DEBORAH G. MAYO

SEVERE TESTS, ARGUING FROM ERROR, AND METHODOLOGICAL UNDERDETERMINATION

(Received in revised form 10 July 1995)

A major problem that has been thought to stand in the way of an adequate account of hypothesis appraisal may be termed the *alternative hypothesis objection*: that whatever rule is specified for positively appraising H, there will always be rival hypotheses that satisfy the rule equally well. Evidence in accordance with hypothesis H cannot really count in favor of H, it is objected, if it counts equally well for some (perhaps infinitely many) other hypotheses that would also accord with H.

This problem is a version of the general problem of underdetermination of hypotheses by data: if data cannot univocally pick out hypothesis H over alternatives, then the hypothesis is underdetermined by the data. Some have considered this problem so intractable as to render any attempt to erect a methodology of hypothesis appraisal impossible. No such conclusion is warranted, however. There is no general argument showing that all rules of appraisal are subject to this objection. At most the argument has been sustained against certain specific rules (e.g., the straight rule, simple hypothetico-deductivism, falsificationist accounts). A more adequate account of hypothesis testing, I will argue, gets around the underdetermination challenge. In this account, evidence is to be taken as a good test of (or good grounds for) a hypothesis only to the extent that it can be seen as the result of passing a *severe test* of that hypothesis. My task in this paper is first to sketch an account of severe tests, and second, to show how it answers the alternative hypothesis objection. To anticipate two of my main theses, I shall be arguing:

(1) The existence of hypotheses alternative to H that accord with evidence *e* as well as H does, does not prevent H from passing a severe test with *e*.

Philosophical Studies **86:** 243–266, 1997.

(2) Even if there are alternative hypotheses that entail or fit evidence *e* as well as H does, there are not always alternatives equally severely tested by *e*.

I

The "alternative hypothesis objection" that concerns me needs to be distinguished from some of the more radical variants of underdetermination. Some of these more radical variants are the focus of a paper by Larry Laudan (1990), "Demystifying Underdetermination". (See also Laudan, 1995.)

". . . [O]n the strength of one or another variant of the thesis of underdetermination," Laudan remarks, "a motley coalition of philosophers and sociologists has drawn some dire morals for the epistemological enterprise. . . . " Several examples follow:

> . . . Quine has claimed that theories are so radically underdetermined by the data that a scientist can, if he wishes, hold on to *any* theory he likes, 'come what may.' Lakatos and Feyerabend have taken the underdetermination of theories to justify the claim that the only difference between the empirically successful and empirically unsuccessful theories lay in the talents and resources of their respective advocates. . . . Hesse and Bloor have claimed that underdetermination shows the *necessity* for bringing noncognitive, social factors into play in explaining the theory choices of scientists." (Laudan 1990, p. 268)

Laudan distinguishes two varieties of underdetermination. The first, Laudan calls *the nonuniqueness thesis*: "It holds that: *for any [hypothesis H] and any given body of evidence supporting H, there is at least one rival (i.e., contrary) to H that is as well supported as H"* (p. 271). A far more extreme position he calls *the egalitarian thesis*: "It insists that: *every hypothesis is as well supported by the evidence as any of its rivals*" (p. 271).

Laudan argues that the Quinean thesis that "any hypothesis can rationally be held come what may" as well as other strong relativist positions are committed to the egalitarian thesis, and that a close look at underdetermination arguments shows that they at most sustain variants of the nonuniqueness thesis: that there are always one or more alternatives to H that are as well supported on the evidence as is H. Laudan denies that the nonuniqueness thesis has particularly dire consequences for methodology; his concern is only with the extreme

challenge "that the project of developing a methodology of science is a waste of time since, no matter what rules of evidence we eventually produce, those rules will do nothing to delimit choice. . . " (p. 281). I agree that the nonuniqueness thesis will not sustain the radical critique of methodology as utterly "toothless", but I am concerned with showing that methodology has some real bite!

II

Even if it is granted that empirical evidence serves *some* role in delimiting hypotheses and theories, the version of underdetermination that still has to be grappled with is the alternative hypothesis objection with which I began, that for any hypothesis H and any evidence, there will always be a rival hypothesis equally successful as H. The objection, it should be clear, is that criteria of success based on methodology and evidence *alone* underdetermine choice. It may be stated more explicitly as the thesis of methodological underdetermination (MUD):

> *Methodological Underdetermination*: any evidence taken as a good test of (or good support for) hypothesis H would (on that account of testing or support) be taken as an equally good test of (or equally good support for) some rival to H.

While not alleging that anything goes, it is a mistake to suppose that the MUD thesis poses no serious threat to the methodological enterprise. The reason formal accounts of testing and confirmation ran into trouble was not that they failed to delimit choice at all, but that they could not delimit choice sufficiently well (e.g., Goodman's riddle). Moreover, if hypothesis appraisal is not determined by methodology and evidence, then when there is agreement in science, it would seem to be the result of extra-evidential factors (as Kuhn and others argue).

Granted, the existence of alternative hypotheses equally well tested by evidence need not always be problematic. For example, it is unlikely to be problematic that a hypothesis about a continuous parameter is about as well tested as another hypothesis that differs by only a tiny fraction. In the following discussion of my account

of severe testing, I will focus on what has seemed to be the most threatening variants of the MUD challenge.

