Deborah Mayo[1]

Learning from Error: The Theoretical Significance of Experimental Knowledge

INTRODUCTION

We learn from our mistakes.

Common though this slogan is, epistemologists of science have given it short shrift. At most, learning from error refers to deductive falsification of the Popperian variety: If a hypothesis is put to the test and fails, we reject it:

H entails O,

Observe ~O.

What is learned is that *H* is false.

By contrast, the *threats of error* are thought to make it difficult if not impossible to implement even a simple *modus tollens*: (1) A predicted observation O is itself a "hypothesis" derived only with the help of "theory laden" auxiliary hypotheses whose own reliability may be questioned. (2) Even if the hypothesized anomaly ~O is assumed sound, the inability to distinguish among the possible sources responsible prevents *H* from being falsified (Duhemian problem). Further, (3) hypothesis *H* typically entails a claim about the probability of an outcome; thus the anomaly (~O) does not contradict *H*, and a statistical falsification rule is required.

But even if all these threats were circumvented, learning merely that there is a flaw in *H* would scarcely capture the truth behind the cliché that we learn from our mistakes: certainly there is something more substantial underlying this intuition. After all, (4) even if *H* is legitimately falsified, experimental evidence in and of itself fails to point us to replacements that are in some sense more correct or less error prone. Yet the reason we intuitively value learning from mistakes is that being forced to reorient ourselves when our claims clash with the world offers a powerful source of objective knowledge. It is this valuable self-correcting activity that would take center stage in an adequate epistemology of experiment.

A key rationale for the "new experimentalism," at least in my view, is to see how far one can go to solving these problems by taking seriously the nitty-gritty

details of low level methods for collecting, modeling, and drawing inferences from experiments. This in turn would require a serious look at the local experimental work of distinguishing effects, ruling out artifacts, and pinpointing blame for anomalies. Merely setting out idealized concepts of evidence will not do; we need to show how to tackle questions such as:

- What are the different types of error, and how do we locate them?
- How does learning from error lead to the growth of experimental knowledge?
- How is the growth of experimental knowledge related to developing and appraising scientific theories?

In the experimental account that I develop, the growth of knowledge proceeds by inferring claims insofar as they pass probative or *severe* tests—tests that would have unearthed a specific error in, or discrepancy from, a hypothesis H, were H false. The spotlight therefore is on those methods that serve as reliable error probes; and a context is deemed "experimental" (whether or not literal manipulation is present) insofar as we are able to assess and control the severity or error-probing capacities of tools. Since, in statistics, the probability that a method would unearth discrepancies and flaws is called an error probability, we may dub this general approach, the *error statistical* approach, even though it is not limited to formal contexts.

The focus of this paper is to elucidate the error-statistical account in the context of appraising, and also building, large-scale theories. Even those who grant that the attention to experiment is "a useful corrective to some of the excesses of the theory-dominated approach" (Chalmers 1999, 206) have seriously shortchanged the reach of experimental knowledge in its efforts to answer these questions.[2] *Even where experimental data do not warrant inferring large-scale theories, the theoretical significance of experimental knowledge, I argue, is at the heart of learning about the world. Considering why, at a given stage, a large-scale theory has not, as a whole, passed severely is crucial to discovering new theories and to designing tests to try next.* The upshot is a more satisfactory and far more dynamic account of large-scale theories*!*

A "life of theory" adequate to the task of supplementing the "life of experiment" would need to go beyond a retrospective sum-up of scientific episodes toward a forward-looking account of the discovery/invention/testing of new theories. (It is not just a miracle, as some claim.)

- It should give insights as to how to discriminate which parts of a theory have and have not been warranted.

This tells the researcher which portions of a theory would be safe to use in probing new domains, as well as what claims, if relied on, would bias and misdirect the growth of knowledge. It does not suffice to reconstruct the problem in terms of

which large-scale theory to accept, or which hypothesis to accord the comparatively highest degree of belief.

- It should account for the stability of experimental effects through theory change, and
- It should capture statistical testing,

which actually is always required even with nonstatistical theories.

Our issue, let me be clear, is not about whether to be "realists" in regard to theories in any of the senses in which this is understood. Allegations that new experimentalists are antirealists, I claim, distract from the core problems that need to be overcome regardless of one's position on the realist debate: how to warrant reliable scientific knowledge. I will try to keep to language that realists and nonrealists alike may use.

2. TAKING EVIDENCE SERIOUSLY

Appealing to the nitty-gritty of experimental practice, it was hoped, would restore the role of empirical data as an objective constraint on inquiry. Powerful experimental arguments could supply robust tests of hypotheses that avoid threats of theory-laden data and provide a stable ground through theory change. Despite these promises, philosophy of science is plagued by general self-doubt, even among philosophers of science I most admire. Achinstein still recounts his "Dean's problem."

2.1 ACHINSTEIN'S DEAN'S PROBLEM

When asked by a skeptical Dean about the relevance of philosophy of science for science, Achinstein conceded that "scientists do not and should not take...philosophical accounts of evidence seriously" (2001, 9) because they (i) make it too easy to have evidence, and (ii) are based on a priori computations; whereas scientists evaluate evidence empirically. Alan Chalmers (1999) similarly claims that "scientists...are not in need of advice from philosophers" (252), whose only hope of generality is limited to "trivial platitudes" such as "take evidence seriously" (171). Examples could be multiplied. The new generation of philosophers of science immerse themselves in fascinating domains of experimental practice, from statistics to cognitive science to biology, but they too rarely seek a general and normative account of evidence that would do justice to their case studies. The thinking is that if it is empirical, then it is best left to the scientists, but this is a mistake: *scientific inference is too important to be left to the scientists!*

2.2 PRINCIPLES FOR SCIENTIFIC EVIDENCE

We might begin by taking seriously Chalmers' platitude to "take evidence seriously." A method does not take evidence seriously in evaluating a claim H, if it fails utterly to discriminate whether H is correct or incorrect. For example, we would deny that a data set $\boldsymbol{x}$ is evidence for some claim H if the observed data are not in any way related to H, such as using data $\boldsymbol{x}$ on the average weight loss in rats treated with the latest anti-obesity drug as grounds for an inference to H: the mean deflection effect in relativistic gravity. It would be misleading even to say that the observed weight loss in rats, $\boldsymbol{x}$, "agrees with" or "fits" hypothesis H unless H explains, entails, or is in some sense rendered more expected under the supposition that H is true rather than false. To regard $\boldsymbol{x}$ as "fitting" H even though the data are just as probable whether or not H is true (or worse, if $\boldsymbol{x}$ is even less probable under H than not-H) would be to give an inadequate account of what it means for data to fit a hypothesis.

