. 321–338.

———. (1978b), "The Ways in Which the Methodology of Scientific Research Programmes Improves on Popper's Methodology", in G. Radnitzky and G. Andersson (eds.), *Boston Studies in the Philosophy of Science*. Vol. 58, *Progress and Rationality in Science*. Dordrecht: Reidel, pp. 45–70.

———. (1985), "Scientific Discovery and Theory-Confirmation", in J. C. Pitt (ed.), *Change and Progress in Modern Science: Papers Related to and Arising from the Fourth International Conference on History and Philosophy of Science*. Dordrecht: Reidel, pp. 301–331.

———. (1989), "Fresnel, Poisson and the White Spot: The Role of Successful Predictions in the Acceptance of Scientific Theories", in D. Gooding, T. Pinch and S. Schaffer (eds.), *The Uses of Experiment: Studies in the Natural Sciences*. Cambridge, England: Cambridge University Press, pp. 135–157.

Zahar, E. (1973), "Why Did Einstein's Programme Supersede Lorentz's? (I) and (II)" *The British Journal for the Philosophy of Science 24*: 95–125, 223–262.