Clearly, not just any rule of evidential appraisal that selects a unique hypothesis will constitute an adequate answer to the challenge. Not just any sort of rule is going to free us from many of the most troubling implications of MUD. That is why the Bayesian Way does not help with my problem. The Bayesian Way of differentially supporting two hypotheses that entail (or equally well fit) the data is by assigning them different prior probabilities.[1] But, prior probabilities, except in very special cases, are matters of personal, subjective choice – threatening to lead to the relativism we are being challenged to avoid (inviting a MUD-slide, one might say).[2]

How does appealing to the notion of severity help? While there are a number of different conceptions of severe tests, we can broadly characterize such accounts as holding the following general severity requirement:

> *Severity Requirement*: Evidence *e* should be taken as good grounds for H only to the extent that H has passed a *severe test* with *e*.

What I want to argue in this paper is that the alternative hypothesis objection loses its sting once the notion of severity is appropriately made out.

III

It is not difficult to see that the MUD charge instantiated for a method of severe testing T is more difficult to sustain than when it is waged against mere entailment or instantiation accounts of inference. The charge of methodological underdetermination (for a method of severe testing T) alleges that: *for any evidence that test T takes as passing hypothesis H, there is always a rival hypothesis to H that passes test T as severely as H does*. On my account of severe testing this charge is false.

The centerpiece of my account is the notion of severity involved. Unlike accounts that begin with evidence *e* and hypothesis H and then seek to define an evidential relationship between them, severity refers to a method or procedure of testing, and cannot be assessed

without considering how the data were generated, modeled, and analyzed in order to obtain relevant evidence in the first place. We can capture this by saying that assessing severity always refers to a framework or context of *experimental inquiry*. While MUD gets off the ground when hypothesis appraisal is considered as a matter of some formal or logical relationship between evidence or evidence statements and hypotheses, it does not in the experimenter's testing framework.

The goal of an experimental strategy is to ensure, with high probability at least, that erroneous attributions of experimental results are avoided. The error of concern in passing hypothesis H is that one will do so while H is not true.[3] Passing a severe test, in the sense I advocate, counts for hypothesis H because it corresponds to having good reasons for ruling out specific versions and degrees of this error.

Even widely different approaches concur that, minimally, for H to pass a test with evidence *e*, *e* should agree with or fit what is expected or predicted according to H. A strong construal of "H fits *e*" would require that H entails *e*; a more useful notion would require that *e* be within some specified distance from H.[4] Although the fit requirement can vary, proponents of some version of the severity requirement agree that something beyond this minimal requirement is needed in order for a test to be genuine – the test must also be severe. A severity requirement stipulates what this "something more" should be. A rough statement of my notion of severity is this: A passing result *e* is a severe test of hypothesis H just to the extent that *e* results from a procedure that would rarely yield such a passing result, were H false. Were H in error, then, with high (frequentist) probability, the test would either have failed hypothesis H or would have produced an outcome more discordant from H than *e* is.

Assessing this probability requires considering the probability a given test procedure has for detecting a given type of error. Such a probability is called an *error probability*, and while error probabilities play a key role in standard statistical practice, they are largely overlooked in philosophical accounts of testing. What they enable us to do is distinguish the well-testedness of two hypotheses – despite their both fitting the data equally well. Two hypotheses, H_1 and H_2, may accord with data equally well, but nevertheless be tested dif-

ferently by the data. The data may be a better, more severe, test of one than of the other. The reason is that the procedure from which the data arose may have had a good chance of detecting one type of error, while it may not have had a good chance of detecting a different type of error. Indeed, a test procedure may be highly capable of detecting one type of error but not capable at all of detecting some other error. What is ostensibly the same piece of evidence is really not the same at all – at least not on the criterion of severe tests.

IV

Let us clarify this criterion of severity. The severity requirement is:

> *Severity Requirement*: Passing a test T (with experimental result *e*) counts as a good test of or good evidence for hypothesis H just to the extent that *e* fits H and T is a *severe test* of H.

and the criterion of severity SC I suggest is this:

> *Severity Criterion (SC)(a)*: There is a very high probability that test procedure T would *not* yield such a passing result, if hypothesis H is false.

By "such a passing result" I mean one that fits H at least as well as H does. Its complement, in other words, would be a result that either fails H or one that still passes H but accords less well with H than does *e*. Very often, it is useful to express SC in terms of the improbability of the passing result. This equivalent expression is given in (b):

> *Severity Criterion (SC)(b)*: There is a very low probability that test procedure T would yield such a passing result, if hypothesis H is false.

The probabilities referred to in our severity criterion are frequentist probabilities. How precisely to understand them is a question whose answer requires a fuller discussion than I can give here. To get a handle on the main idea, a high severity assignment can be compared to a high score on an exam of some sort. A high severity assignment asserts that were we experimenting on a system where

hypothesis H is false, then, in a long series of trials of this experiment, hypothesis H would very rarely be accorded such a good score; the overwhelming preponderance of experimental trials would yield outcomes that accord H a worse fit or a lower score.

Arguing from passing a severe test, as I see it, corresponds to an informal pattern of argument with which we are very familiar. This informal argument might be called an *argument from error* or *learning from error*. The overarching structure of the argument is guided by the following thesis:

> It is learned that an error is absent when (and only to the extent that) a procedure of inquiry (which may include several tests taken together) that has a very high probability of detecting an error if (and only if[5]) it existed, nevertheless detects no error.