But satisfying this minimal "fit" criterion scarcely amounts to taking evidence seriously. One is not taking evidence seriously in appraising hypothesis H if, either through selective searches or deliberate constructs, so good a fit with H is assured even if H is false. A drug company that refused to construe repeated observations of lack of weight loss in rats as indicative of the ineffectiveness of their obesity drug would be following a procedure with little or no chance of inferring the *in*effectiveness of their drug. Similarly, if researchers selectively reported only those data $\boldsymbol{x}$ that showed weight loss, ignoring the other data, we would deny that they had provided evidence for the effectiveness of their drug.

We deny that data $\boldsymbol{x}$ is evidence for H if, although $\boldsymbol{x}$ fits H, the inferential procedure had very little capacity of providing a worse fit with H, even if H is false. Such a test, we would say, is insufficiently stringent or severe.

Although one can typically take one's pick in criticizing purported evidence—deny that there is an adequate fit, or show that the observed fit is easy to achieve even if H is false—it is useful to have them both on hand for scrutinizing evidence. These two minimal conditions—others could be supplied as well—might be seen as too obvious to bear explicit notice. But ruling out such flagrantly unserious treatments of evidence, I claim, already lets us make progress with some of the most skeptical doubts about evidence.

3. SEVERITY PRINCIPLE: SOME QUALIFICATIONS

From the considerations in section 2, we arrive at what may be called the severity principle.

***Severity Principle (Weak)*:** An accordance between data $\boldsymbol{x}$ and H provides poor evidence for H if it results from a method or procedure that has little or no ability of finding discordant data, even if H is false.

Another way to talk of $\boldsymbol{x}$ "fitting" or "being in accord" with H is by saying that the data $\boldsymbol{x}$ enable a hypothesis H to pass a test (or to pass with a score corresponding to how well it agrees with H). But in order for the test that hypothesis H passes to be severe, so good a fit must not be something that is easy to achieve, i.e., probable, even if H is false. As weak as this is, it is stronger than a mere falsificationist requirement: it may be logically possible to falsify a hypothesis, whereas the procedure may make it virtually impossible for such falsifying evidence to be obtained. Although one can get considerable mileage by going no further than this negative conception (as perhaps Popperians would), we will continue on to the further positive conception, which will comprise the full severity principle:

***Severity Principle (Full)*: Data $\boldsymbol{x}$ provide a good indication of or evidence for hypothesis H (just) to the extent that H passes experimental test E severely with $\boldsymbol{x}$.[3]**

3.1 SOME QUALIFICATIONS MUST BE KEPT IN MIND.

First, a severity assessment is a function of a particular set of data or evidence $\boldsymbol{x}$ and a particular hypothesis or claim H.

Severity has three arguments: a test E, a result $\boldsymbol{x}$, and an inference or a claim H. "The severity with which H passes test E with data $\boldsymbol{x}$" may be abbreviated by:

SEV(Test E, data $\boldsymbol{x}$, claim H).

When $\boldsymbol{x}$ and E are clear, we may write SEV(H). Setting out a test E calls for its own discussion, which I put to one side here.[4]

Defining severity in terms of three arguments is in contrast to a common tendency to speak of "a severe test" divorced from the specific inference at hand. This common tendency leads to fallacies we need to avoid. A test may be made so sensitive (or powerful) that discrepancies from a hypothesis H are inferred too readily. However, the severity associated with such an inference is *decreased* the more sensitive the test (not the reverse). Suppose that any observed weight decrease, regardless of how small, is taken to signal evidence for

H: Drug x results in weight loss (in rats).

H would be inferred with low severity. On the other hand, if no observed difference is found with such a sensitive test, high severity could be accorded to the denial of H, which we may write as the null hypothesis H_0:

H_0: The drug fails to result in weight loss.

Or perhaps, it may set an upper bound:

H_0: Any weight loss due to this drug is less than δ.

Hypothesis H_0, in this second form, would pass with high severity insofar as the test had a very high probability of detecting a weight loss greater than δ and yet no such loss was observed.

The main point here is that since the same experimental test *E* will have some hypotheses passing severely and others inseverely, we are prohibited from speaking generally about a test being severe. We must specify the hypothesis being considered for a severity assessment, as well as the data $\boldsymbol{x}$.[5] Hypothesis *H* is generally a claim about some aspect of the data-generating procedure in the experimental test *E*.

3.2 TWO OTHER QUALIFICATIONS.

Second, although it is convenient to speak of a severe experimental test *E*, it should be emphasized that test *E* actually may, and usually does, combine several tests and inferences together; likewise, data $\boldsymbol{x}$ may combine observed results of several experiments. So long as one is explicit about the test *E* being referred to, no confusion results. The third point I want to make is that I will use testing language even for cases described as estimation, because any such question can be put in testing terms. Still, I am not assuming some classic conceptions of "testing hypotheses" such as the assumption that hypotheses to be inferred must be set out pre-data. Fourth, that "*H* is severely tested" will be understood as an abbreviation of the fact that *H* has *passed* the severe or stringent probe, not, for example, merely that *H* was subjected to one. Data $\boldsymbol{x}$ provides a good indication of or evidence for hypothesis *H* if and only if $\boldsymbol{x}$ results from a test procedure that would have, at least with very high probability, uncovered the falsity of, or discrepancies from *H*, and yet no such error is detected. To encapsulate:

Hypothesis H passes a severe test E with $\boldsymbol{x}$ *if,*

(i) $\boldsymbol{x}$ agrees with or "fits" *H,* and,

(ii) test *E* would have (with high probability) produced a result that fits *H* less well than $\boldsymbol{x}$ does, if *H* were false or incorrect.

While this affords a succinct summary, it should be regarded as merely a placeholder for the real-life, flesh and blood, arguments from severity that are the cornerstone of experimental learning. The multiplicity of methods and standards that lead many philosophers to view themselves as pluralists about evidence are better seen as diverse ways to satisfy or appraise severity criteria (of course different terms may be substituted).

4. THE SEVERITY REQUIREMENT: TWO EXAMPLES

I consider two examples, the first very informal and the second alluding to the arena of theory testing to be considered in the second half of this paper.

EXAMPLE 1. *My weight.* To move away from weighing rats, suppose I am testing whether I have gained weight—and if so how much—between the day I left for London and now using a series of well-calibrated and stable weighing methods. If no change registers on any of these scales, even though, say, they easily detect a difference when I lift a standard one-pound potato, then we may regard this as grounds for inferring that my weight gain is negligible within the limits set by the sensitivity of the scales.