Such a procedure of inquiry, we can say, is one with a high capability of severely probing for the error in question – we may call it a *reliable (or highly severe) error probe*. According to the above thesis, we can argue that an error is absent if it fails to be detected by a highly reliable error probe. The informal argument from error converts to the corresponding argument from severity when hypothesis H is written as an assertion that a given error is absent:

> H: the error is absent

while not-H asserts that the error is present. The evidence indicates the correctness of hypothesis H (the correctness of the claim that the error is absent), when H passes a severe test. An analogous argument can also be given to infer the presence of an error.[6]

Tools for medical diagnoses (e.g., ultrasound probes) offer useful analogies to extract these intuitions about severity: If a diagnostic tool had little or no chance of detecting a disease, even if it is present (low severity), then a passing result – a clean bill of health – with that instrument fails to provide grounds for thinking the disease is absent. That is because the tool has a very high probability of issuing in a clean bill of health even when the disease is present. It is a highly unreliable error probe. Alternatively, suppose a diagnostic tool had an overwhelmingly high chance of detecting the disease just in case

it is present – suppose it is a highly severe error probe. A clean bill of health with that kind of tool provides strong grounds for thinking the disease is not present. For if the disease were present, our probe would almost certainly have detected it.

It is important to stress that my notion of severity always attaches to a particular hypothesis passed or a particular inference reached. A procedure may be highly severe for arriving at one type of hypothesis and not another. Consider again a diagnostic tool with an extremely high chance of detecting a disease. Finding no disease (a clean bill of health) may be seen as passing hypothesis H_1: no disease is present. If H_1 passes with so sensitive a probe, then H_1 passes a severe test. However, the probe may be so sensitive as to have a high probability of declaring the presence of the disease, even if no disease exists. Declaring the presence of the disease may be seen as passing hypothesis H_2: the disease is present. If H_2 passes a test with such a highly sensitive probe, then H_2 has *not* passed a severe test. That is because there is a very low probability of *not* passing H_2 (not declaring the presence of the disease) even when H_2 is false (and the disease is absent). The severity of the test that hypothesis H_2 passes is very low.

V

Accounts of testing that are based on severity requirements may themselves be the subject of alternative hypotheses objections. There are two ways in which such an objection can be and have been raised. The first way is to object that passing a severe test cannot really count in favor of a hypothesis H because there are always other alternative rival hypotheses that are equally severely tested by given evidence *e*. The second way is to argue that the existence of alternative hypotheses that accord with evidence *e* as well as H does precludes high severity from being obtained for H in the first place. Examining these two types of objections allows me to address anticipated misunderstandings of my severity criterion, and will be relevant later in dealing with MUD. I consider them in turn:

a. *The alternative hypothesis objection against Popperian severity*.

Karl Popper is well-known for stressing the importance of severe tests. For Popper, hypothesis H passes a severe test with *e* if all alternatives to H that have so far been considered or tested entail not-*e* (or render not-*e* highly probable).[7] Suppose outcome *e* is observed. The hypotheses that entail not-*e* are thereby rejected, and H, which entails *e*, passes the test severely. The problem is that there are generally many not-yet-considered alternative hypotheses that also entail *e*.[8] So it would seem that *e* counts as much for these other hypotheses as it does for H. To put this in other words, if hypothesis H passes the tests failed by all of the existing rivals, then H gets the badge for "best-tested theory of the moment". Yet, any other hypothesis that would also pass the existing tests would have to be said to do as well as H – by Popper's criteria of judging tests. Popper's problem here is that the grounds for the "best tested" badge would also be grounds for giving the badge to countless many other (not yet even though of) hypotheses, had they been the ones considered for testing. So this alternative hypothesis objection goes through for Popper's account.

This is not the case for the severity criterion I have set out. A non-falsified hypothesis H that passes the test failed by each rival hypothesis H′ that has been considered, has passed a severe test for Popper – but not for me. Why not? Because for H to pass a severe test in my sense it must have passed a test with high power at probing the ways H can err. And the test that alternative hypothesis H′ failed need not be probative in the least so far as H's errors go. So long as two different hypotheses can err in different ways, different tests are needed to probe them severely.

b. *An alternative hypothesis objection to the severity criterion SC*

It might seem that the possibility of alternative hypotheses that equally well accord with available evidence prevents my severity requirement from ever being met. Demanding high severity in my sense, it might be charged, is too demanding. This is the second type of objection to accounts based on severity. John Earman (1992, p. 117) raises it against my severity criterion SC.

Earman's charge, however, stems from appraising severity from a Bayesian perspective. His criticism of my severity requirement

seems to be that it requires a low probability to what I call the *Bayesian catchall factor*. The Bayesian catchall factor (in assessing H with evidence *e*) is:

P(*e*/not-H).

The catchall[9], not-H, refers to all possible hypotheses other than H, including those that may be conceived of in the future. It is a disjunction of all hypotheses that could explain the evidence other than hypothesis H. To ask, What is the value of the Bayesian catchall factor? is to ask, What is the probability that evidence *e* would arise given that hypothesis H is false and some one of the possible alternative hypotheses were true? On the Bayesian scheme of inference evidence *e* confirms hypothesis H to the extent that *e* is made more probable under H than under not-H (i.e., to the extent that P(*e*/H) > P(*e*/not-H)). So the lower the assignment to the Bayesian catchall, the higher the Bayesian confirmation of H.

Assessing the probability of *e* on the catchall hypothesis requires a prior probability assignment to the catchall, and this requires a probability assignment to each possible hypothesis other than H. These probability assignments are illegitimate for frequentists such as myself.[10] In addition, without knowing what other hypotheses might someday be entertained, we cannot objectively assess how probable *e* is, given the catchall. As Wesley Salmon puts it:

> What is the likelihood of any given piece of evidence with respect to the catchall? This question strikes me as utterly intractable; to answer it we would have to predict the future course of the history of science. (Salmon 1990, p. 329)

Earman grants the *desirability* of a low assignment to the Bayesian catchall factor, because, as we said, the lower its value, the more Bayesian confirmation accrues to H. The difficulty he sees is in obtaining it. While I quite agree that this presents an obstacle for the Bayesian approach to support, satisfying the severity criterion SC does not require computing the Bayesian catchall factor. And because it does not, alternatives in the catchall that might also fit the evidence do not present the obstacle to obtaining high severity that Earman thinks they do.

Consider the example Earman raises in this connection:

> If we take *H* to be Einstein's general theory of relativity and *e* to be the outcome of the eclipse test, then in 1918 and 1919 physicists were in no position to be confident

> that the vast and then unexplored space of possible gravitational theories [−GTR] does not contain alternatives to GTR that yield the same prediction for the bending of light as GTR. (Earman 1992, p. 117)

(I substitute his *E* with *e* for consistency with my notation.) In fact, he continues, there is an endless string of such alternative theories. *The presumption is that alternatives to GTR that also predict light bending would prevent high severity in the case of the eclipse test.*

But alternatives to GTR did not prevent the eclipse results from being used to test severely the hypotheses for which the eclipse experiments were designed. Those tests, as I discuss elsewhere (Mayo 1991), proceeded by asking specific questions: Is there a deflection of light of about the amount expected under Einstein's law of gravitation? Is it due to gravity? Are alternative factors responsible for appreciable amounts of the deflection? Finding the answers to these questions in a reliable manner did not call for ruling out any and all alternatives to the GTR.

Take the question of the approximate deflection of light. If this is the primary question of a given inquiry then alternative answers to it are alternative values of the deflection, not alternatives to the general theory of relativity. If alternative theories predict the same results, so far as the eclipse experiments go, as Earman says they do, then these alternatives are not *rivals* to the particular hypotheses under test. If the endless string of alternative theories would, in every way, give the same answers to the questions posed in the 1919 tests, then they all agree on the aspects of the gravitation law that was tested. They are not members of the space of alternatives that are relevant for the primary question being addressed.

This reply depends on a key feature of my account of testing. In this account, experimental inquiry is viewed in terms of a series of models, each with different questions, stretching from low-level theories of data and experiment to higher level hypotheses and theories of interest. The hierarchy of models set out by Patrick Suppes (1962) serves nicely for my purposes.[11] The model placed at the "top" of my hierarchy is the primary hypothesis or question of interest. Suppose a particular inquiry has as its primary focus the deflection effect. Einstein's gravitational hypothesis makes certain claims about the value of the deflection of light as well as about its causes. The eclipse effect was used to test the extent to which these

claims were correct. The concern was to rule out rival (Newtonian) values of the deflection and rival causal explanations of any deflection found. In relation to these hypotheses about the deflection effect, alternatives to the general theory of relativity are on a higher level. The higher-level alternatives are not even being tested by the test at hand. Most importantly, higher-level alternatives pose no threat to learning with severity what they needed to learn in the specific 1919 experiments.

The general lesson goes beyond answering Earman. It points up a strategy for dispelling a whole class of equally good fitting alternatives to a hypothesis H. The existence of alternatives at a higher level than H is no obstacle to finding high severity for H. Asking the higher level questions, just like asking about the correctness of the whole of the GTR, is tantamount to *asking the wrong question* – relative to the test at hand.

Our approach to experimental learning recommends proceeding in the way one ordinarily proceeds with a complex problem: break it up into smaller pieces, some of which, at least, can be tackled. One is led to break things down if one wants to learn. For we learn by detecting and ruling out specific errors. Satisfying the severity requirement demands that we make our questions appropriately small or local. To put this in other words, it directs us to learn by solving local experimental problems. By using simple local contexts in which the assumptions may be shown to hold sufficiently, it is possible to ask *one question at a time*. Setting out all possible answers to this one question becomes manageable, and that is all that has to be "caught" by our not-H.

Within an experimental testing model, the falsity of a primary hypothesis H takes on a very specific meaning. If H states that a parameter is greater than some value c, not-H states it is less than c; if H states that factor x is responsible for at least p% of an effect, not-H states it is responsible for less than p%; if H states an effect is caused by factor f, not-H may say it is caused by some other factor possibly operative in the experimental context; if H states the effect is systematic – of the sort brought about more often than by chance – then not-H states it is due to chance. To determine what, if anything, is learned from an experimental result, we must ask: *what, if anything, has passed a severe test*?