H : My weight gain is no greater than δ,

where δ is an amount easily detected by these scales. H, we would say, has passed a severe test: Were I to have gained δ pounds or more (i.e., were H false), then this method would almost certainly have detected this (Mayo and Cox 2010).

If the scales work reliably on test objects with known weight, what extraordinary sort of circumstance could systematically make them all go astray only when I do not know the weight of the test object: Can the scales read my mind? I could keep claiming that all the scales are wrong—they work fine for vegetables with known weights, but conspire against me when the weight is unknown—but this tactic would keep me from correctly finding out about weight. It is the learning goal that precludes discounting results based on conspiracies and other "rigging" gambits.

EXAMPLE 2: *Data on light bending as tests of the deflection effect λ given in Einstein's gravitational theory (GTR).* Data based on very long baseline radio interferometry (VLBI) in the 1970s teaches us much more about, and provides much better evidence for, the Einsteinian predicted light deflection (now often set at 1) than did the passing result from the celebrated 1919 eclipse test. The interferometry tests are far more capable of uncovering a variety of errors and discriminating among values of the deflection λ than the crude eclipse tests were. The results set more precise bounds on how far a gravitational theory can differ from the GTR value for λ.

When we evaluate evidence along the lines of these examples, we are scrutinizing inferences according to the severity of the tests they pass.

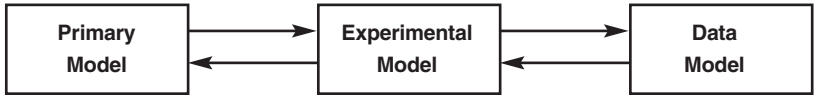
Except for formal statistical contexts, "probability" in defining severity may serve merely to pay obeisance to the fact that all empirical claims are strictly fallible. We would infer that my weight gain does not exceed such and such amount; without any explicit probability model.

Indeed, the most forceful severity arguments usually do not require explicit reference to probability or statistical models. So it would be incorrect to dismiss this approach by claiming that scientists do not always use explicit statistics, or by noting that scientists were learning even before modern statistics was developed. It is really only in the more difficult and less blatant cases that we appeal to formal statistical tests to ensure that errors will be correctly detected (i.e., signaled) with high probabilities.

Nowadays, we hear of the stress testing of banks to indicate how well they would survive various economic disruptions. We may criticize a test as too readily giving a pass to Bank B, if banks on the brink of collapse are given high passing scores; there is no need to appeal to a formal probability model.

5. MODELS OF INQUIRY, MODELS OF ERROR

This account of testing cannot abide oversimplifications of accounts that begin with statements of evidence and hypotheses. Such accounts overlook the complex series of models required in inquiry, stretching from low-level theories of data and experiment to high-level hypotheses and theories. To discuss these different pieces, questions, or problems, we need a framework that lets us delineate the steps involved in any realistic experimental inquiry and lets us locate the necessary background information: how much more so when attempting to relate low-level experimental tests to high-level theories, as is my focus here.



To organize these interconnected pieces, let us view any given inquiry as involving a *primary question* or *problem* that is then embedded and addressed within one or more other models, which we may call "experimental." *Secondary* questions would include a variety of inferences involved in probing answers to the primary question (e.g., How well was the test run? Are its assumptions satisfied by the data in hand?). The primary question may be investigated by means of properly modeled rather than "raw" data. Only then can we adequately discuss the inferential move (or test) from the data (data model) to the primary claim H (through the experimental test E).

Take the interferometric example. The primary question, determining the value of the GTR parameter, λ, is couched in terms of parameters of an astrometric model M that (together with knowledge of systematic and nonsystematic errors and processes) may allow raw data, adequately modeled, to estimate parameters in M to provide information about λ (the deflection of light). Having passed, with severity, a hypothesis about the value of parameter λ, the assessment is not altered by the introduction of new theories, say T' and T'', that agree with respect to λ, even if T' and T'' disagree as regards the explanation for the severely affirmed values. Theories T'

and T'', we would say, *are not rivals* so far as the effect of or hypothesis about λ, no matter how much they differ from each other in their full theoretical frameworks.

Experimental knowledge can be, and often is, the basis for inferring theoretical hypotheses. When I speak of the ways a hypothesis H may be false, I am including erroneous claims about underlying causes and mistaken understandings of any testable aspect of a phenomenon of interest. I am not drawing distinctions between "experimental" and theoretical, as some might. Often the parameter in a statistical model directly parallels the theoretical quantity in a substantive theory or proto-theory.

Although there are myriad types of mistakes in inference, I propose that there are only a handful of error types and strategies for checking, avoiding, and learning from them. I term these "canonical" errors:[6]

1. mistaking chance effects or spurious correlations for genuine correlations or regularities
2. mistakes about a quantity or value of a parameter
3. mistakes about a causal factor
4. mistakes about the assumptions of the data for the experimental inference
5. mistakes about the theoretical significance of the experimental inference

There is a corresponding localization of what one is entitled to infer severely: To infer H is to infer the absence of the particular error that H is *denying*. Note that to falsify H is to pass not-H. Nothing turns on this particular list, it might be condensed or expanded; I simply find it a useful taxonomy for experimental strategies now in use.

6. SOME CONTRASTS WITH FAMILIAR ACCOUNTS

Viewing experimental inference in terms of severe testing is a departure from familiar philosophical accounts with respect to scientific aims, methods, and knowledge. I identify two main contrasts.

6.1 SKEPTICAL ASSUMPTIONS ABOUT RELIABLE EVIDENCE

It is commonly supposed that evidence claims are only as reliable as are the intermediary inferences involved. But this is false. We can obtain rather accurate claims from far less accurate ones. Individual measurements may each be highly inaccurate, while the estimate inferred is far more accurate (e.g., based on tech-

niques of averaging, and by varying assumptions). This allows inductive inference—understood as severe testing—to be genuinely ampliative (achieve lift-off!)

For example, although the "inferred" deflection of light required piling on one inference after the other, the inferred result was far more reliable than any of the "premises." In the 1970s, each yearly estimate of λ using quasars was scarcely more accurate than the eclipse tests of 1919, but multiple years taken together resulted in a highly accurate inference. As knowledge of interferences grows, scientists learn to subtract them out, as they did with the corona effect in the 1970s. Better still, they learned from these errors how sufficiently to amplify the deflection effect so that it can be discerned with precision anywhere in the sky. Using intercontinental quasar observations made to monitor the Earth's rotation, we get to .1% accuracy, rather than the 10 or 20% accuracy of earlier experiments, and this precision increases every year.

6.2 ASSUMPTIONS ABOUT SCIENTIFIC INFERENCE AND PROGRESS

In the experimental account I favor, the growth of knowledge depends neither on probabilifying nor on "rationally accepting" large scale theories or paradigms. Instead, we learn to test specific hypotheses in such a way that there is a good chance of learning something—whatever theory it winds up as part of (Mayo 2010a, 28). The role of probability (which we only assign to events or outcomes, not hypotheses) is *not* to assign degrees of confirmation or belief to hypotheses, but to characterize how frequently methods are capable of detecting and discriminating errors: these are called error frequencies or *error probabilities.*

Error probabilities, whether informally arrived at, or derived from formal statistical models, may be used to arrive at severity assessments. I call any such account an *error-statistical account based on the idea of severity*. Two pieces of data that equally well fit (and thus "confirm") a hypothesis, according to existing measures of confirmation, may differ greatly in their evidential value due to differences in the probativeness of the tests from which they arose. This is an important way of addressing underdetermination worries. Even if rival hypotheses and theories may be found equally to fit the data $\boldsymbol{x}$, they will not be equally well tested by $\boldsymbol{x}$. If hypothesis H passes with severity, then no (genuine) rivals to H can pass severely (Mayo 1997b). (I will return to the question of rival hypotheses at a given "level" of inquiry.)

At minimum, a claim of experimental knowledge is a claim about what outcomes or events would occur with specified probabilities were a given experiment carried out. Since we can reliably check if the experimental assumptions are satisfied, the methods offer a reliable basis for "ampliative" or inductive inference.

Although the idea of using probability to evaluate well testedness has its own tradition, in the work of C.S. Peirce, Popper and a few others, philosophers are

generally wedded to the idea that the role for probability must be to assign a degree of confirmation determined by conditional probability and Bayes's theorem. Since I discuss those approaches at length elsewhere, I will provide only a brief summary in section 7 of the current state of play in Bayesian statistics.

7. HIGHLY PROBABLE VERSUS HIGHLY PROBED HYPOTHESES

Some philosophers profess not to understand what I could be saying if I am prepared to allow that a hypothesis H has passed a severe test T with $\boldsymbol{x}$ without also advocating (strong) belief in H, understood Bayesianly. "If Mayo refuses to assign a posterior probability to H...then I...would have a problem understanding what passing a severe test has to do with...providing a good reason to believe H'' (Achinstein 2010, 183).

The truth is that a high posterior probability (in any of the ways it may be obtained) is neither necessary nor sufficient for passing severely in my sense. Interestingly, when Achinstein looks to Mill and Newton as exemplars for his account of evidence, he discovers that neither fits the Bayesian mold of assigning a probability to an inductively inferred hypothesis (e.g., Achinstein 2010, 176). He notices that they speak of claims being approximately true, but not probably true. Yet rather than concede this as casting doubt on his probabilism, Achinstein goes to great lengths to imagine that what Mill says about the probability of events is also meant to apply to hypotheses! (Mayo 2005, 2010b; 2011).

7.1 THE CURRENT STATE OF PLAY IN STATISTICAL PRACTICE

Formal epistemologists should be aware that in current statistical practice, the program for getting scientists to assign priors by posing a series of bets has been largely rejected. Bayesian practitioners instead appeal to one or another form of "default" prior defined by convention. The central idea that prior probabilities are expressions of uncertainty or degrees of belief has gone by the board. The priors are conventions that may not even be probabilities (they are often improper); rather than reflect knowledge prior to data, the default priors are model-dependent. Given a Bayesian report of a high posterior degree of belief, say .95, in a hypothesis H, the severe tester always demands to know: how often would such a high assignment occur even if H were false? While in special cases Bayesian methods control long-run relative frequencies of error, in general the ability to assess and control the severity associated with particular inferences is absent (Mayo and Cox 2010; Mayo and Spanos 2011).[7] In any event, it has not been demonstrated.

7.2 PROBABILITY LOGIC IS THE WRONG LOGIC FOR INDUCTIVE INFERENCE

Moreover, probability logic seems to be the wrong logic for scientific inference. When SEV*(H)* is high there is no problem in saying that ***x*** warrants *H*, or, if one likes, that ***x*** warrants believing in *H*, even though that would not be the direct outcome of a statistical inference. The reason it is unproblematic in the case where SEV*(H)* is high is:

If SEV*(H)* is high, its denial is low, i.e., SEV*(~H)* is low.

But it does not follow that a severity assessment should obey the probability calculus, or be a posterior probability—it should not, and is not.

After all, a test may poorly warrant both a hypothesis *H* and its denial, violating the probability calculus. That is, SEV*(H)* may be low because its denial was ruled out with severity, i.e., because SEV*(~H)* is high; but SEV*(H)* may also be low because the test is too imprecise to allow us to take the result as good evidence for *H*.

Even if one wished to retain the idea that degrees of belief correspond to (or are revealed by?) the bets an agent is willing to take, this still would not have shown the relevance of a measure of belief to the objective appraisal of what has been learned from data. Even if I strongly believe a hypothesis, I will need a concept that allows me to express whether or not the test with outcome ***x*** warrants *H.* That is what a severity assessment would provide.[8] In this respect, a dyed-in-the-wool subjective Bayesian could in principle accept the severity construal for science, and still find a home for his personalistic conception.

Scientific inference is about obtaining new knowledge; we do not want a logic that just to get things started requires delineating all possible hypotheses and probability assignments to them. Moreover, this kind of closed system is at odds with learning in science.

7.3 TACKING PARADOX IS SCOTCHED.

Severity logic avoids classic problems facing both Bayesian and hypothetical-deductive accounts in philosophy. For example, tacking on an irrelevant conjunct to a well-confirmed hypothesis *H*, on these accounts, seems magically to allow confirmation for some irrelevant conjuncts. Not so in a severity analysis.

If SEV(Test *E*, data ***x,*** claim *H*) is high, but *J* is not probed in the least by the experimental test *E*, then SEV (*E*, ***x***, (*H* & *J*)) is very low or minimal.

For example, consider:

H: GTR and *J*: drug Y causes weight loss in rats,

and let data $\boldsymbol{x}_0$ be a value of the observed deflection in accordance with the general theory of relativity, GTR. The two hypotheses do not refer to the same data mod-

els or experimental outcomes, so it would be odd to conjoin them; but if one did, the conjunction gets minimal severity from this particular data set. Note that we would distinguish *H* severely passing by dint of data ***x,*** and its severely passing based on all evidence in science at a time.

A severity assessment distinguishes the well testedness of a portion or variant of a larger theory; it *partitions* the theory. It directs us to exhaust the space of alternatives to any claim to be inferred. These are the "relevant" rivals to *H*—they must be at "the same level" as *H*.