This does not mean we are precluded from severely testing higher-level theories by other tests. The results from local tests may be accumulated so as to allow us to say that several related hypotheses are correct, or that a theory solves a set of experimental problems correctly. Indeed, Earman himself discusses the progress that has been made in showing which available experiments can serve to eliminate whole chunks of theories of gravity (e.g., so-called "non-metric" theories), which sets of theories are still not distinguished by known experiments, and how further progress along these lines might be made (Earman 1992, p. 177). Something like this kind of program of partitioning and eliminating of chunks of theories is what the present program would call for at the level of large-scale theories.

VI

It may be asked how we obtain a high severity assignment even limiting ourselves to a particular primary question. Since the key to avoiding methodological underdetermination is to be able to distinguish the severity of tests, it is necessary to address this question before tackling MUD. It is useful to consider a specific example. One kind of example with which we are familiar comes from standard statistical practice. We are interested in determining whether a substance or practice increases a risk of some sort in a given population. To this end we may perform a clinical trial and at the end of the study record the difference between the incidence of the risk among those treated with the substance or drug as compared with the risk incidence among those not treated. (This is a standard treatment-control clinical trial.) An example of interest might be the effect of oral contraceptives on the risk of blood-clotting disorders in women.[12]

A central error of concern is that any difference in the rate of clotting disorders observed in a given study might be simply spurious, due to chance, or not statistically significant. This error is formally expressed in a standard null (or "no-effect") hypothesis, H_0. H_0 asserts there is no real increase in risk among women treated with the Pill (for the specified time period in the study). That is:

H_0: The risk of clotting disorders in women who use the Pill is no higher than among women who do not.

A standard or "canonical" test procedure is to reject null hypothesis H_0 just in case the observed difference in risk rates exceeds 2 standard deviations.[13] This is the same as rejecting the null whenever the observed difference is significant at the 0.03 level. If we reject the null hypothesis we affirm this error is absent and pass the hypothesis H:

H: The risk of clotting disorders is higher among Pill users than among women who do not use the Pill.

According to my account we have to ask: Suppose I were to pass H (the risk of clotting is higher among Pill users) whenever this statistical test procedure tells me to reject the null hypothesis H_0. How severe would that test procedure be? Would it often lead me to mistaking chance effects for systematic or "real" ones?

We then ask the above questions more formally in terms of the significance level of the test: What is the probability of the experiment producing so large a difference from what is expected under the null hypothesis, if in fact the null hypothesis is true? The answer, we said, was 0.03. We have,

$$P(\text{Test T passes H, given H is false } (H_0 \text{ is true})) = 0.03.$$

The severity of a test that passes H (rejects H_0) is 1 minus the significance level, namely, 0.97. (That is, the severity in this case is 1 minus the probability of erroneously passing H.) So, the severity for H is high (0.97). This means that if the null hypothesis were true, and the observed increase in risk merely "due to chance", then, with high probability, I would have detected this. I did not detect this, hence, by an argument from error, the result is a good indication that a genuine, and not a mere chance, increase exists.

By rejecting the null hypothesis H_0 only when the significance level is low, we automatically ensure that any such rejection constitutes a case where the non-chance hypothesis H passes a severe test. Such a test procedure T can be described as follows:

> *Test Procedure T*: Pass H whenever the statistical significance level of the observed difference is less than or equal to α (for some very small value of α).

Calculating severity for passing H is one minus the probability of such a passing result, when in fact the results are due to chance, i.e., when H_0 is true. By definition,

> P(Test T yields a statistical significance level $\leq \alpha$, given H_0 is true) = α.

That "test T yields a statistical significance level $\leq \alpha$" is identical to test T passing hypothesis H (and rejecting H_0). Thus, the severity of the test that hypothesis H passes equals $1 - \alpha$.

Sustaining an argument from error does not require computing a precise value of the probability, nor need one identify a specific statistical model that applies. It is enough to be able to argue that the severity is high in some more qualitative manner. Take an example used by Hacking (1983). Hacking asks, What convinces someone that an effect is real? Low-powered electron microscopy reveals small dots in red blood platelets, called dense bodies. Are they merely artifacts of the electron microscope?

> One test is obvious: can one see these selfsame bodies using quite different physical techniques? . . . In the fluorescence micrographs there is exactly the same arrangement of grid, general cell structure and of the 'bodies' seen in the electron micrograph. It is inferred that the bodies are not an artifact of the electron microscope . . . It would be a preposterous coincidence if, time and again, two completely different physical processes produced identical visual configurations which were, however, artifacts of the physical processes rather than real structures in the cell (Hacking 1983, pp. 200–201).

Hacking's argument exemplifies my argument from error. The error of concern is to take as real structure something that is merely an artifact. The evidence is the identical configurations produced by completely different physical processes. Such evidence is extremely unlikely if the results were due to "artifacts of the physical processes rather than real structures in the cell" (ibid.). This is justified by a good deal of background knowledge, e.g., we made the grid, we know all about these grids, etc.[14] It is overwhelmingly improbable that all the instruments and techniques conspire together to make

the evidence appear as if the effect is real, when it is actually an artifact.

It may be objected that with substantive questions all the possible alternatives even to a single primary hypothesis cannot be set out. Even where this is so, it does not present an insurmountable obstacle to experimental testing. In such cases we may often manage to find a more general or less precise hypothesis such that when *it* is severely tested there are at the same time grounds for rejecting all the alternatives in a manner that meets the severity requirement. What we try to do is emulate what is possible in canonical experimental tests.