8. IS SEVERITY TOO SEVERE?

Some critics maintain that the severity requirement is too severe, alleging that scientists accept a theory even if it has passed inseverely, so long as it is the "best tested" so far.

8.1 AGAINST COMPARATIVISM

This position—I dub it "comparativism"—is often attributed to Popper. Though Popper did at times suggest that accepting (or preferring) the best-tested theory is the very definition of scientific rationality, this was one of his weakest positions. Accepting a theory *T* as a whole, when it is given that *T* is inseverely tested, is at odds with the classic Popperian demand that only claims that have survived "serious" tests of error have the right to be accepted, even provisionally. The comparativist-testing procedure licenses the inference from passing hypothesis *H* (perhaps severely in its own right) to inferring all of *T*—but this is a highly *un*reliable method. "Best tested" is not only relative to existing theories but to existing tests: they may all be poor tests for the inference to *T* as a whole. Even adding a requirement that the predictions are novel, or sufficiently varied to all be coincidental, do not yield what is required (Mayo 1996, 1997b; Chalmers 2010).

8.2 DENYING THAT NO THEORIES CAN PASS WITH SEVERITY

The most serious complaint I have received about why I should lower the bar for what may be inferred with severity asserts that no theory or generalization can satisfy my stipulation! According to Chalmers, "theories, not just high-level theories but theories in general, cannot be severely tested in Mayo's sense....Low-level experimental laws cannot be severely tested either" (Chalmers 2010, 62). Musgrave (2010), who takes himself to be following Chalmers (1999, 2010) and Laudan (1997), appears to concur. Why? First, they say, because theories transcend empirical evidence and thus could be wrong. But high severity never

demanded infallibility! Their second reason is based on adducing what I call "rigged alternatives."[9] After all, they say one is always free to argue that any hypothesis *H*, however well probed, is actually false, and that some unknown (and unnamed) rival is responsible for our repeated ability to generate results that *H* passes.

> *Rigged hypothesis H**: An (unspecified) rival to (primary) hypothesis *H* that by definition would be found to agree with any experimental evidence taken to pass *H*.

The fact that one is always free to a rigged hypothesis scarcely prevents me from discrediting the gambit. Such an argument has a high if not a maximal probability of erroneously failing to discern the correctness of *H*, even where *H* is true. This general procedure would always sanction the argument that all existing experiments, however probative, were affected in such a way as to mask the falsity of *H*. Even if one cannot argue in a particular case that *H* has passed severely, the error-statistical tester can condemn this general underdetermination gambit.

Like C.S. Peirce, the current account need only assume that "the supernal powers withhold their hands and let me alone" when I apply and test error correcting methods (CP 2.749), Much as in the case of my scales, I deny they read my mind and conspire to trick me just when the weight of the object is unknown. Whenever it can be shown that the skeptic's position reduces to "an argument from conspiracy," it is discounted by the error statistician—as it can be shown to preclude correctly inferring a claim *H*, even if true.

Philosophers of science, perhaps unlike analytic epistemologists, do not question the success of science; the difficulty is in explaining its success, as well as how we might acquire more of the kind of knowledge we already have, more quickly and efficiently.

8.3 ALTERNATIVE HYPOTHESES OBJECTIONS

Without being guilty of the blanket "rigged" alternative, a critic may claim there is always a specific rival to *H* that can be erected to accommodate or fit the data. Two features of the severity account prevent this: First, the test must be designed to ask a specific question so that *H* and its alternative(s) exhaust the space of possibilities. Second mere fitting is not enough, and we can distinguish the severity of hypotheses passed by considering the reliability of the accommodation method.[10] Admittedly, this demands examining the detailed features of the data recorded (the data models). It sounds plausible to say there can always be some rival—when that rival merely has to "fit" already known experimental effects. Things are very different if one takes seriously the constraints imposed by the information in the detailed data, coupled with the need to satisfy severity (Mayo 1997b).

8.4 A BASKET OF OTHER CHARGES

Some argue that unless scientists accept whole theories even when only subsets are severely tested, they cannot draw out testable predictions. But assuming, hypothetically, one or another theory for the purposes of testing is scarcely to accept the theory! Then there is the charge that my account precludes the inevitable appeal to background theories in the course of testing some primary theory *T*. Not at all. I only require that those assumptions not introduce bias. Fortunately, this can be satisfied by numerous experimental strategies: for example, that the background theories are themselves severely tested, and that enough is known to subtract out any errors introduced by uncertain background claims. Where such defenses are unavoidable, the scientist must report those assumptions that prevent various claims from passing with severity. Far from obstructing progress, pinpointing potential threats to severity is a crucial source of progress for the error-statistical tester!

Finally, there is nothing that precludes the possibility that so-called low-level hypotheses *could* warrant inferring a high-level theory with severity. Some suppose that inferring *T* with severity would demand that we slog through all rival theories and eliminate them one at a time. That is not an astute way to proceed. Large shake-ups, even on the order of paradigm changes, may result from local effects affirmed with severity. Even GTR, everyone's favorite example, it is thought, predicts a unique type of gravitational radiation, such that affirming that particular "signature" with severity would rule out all but GTR (in its domain). In the remainder of this paper I consider some of the ways these issues have arisen concerning the theoretical significance of research in experimental relativity.

9. EXPERIMENTAL GRAVITATION

The case of GTR has figured in challenging this account of severity—one reason I delved in it further. Take John Earman (1993):

> When high-level theoretical hypotheses are at issue, we are rarely in a position to justify a judgment to the effect that [such a passing result is improbable under the assumption that *H* is false]. 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…does not contain alternatives to GTR that yield the same prediction for the bending of light as GTR. (Earman 1993, 117)

Far from posing a problem for the severity account, this is just what an adequate account should say, since all of GTR had scarcely passed with severity in 1918 and 1919!

Moreover, this situation does not preclude piecemeal learning with severity. "*H* is false"—as it enters a severity assessment— is not a general *catchall* hypothesis, but refers to specific errors or discrepancies from a hypothesis *H*. Even large-scale theories, when we have them, are applied and probed only by a piecemeal testing of local hypotheses. Rival theories need to be applicable to the same data models, particularly if *one is to be a possible replacement for the other.*[11]

9.1 FOUR PERIODS OF EXPERIMENTAL GTR

Experimental testing of GTR nowadays is divided into four periods: 1887-1919, 1920-1960, 1960-80, and 1980 onward. Following Clifford Will, they may be called the periods of *genesis, stagnation, golden era*, and *strong gravity*. The first period, that of *genesis,* encompasses experiments on the foundations of relativistic physics: the Michelson Morley and the Eotvos experiments, as well as the so-called "classical tests" of GTR on the deflection of light and the perihelion of Mercury. Through the second period, 1920 to 1960, GTR enjoyed its status as the "best tested" theory of gravity, while in the third period, 1960 to 1980, a veritable zoo of rivals to GTR were erected, all of which could be constrained to fit these classical tests.

The large-scale theory tester might regard the latter period as one of crisis and uncertainty, but in practice it is widely hailed as "the golden era" or "renaissance" of GTR.

It is the earlier period, when GTR was the best tested, that is bemoaned as one of "stagnation" or at least "hibernation." The dearth of linkups between the very mathematical GTR and experiment made it an unappealing research area. Clifford Will, one of the founders of experimental relativity, describes how young researchers then were routinely told to look elsewhere for promising work. Only when the experiments of 1959 and 1960 enabled confronting GTR's predictions in new ways did the golden age of GTR ensue. The goal of the testing, however, was not to decide whether GTR, in all its implications, was correct, but was rather, in the first place, to learn more about GTR (what does it really imply about experiments we can perform?), and in the second place, to build models for phenomena that involve relativistic gravity: quasars, pulsars, gravity waves, and such. The goal was *to learn more about gravitational phenomena*.

9.2 THE PARAMETERIZED POST-NEWTONIAN (PPN) FRAMEWORK

Far from arguing that we accept all of GTR before its time, our severe tester would explore just *why* it would be wrong to regard GTR—as a whole—as well

tested, in all the arenas in which gravitational effects may occur. This turns on distinguishing between those portions of GTR that were and those that were not well tested. Even without full-blown alternative theories of gravity in hand we can ask (as they did in 1960): *how could it be a mistake to regard the existing evidence as good evidence for GTR (both within the solar system and beyond)*?

To this end experimental relativists developed a kind of *theory of theories* for delineating and partitioning the space of alternative gravity theories, called the Parameterized Post Newtonian (PPN) framework. The PPN framework was deliberately designed to prevent researchers from being biased toward accepting GTR prematurely (Will 1993, 10), while allowing them to describe violations of GTR's hypotheses—discrepancies with what it said about specific gravitational phenomena. It set out a list of parameters that allowed describing systematically violations of GTR's hypotheses. "The PPN framework takes the slow motion, weak field, or post-Newtonian limit of metric theories of gravity, and characterizes that limit by a set of 10 real-valued parameters. Each metric theory of gravity has particular values for the PPN parameters" (Will 1993, 10).

Parameter	What it measures relative to GTR	Values in GTR
λ	How much space-curvature produced by unit rest mass?	1
β	How much "non-linearity" in the superposition law for gravity?	1
ξ	Preferred location effects?	0
α_1	Preferred frame effects?	0
α_2		0
α_3		0
α_3	Violation of conservation of total momentum?	0
ζ_1		0
ζ_2		0
ζ_3		0

Table 1: *The PPN Parameters and their significance. Adapted from C. Will 2005.*

9.3 HIGH-PRECISION NULL-HYPOTHESES EXPERIMENTAL TESTS

The PPN framework permitted researchers to compare ahead of time the relative merits of various experiments that probed the solar system approximation, or solar system variant, of GTR. Appropriately modeled astronomical data supplied the "observed," i.e., estimated, values of the PPN parameters, which could then be compared with the different values hypothesized by the diverse theories of gravity.

In relation to the series of models, the PPN framework gets all the candidates for relativistic theories of gravity to be talking about the same things and to connect to the same models of data and experiment. This allows measuring, i.e., *inferring*, the values of PPN parameters by means of complex, statistical least squares fits to parameters in models of data. In general, the GTR value for the PPN parameter under test serves as the *null hypothesis* H_0 from which discrepancies are sought. For example, for the deflection parameter λ:

$$H_0: \lambda = \lambda_{\text{GTR}}$$

By identifying the null with the prediction from GTR, any discrepancy has a very good chance of being detected, so if no significant departure is found, this constitutes evidence for the GTR prediction with respect to the effect under test, i.e., λ.

Without warranting an assertion of zero discrepancy from the null GTR value (set at 1 or 0), the tests are regarded as ruling out GTR violations exceeding the bounds for which the test had a very high probative ability. For example, λ, the deflection of light parameter, measures "spacial curvature"; setting the GTR predicted value to 1, experimental tests infer, with high severity, upper bounds to violations. Equivalently, this can be viewed as inferring a confidence interval estimate $\lambda = L \pm \varepsilon$), for are estimated deflectin L.

Some elements of the series of models, for the case of λ, are sketched in Table 2 on the following page.

10. THEORETICAL SIGNIFICANCE OF EXPERIMENTAL KNOWLEDGE IN THE CASE OF GTR

Experimental general relativity need be respresentative of all theory testing in order to yield important insights into the theoretical significance of experimental knowledge. Its particular enlightenment stems from the difficulty of obtaining robust or severe experiments on gravitational effects. This difficulty led physicists to develop a theoretical framework in which to discuss and analyze rivals to GTR, and in which experiments could be compared.

The PPN framework is not merely a set of statistical parameters: it provides a general way to interpret the significance of piecemeal tests for primary gravitational questions. Such an inferential move, however, requires that we rule out mistakes in connecting experimental inferences to substantive physical questions. (This is error

Table 2
PRIMARY: Testing the Post-Newtonian Approximation of GTR: Parametrized Post-Newtonian (PPN) formalism Delineate and test predictions of the metric theories using the PPN parameters: Use estimates to set new limits on PPN parameters and on adjustable parameters in alternatives to GTR e.g., λ How much spatial curvature does mass produce?
EXPERIMENTAL MODELS: PPN parameters are modeled as statistical null hypotheses (relating to models of the experimental source) Failing to reject the null hypothesis (identified with the GTR value) leads to setting upper and lower bounds, values beyond which are ruled out with high severity. e.g., hypotheses about λ in optical and radio deflection experiments
DATA: Models of the experimental source (eclipses, quasar, moon, earth-moon system, pulsars, Cassini) Least squares fits of several parameters, using a function of the observed statistic and the PPN parameter of interest (with known distribution) e.g., least squares estimates of from "raw" data in eclipse and radio interferometry experiments.
DATA GENERATION & ANALYSIS, EXPERIMENTAL DESIGN How to collect and model data.

#5 on my list.) This issue—not a trivial one—is illustrated in our framework with two-way arrows linking experimental models to primary questions.