For example, Jean Perrin, in his experiments on Brownian motion, was able to rule out, as causes of the motion, all factors outside a certain liquid medium. He did so by arguing that if the observed Brownian motion were due to such external factors – *whatever they might be* – the motion of Brownian particles would follow a specified coordinated pattern. His experimental tests, he went on to argue, would almost surely have detected such a pattern of coordination, were it to exist; but only uncoordinated motion was found.

VII

In addition to showing how high severity may be attained, the above examples of severe tests let us make short work of the variants of the alternative hypothesis objection. Let us see how.

(a) *Alternatives That Ask the Wrong Question*

For starters, the examples in Section V exemplify the point I made in grappling with Earman's criticism. We can avoid pronouncing as well-tested a whole class of hypotheses that, while implying (or in some other way fitting) a given result, are nevertheless not part of the hypothesis space of the primary test. They are simply asking after the *wrong question*, at least so far as the given test is concerned.

In the case of the treatment-control experiment above, examples of wrong question hypotheses would be any of the various causal hypotheses that might be given to explain any increased risk rate. That these other hypotheses predict the increased risk that is observed does not redound to their credit in the same way that the results

count in favor of H (merely that a genuine increased risk exists). This shows up in a difference in assessing the severity for passing H, a correlation hypothesis, as opposed to the severity for passing a causal hypothesis. The test designed to test severely if the effect is easily explained by chance does not automatically have a good chance of detecting mistakes about the effect's cause. With regard to questions about the cause of a systematic effect, a whole different set of wrong answers needs to be addressed.

This same argument can be made quite generally to deal with alternatives often adduced in raising the alternative hypothesis objection. While these alternatives also fit or accord with H, they may be shown to be less well tested than is H. There are two main points: First, these alternative hypotheses do not threaten a high severity assignment to the primary hypothesis. Second, it can be shown that these alternatives are not equally severely tested. *Because they ask a different question, the ways in which they can err differ, and this corresponds to a difference in severity*. Moreover, if the primary hypothesis is severely tested, then these alternatives are less well tested.

(b) *Alternative Primary Hypotheses*

It will be objected that I have hardly answered the alternative hypothesis objection when it becomes most serious: the existence of alternative hypotheses to the primary hypothesis (alternative answers to the same primary question). This is so. But we can handle such cases in much the same fashion as the previous ones – via a distinction in severity.

One point that bears noting at this juncture is that I am surely not aiming to show that all alternatives can always be ruled out. Experimental learning is never guaranteed. What I do claim to show, and all that avoiding MUD requires, is that there are not always equally well tested alternatives that count as genuine rivals, and that there are ways to discriminate hypotheses on grounds of well-testedness that get around objections. Let us consider some alternative primary hypotheses.

Maximally likely alternatives. A type of alternative that is often adduced in raising the alternative hypothesis objection is one constructed after-the-fact to perfectly fit the data in hand. By perfectly

fitting the data, by entailing it, the data make the hypothesis maximally likely[15] (i.e., P(*e*/H) = 1). The corresponding underdetermination argument is that for any hypothesis H, there is a maximally likely alternative that is tested as well as or better than H.

To demonstrate the ease with which such hypotheses may be constructed, consider an example of fitting the outcomes of a coin tossing experiment. The outcome of n trials is a series of n heads and tails, where we can call the outcome "heads" a "success", and "tails" a "failure". For any sequence of the n dichotomous outcomes it is possible to construct a hypothesis after-the-fact that perfectly fits the data. The primary hypothesis here concerns the value of the parameter p – the probability of success on each coin-tossing trial. The standard null hypothesis H_0 is that the coin is "fair" – that p is equal to 0.5 on each coin-tossing trial. Thus, any alternative hypothesis about this parameter can be considered an alternative primary hypothesis. In any event, this is what our imaginary alternative-hypothesis challenger alleges.

Let G(*e*) be some such hypothesis that is constructed so as to perfectly fit data *e*. (G(*e*) is constructed so that P(*e*/G(*e*)) = 1.) To make out the bare-bones of this type of alternative hypothesis, let G(*e*) assert that p, the probability of success, is 1 just on those trials that result in heads, and 0 on the trials that result in tails.[16] It matters not what, if any, story accompanies this alternative hypothesis. Hypothesis G(*e*) says:

G(*e*): p equals 1 on just those trials that were successes, 0 on the others.

The test procedure, let us suppose, is to observe a series of trials, find a hypothesis G(*e*) that makes the result *e* maximally probable, and then pass that hypothesis. In passing G(*e*), the test rejects the null hypothesis H_0 that the coin is fair. The particular hypothesis G(*e*) erected to perfectly fit the data will vary in different trails of our coin-tossing experiment, but for every data set, some such alternative may be found. Therefore, for any experimental result *e*, *e* is taken to fail null hypothesis H_0 and pass the hypothesis G(*e*) – even when G(*e*) is false and H_0 is true (i.e., when the coin *is* "fair"). In a long-run sequence of trials on a fair coin, this test would always

fail to correctly declare the coin fair. Hence, passing G(*e*) with this kind of test procedure is minimally severe.

To calculate severity in cases where the hypothesis is constructed on the basis of data *e*, it is important to see that two things may vary: the hypothesis tested as well as the value of *e*. When the special nature of this type of testing procedure is taken into account, our severity criterion SC becomes:

(SC) There is a very high probability that test procedure T would *not* pass the hypothesis it tests, given that hypothesis is false.[17]

Let the test procedure T be the one just described. The hypothesis that T tests on the basis of outcome *e* is G(*e*). There is no probability that test T would *not* pass G(*e*), even if G(*e*) were false. Hence, the severity is minimal (i.e., 0).