10.1 LINKING EXPERIMENTAL (STATISTICAL) MODELS TO PRIMARY (SUBSTANTIVE) QUESTIONS

A central concern in forging the experimental (statistical)-substantive link is how to determine to which questions a given test is discriminating answers. Notably, it was determined that one of the classic tests of GTR (test of redshift) "was not a true test" of GTR but rather tested the *equivalence principle*—roughly the claim that bodies of different composition fall with the same accelerations in a gravitational field. This principle is inferred with severity by passing a series of its own null hypotheses tests (e.g., Eotvos experiments), which assert a zero difference in the accelerations of two differently composed bodies. The precision with which these null hypotheses passed warranted the inference that "gravity is a phe-

nomenon of curved spacetime, that is, it must be described by a 'metric theory' of gravity" (Will 1993, 10). What had earlier been taken as a test of GTR had to be re-described once it was realized that red-shift tests could only discriminate metric from nonmetric theories (all metric theories view gravity as "curved space-time" phenomena). This recognition emerged with the discovery that all metric theories say the same thing (with respect to the equivalence principle); they were not rivals with respect to this principle.

More generally, an important task is to distinguish among classes of experiments according to the specific aspects each probed and thus tested. An adequate account of the role and testing of theories must include this. The equivalence principle itself, more correctly called the Einstein Equivalence Principle, admitted of new partitions (e.g., into Strong and Weak), leading to further progress.[12]

The experimental knowledge gained permits us, not merely to infer that we have a correct parameter value, but also to correctly *understand* gravity or how gravity behaves in a given domain. Different values for the parameters correspond to different mechanisms. For example, in one of the most promising GTR rivals, the Brans-Dicke theory, gravity couples both to a tensor metric and a scalar, and the latter is related to a distinct metaphysics (Mach's principle). Although clearly theoretical background is what provides the interpretation of the theoretical significance of the experimental effects (for gravity), there is no one particular theory that needs to be accepted to employ the PPN framework—this is at the heart of its robustness. Even later when this framework was extended to include nonmetric theories ("the search for strong gravitational effects"), those effects that had been vouchsafed with severity remain (even where they may demand reinterpretation).

10.2 NORDVEDT EFFECT η

The rival Brans-Dicke theory, with its adjustable parameter, was able to fit the existing data, but it was not severely tested by the data. Nor, however, did the data count as evidence against it; that came later. In particular, Nordvedt discovered during the 1960s that Brans-Dicke theory would conflict with GTR by predicting a violation of what came to be known as the Strong Equivalence Principle (basically the Weak Equivalence Principle for massive self-gravitating bodies, e.g., stars and planets; see note 12). Correspondingly, a new parameter to describe this effect, the Nordvedt effect, was introduced into the PPN framework, i.e., η.

Following the general pattern for these tests, η was set at 0 for GTR, so the null hypothesis tested is that $\eta = 0$ as against non-0 for rivals. Measurements of round trip travel times between the earth and moon (between 1969 and 1975) enabled the existence of such an anomaly for GTR to be probed severely (the

measurements continue today). Again, the "unbiased, theory-independent viewpoint" of the PPN framework (Will 1993, 157) allowed the conflicting prediction to be identified. Because the tests were sufficiently sensitive, the measurements provided good evidence that the Nordvedt effect was absent, set upper bounds to the possible violations, and provided evidence for the correctness of what GTR says with respect to this effect—once again instantiating the familiar logic.[13]

11. PIECEMEAL TESTS, INTERCONNECTED PARAMETERS, AND SQUEEZING THEORY SPACE

Although the tests are conducted piecemeal, it does not follow that they present us with a disconnected array of local results, as some fear:

> According to [Mayo], a test, even a severe test, of the light-bending hypothesis leaves us in the dark about the ability of GTR to stand up to tests of different ranges of its implications. For instance, should GTR's success in the light-bending experiments lend plausibility to GTR's claims about gravity waves or black holes? (Laudan 1997, 313)

In the error-statistical account of experiment, whether a theory's success in one range signifies likely success in another is an empirical question that must be answered case by case. In the current view, a single context-free answer would not even be desirable. As experimental knowledge of GTR grows, the astrometric (experimental) models show that many of the parameters are functions of the others. For example, it was determined that the deflection effect parameter λ measures the same thing as the so-called time delay, and the Nordevdt parameter η gives estimates of several others. These offer powerful and interconnected checks that fortify and check existing inferences (Mayo 1997a).

11.1 CLEAN TESTS

For instance, hypotheses about λ, as well as how λ constrains other parameters, are now known to have passed with severity, or, as the experimental relativist would put it, with "clean" tests. What clean (severe) tests enable us to do is to *detach inferences* (in this case about gravity) and thereby shrink the possible alternative theories of gravity—in what Clifford Will calls a "gravitation theory independent way." He continues: "The use of the PPN formalism was a clear example of this approach of squeezing theory space" (1993, 303). That is, putting together the interval estimates, it is possible to constrain the values of the PPN parameters and thus "squeeze" the space of theories into smaller and smaller volumes. In this

way entire chunks of theories are ruled out at a time (i.e., all theories that predict the values of the parameter outside the interval estimates). By getting increasingly accurate estimates, more severe constraints are placed on how far theories can differ from GTR, in the respects probed. By 1980 it could be reported that "one can now regard solar system tests of post Newtonian effects as measurements of the 'correct' values of these parameters" (Will 1993).

11.2. GOING BEYOND SOLAR SYSTEM TESTS: GRAVITY WAVES

Even as experimental relativists break out of the "metric theory" framework, progress is continually made by recognizing the general errors that research at any stage is prey to.

> All tests of GTR within the solar system have this qualitative weakness: they say nothing about how the "correct" theory of gravity might behave when gravitational forces are very strong such as near a neutron star. (Will 1996, 273)

The discovery (in 1974) of the binary pulsar 1913+16 opened up the possibility of probing new aspects of gravitational theory: the effects of gravitational radiation.

"The discovery of PSR 1913 + 16 caused considerable excitement in the relativity community...because it was realized that the system could provide a new laboratory for studying relativistic gravity." In general, "the system appeared to be a 'clean' laboratory, unaffected by complex astrophysical processes" (W 1993, 284). Here, "relativistic gravitational theory"—but no one theory within the viable set—is used as a tool to estimate statistically such parameters as the mass of the pulsar. Learning about relativistic effects without assuming the truth of any one theory of gravity, researchers opportunistically used the severely passed relativistic hypotheses to increase knowledge of novel phenomena such as gravity waves.[14] Using GTR as a tool for measuring astrophysical parameters in the binary pulsar, we become "applied relativists" in Will's terminology.

11.3 SQUEEZING PHYSICS SPACE

In particular, experimental relativists were able to arrive at qualitative contrasts between the predictions of the effects of gravity waves on the pulsar's orbit (just in time for the 1979 centenary of Einstein's birth). Hypothetically assuming alternative theories of gravitation, they discovered that one theory, Rosen's bimetric theory, "faces a killing test" by yielding a qualitatively different prediction—the orbit should slow down rather than speed up (Mayo 2000).