Practically indistinguishable alternatives. What about alternatives that cannot be distinguished from a primary hypothesis H on ground of severity because they differ too minutely from H? As we have already noted, such alternatives are unlikely to count as substantive rival hypotheses. However, suppose in a given context that such an alternative is a substantive rival and yet it cannot be distinguished on grounds of severity. In that case, there are good grounds for criticizing the experimental test specifications (the test was insufficiently sensitive). It is not grounds for methodological underdetermination.

(c) *Empirically Equivalent Alternatives*

We have yet to take up what some might consider the most serious threat to a methodology of testing: the existence of rival primary hypotheses that are empirically equivalent to H, not just on existing experiments but on all possible experiments. In the case where the alternative H′ was said to ask the wrong question, it was possible to argue that the severity of a test of primary hypothesis H is untouched. But the kind of case we are to imagine now is not like that. In this kind of case it is supposed that, although two hypotheses give different answers to the same primary question, they have all of the same

testable consequences. Does it follow that a severity assessment is unable to discriminate between any tests they both pass?

That depends. If it is stipulated that any good test is as likely to pass H, although H′ is true, as it is to pass H′ although H is true – if it is stipulated that any test must have the same error probabilities for both hypotheses – then it must be granted. In that case no severe test can indicate H *as opposed to* H′. The best example is a mathematical one, the two hypotheses being Euclidean and non-Euclidean geometry. Apart from certain, not entirely uncontroversial, cases in physics, however, there is no reason to suppose such pairs of rivals exist in science. But even if we grant the existence of these anomalous cases, this would fail to sustain MUD, which, recall, alleges that the problem exists for *any* hypothesis. There is no reason to suppose every hypothesis has such a rival.

We can go further. When one looks at attempts to give a *general* argument for the existence of such empirically equivalent rivals, one finds that severity considerations serve to discriminate them after all. As with maximally likely alternatives of form G(*e*) above, they too turn out to be "rigged", and if countenanced, lead to highly unreliable test procedures.

This is illustrated in an example offered by Richard Miller (1987) against alleged empirically equivalent, "just-as-good as" alternatives. He asks "What is the theory, contradicting elementary bacteriology, that is just as well confirmed by current data?" (1987, p. 425). Granted, an alternative that can be constructed is "that bacteria occasionally arise spontaneously but only when unobserved". However, the severe testing theory dismisses such an alternative in just the same way it dismisses an alleged parapsychologist's claim that his powers fail to operate when scientists are watching. The tactic followed in constructing these alternative hypotheses allows the alternative to pass the test, but only at the cost of having no chance of failing, even if it is false, i.e., at the cost of adopting a minimally severe test. Such an alternative is a rigged alternative – rigged up so as to make it in principle impossible to distinguish it experimentally from the hypothesis of interest, H. We might define:

> *Rigged Hypothesis R*: a (primary) alternative to H which, by definition, would be found to agree with any experimental evidence taken to pass H.

Consider the general procedure of allowing, for any hypothesis H, that some rigged alternative or other is as warranted as H is. Even where H had repeatedly passed highly severe tests, this general procedure would sanction the argument that all existing experiments were affected in such a way as to systematically mask the falsity of H. That argument procedure is a highly unreliable one. It has a very high (if not a maximal) probability of erroneously failing to discern the correctness of H.

VIII

Let us recapitulate how my account of severe testing deals with alternative hypothesis objections that are thought to be the basis for MUD. The MUD charge (for a method of severe testing T) alleges that for any evidence test T takes as passing hypothesis H severely, there is always a substantive rival hypothesis H′ that test T would regard as having passed equally severely. We have shown this claim to be false, for each type of candidate rival that might otherwise prevent the evidence from genuinely counting in favor of H. Although H′ may accord with or fit the evidence as well as H does, the fact that the two hypotheses can err in different ways and to different degrees shows up in a difference in the severity of the test that each can be said to have passed. The same evidence effectively rules out H's errors to a different extent than it rules out the errors of H′.

This solution rests on the chief strategy associated with my experimental testing approach. It instructs one to proceed in carrying out a complex inquiry by breaking it down into pieces, some of which, at least, will suggest a question that can be answered by means of one of the standard or "canonical" models of arguing from error. In some cases one actually carries out the statistical modeling involved, but in others it is sufficient to carry out an argument from error informally. With regard to the local hypotheses involved in asking specific experimental questions, the task of setting out all possible answers

is not a daunting one. Although it may be impossible to rule out everything at once, we can and do rule out one thing at a time.

Naturally, even if all threats are ruled out, and H is accepted with a severe test, H may be false. The high severity requirement, however, ensures that this erroneous acceptance is very improbable, and that in future experiments, it is probable that the error will be revealed. The thrust of experimental design is deliberately to create contexts which enable questions to be asked one at a time in this fashion. By attempting to talk about data and hypotheses in some general way, apart from the specific context in which the data and hypothesis are generated, modeled, and analyzed to answer specific questions, philosophers have missed the power of such a piecemeal strategy, and underdetermination arguments have been allowed to flourish.

ACKNOWLEDGMENT

Research for this paper was carried out during tenure of an NSF grant, I gratefully acknowledge that support. For useful comments on parts of this paper, I thank Ronald Giere, Valerie Hardcastle, David Hull, Henry E. Kyburg, Jr., Harlan Miller, Wesley Salmon, and an anonymous referee for this journal.

NOTES

[1] Indeed, if two hypotheses entail the evidence, the only way they can be differently confirmed by that evidence by Bayes's theorem is if their prior probability assignments differ.

[2] Nor have attempts to arrive at objective prior probabilities, either through logical probabilities or by other means, met with success.

[3] Speaking of the truth of hypotheses does not commit us to any kind of realism. The truth of H could simply mean that what H says about experimental effects is correct or that specific applications of H will be reliable.

[4] It is part of the test specifications to say how close the evidence must be to what is expected under H in order for the evidence to be classified as fitting H. A minimal requirement is that a fit with H is probable given that H is true, or that it is more probable than under alternatives to H.

[5] The "only if" clause is actually already accommodated by the first requirement of passing a severe test, namely, that the hypothesis H fit the data (see Note 4). I repeat it to ensure misinterpretations are avoided. Equivalently, the fit requirement ensures that when a result is classified as "failing to detect the error" then that result (or one even closer to H) is probable assuming the error is absent. If the fit required is entailment, then if result *e* fits H then P(*e*/H) is maximal.

[6] It is learned that an error is present when a procedure of inquiry, having a very high probability of not detecting an error if none exists, nevertheless detects an error.
[7] A weaker construal requires only that the alternatives say nothing about whether *e* or not-*e* will occur. It is assumed, of course, that H is non-falsified.
[8] This criticism is explicitly raised by Grünbaum (1978).
[9] The term "catchall" hypothesis, I believe, was introduced by L. J. Savage. In the philosophical literature Abner Shimony seems to have been responsible for introducing this term.
[10] An exception is the special case where the truth of a hypothesis can be seen as the outcome of a random trial or game of chance. In all other cases, the only probabilities that a frequentist can assign to a hypothesis are the trivial ones, 1 and 0, according to whether it is true or false, respectively.
[11] Suppes' hierarchy includes canonical models of data, models of experiment, and models of theories. Such a series of models is analogous to the way data are linked to experiments and hypotheses in statistics.
[12] I discuss the details of an early study on oral contraceptives in Mayo (1985 and 1996).
[13] The standard deviation is often estimated from the sample, and in that case it may more properly be called the standard error.
[14] It is important to distinguish carefully between "real" as it is understood here; namely as genuine or systematic, and as understood by various realisms.
[15] I am referring to the technical notion of likelihood.
[16] For example, suppose *e*, the result of four tosses of a coin, is heads, tails, tails, heads. That is, $e = \langle s,f,f,s \rangle$ where s,f abbreviate "success" and "failure" respectively. Then G(*e*) would be: the probability of success equals 1 on trials one and four, 0 on trials two and three. The null hypothesis, in contrast, asserts that the probability of success is 0.5 on each trial.
[17] It is often thought that whenever a hypothesis is constructed to fit the data *e*, and then *e* is used again to test that hypothesis, the result is a test with minimal severity – or no test at all. This definition of (SC) should enable us to see that this need not be so. I discuss this explicitly in Mayo (1991 and 1996).

REFERENCES

Earman, J. (1992), *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.
Grünbaum, A. (1978), "Popper vs. Inductivism", in G. Radnitzky and G. Andersson (eds.), *Progress and Rationality in Science*. Boston Studies in the Philosophy of Science, Vol. LVIII. Dordrecht: Reidel, pp. 117–142.
Hacking, I. (1983), *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.
Laudan, L. (1990b), "Demystifying Underdetermination", in C. W. Savage (ed.), *Scientific Theories*. Minnesota Studies in the Philosophy of Science, Vol. XIV. Minneapolis: University of Minnesota Press, pp. 267–297.
Laudan, L. (1995), *Beyond Positivism and Relativism*. Boulder: Westview Press.

Mayo, D. (1985), "Increasing Public Participation in Controversies Involving Hazards: The Value of Metastatistical Rules", *Science, Technology, and Human Values* 10: 55–68.
Mayo, D. (1988), "Brownian Motion and the Appraisal of Theories", in A. Donovan, L. Laudan and R. Laudan (eds.), *Scrutinizing Science*. Dordrecht: Kluwer, pp. 219–243.
Mayo, D. (1991), "Novel Evidence and Severe Tests", *Philosophy of Science* 58: 523–552.
Mayo, D. (1996), *Error and the Growth of Experimental Knowledge*. Series in Conceptual Foundations of Science. Chicago: The University of Chicago Press.
Miller, R. (1987), *Fact and Method: Explanation, Confirmation and Reality in the Natural and the Social Sciences*. Princeton: Princeton University Press.
Salmon, W. (1990), "The Appraisal of Theories: Kuhn Meets Bayes", in A. Fine, M. Forbes, and L. Wessels (eds.), *PSA 1990*, Vol 2, East Lansing: Philosophy of Science Association, pp. 325–332.
Suppes, P. (1962), "Models of Data", in E. Nagel, P. Suppes and A. Tarski (eds.), *Logic, Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress*. Stanford: Stanford University, pp. 252–261.

Department of Philosophy
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0126
USA