The estimated orbital decay is in sync with GTR, but this is not regarded as providing reliable evidence for all of GTR; at most it provides indirect evidence for the existence of gravity waves. The adjustable parameter in Brans–Dicke theory prevents the binary results from discriminating between them: "[T]he theoretical predictions are sufficiently close to those of general relativity, and the

uncertainties in the physics still sufficiently large that the viability of the theory cannot be judged reliably" (Will 2004, 307).

In interpreting the results, in other words, there must be a careful assessment to determine what is and is not ruled out with severity. Pinpointing the domains where existing tests have poor discriminatory power is an essential part of the evidential report, indicating potential rivals not ruled out. Far from providing grounds that all of a theory must be accepted as true, even where it has failed to pass with severity as a whole, a correct understanding of how progress is made reinforces my conjecture that "enough experimental knowledge will do" (Mayo 1996).

12. CONCLUDING REMARKS

The error-statistical account of evidence has been challenged to supplement its account of the life of experiment with an adequate account of the life of theory. Having accepted the challenge, I would advocate an account of theory more dynamic than those extant. An adequate account of theory needs to go beyond a retrospective sum-up of scientific episodes toward a forward-looking account of discovering/inventing/testing new theories: (It is not just a miracle, as some claim.) It should give us insights as to how to discriminate between those parts of a theory that have and have not been warranted. It should account for the stability of experimental effects through theory change. And it should capture statistical testing, which actually is always required even with nonstatistical theories. Rather than weaken the severity requirement when it comes to testing theories as some have recommended, I have argued that the current conception better accounts for how scientists develop, use, and link experiments to theories.

12.1 THE DESIRE IS NOT TO HAVE THINGS SETTLED

Those who advocate accounts that permit inferring or accepting a theory T, even knowing that many of the ways that T can be wrong have yet to be probed, evidently view the scientist as behaving this way, so that any account that demanded severe tests would be at odds with actual practice. I deny their take on the scientific researcher.

Doubtless there are contexts where the scientist wishes to justify adhering to the status quo—one is reminded of Kuhn's famous "normal scientist." We may grant that normal scientists are being perfectly "rational" and are earning their keep, while denying that such a stance can spark the kinds of probing and criticism that extend the frontiers of knowledge. As it is, I fail to see such conser-

vatism in scientific practice; even scientists engaged in day-to-day (normal) practice are keen to push the boundaries, if only on a local question or method of investigation. As much as experimental relativists revered Einstein, they could not have been happier to quit their hibernation and take on the challenge of finding an anomalous effect that would not go away. In my view, the scientist is driven to replace hypotheses and methods with those that show us where our earlier understanding was in error, and teach us more.

An error statistical account of evidence may well go to show the inadequacy of certain distinctions in analytic epistemology. Whether and how severely data $\boldsymbol{x}$ from test $\boldsymbol{E}$ passes hypothesis $\boldsymbol{H}$, is an objective matter. Even though setting standards for required severity levels vary in different contexts and stages of inquiry, the analysis of how well or poorly those standards are met is a matter of the properties of the test and data. In order to warrant a claim that $\boldsymbol{H}$ has passed with severity, satisfaction of the severity requirements need to be shown. Indeed, the ability to implement this account, with the kind of information experimenters tend to have, is an important asset. The actual level of severity need only be approximate, and reporting benchmarks is typical, even in formal statistical settings (e.g., .01, .95). Severely tested hypotheses may prove to have been wrong, there is no claim to infallibility; but exploiting the same reasoning promotes identification of failed assumptions, as well as erroneous conceptions that need revision or reinterpretation.

12.2 UNDERSTANDING A THEORY

The current account of the life of theory, as I have set it out here, still adheres too closely to the standard conception that scientific learning is all about inferring models, theories, and equations. That is because it is difficult to adequately capture what I mean by "the theoretical significance of experimental knowledge." Experimental relativists allow that it is only when they began experimentally testing GTR that they began to really *understand* relativistic gravity. Note that even if scientists had somehow known in 1930 that GTR was true (despite the limited evidence), they still could not have been said to have correctly understood relativistic gravity—how it behaves in its variety of interactions and domains. At the root of this claim is the idea that what is learned with severity is experimental knowledge, which is not limited to knowledge of how to generate a distribution of observable outcomes. Nor need what is learned with severity be well captured by writing down some theory, hypotheses or equations, at least not as that is usually understood.

What is learned may reasonably be captured by means of one or more hypotheses couched in one of the intermediate, experimental models linking data and theory, often statistical. But even those hypotheses seem like little more than placeholders for the full experimental comprehension of the mechanism or process involved. How to arrive at a fuller explication is a task for the future.

REFERENCES

Achinstein, P. 2001. *The Book of Evidence*. Oxford: Oxford University Press.

_____. 2010. Mill's Sins or Mayo's Errors? In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, edited by D. Mayo and A. Spanos, 170–188. Cambridge: Cambridge University Press.

_____. 2011. Philosophy of Science Matters (The Philosophy of Peter Achinstein), G. Morgan ed.; Oxford.

Chalmers, A.F. 1999. *What Is This Thing Called Science?* 3rd ed. Queensland, Australia: University of Queensland Press.

_____. 2010. Can Scientific Theories Be Warranted? In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, edited by D. Mayo and A. Spanos, 58–72. Cambridge: Cambridge University Press.

Earman, J. 1992. *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory.* Cambridge, MA: MIT Press.

_____. 1993. Underdetermination, Realism, and Reason. In *Midwest Studies in Philosophy* 18, edited by P. French, T. Uehling Jr., and H. Wettstein. Notre Dame: University of Notre Dame Press.

Laudan, L. 1997. How about bust? Factoring explanatory power back into theory evaluation. *Philosophy of Science* 64:303–16.

Lightman, A., & D. Lee. 1973. Restricted proof that the weak equivalence principle implies the Einstein equivalence principle. *Phys. Rev. D* 8:364.

Mayo, D. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press (Science and its Conceptual Foundations Series).

Mayo, D. 1997a. Duhem's problem, the Bayesian Way and Error statistics, or 'what's belief got to do with it?' and Response to Howson and Laudan. *Philosophy of Science* 64: 222-44; 323-33.

Mayo, D. (1997b), Severe tests, arguing from error, and methodological underdetermination. *Philosophical Studies* 86: 243-66.

_____. 2000. Experimental practice and an error statistical account of evidence. *Philosophy of Science* 67(3):193–207.

_____. 2005. Evidence as Passing Severe Tests: Highly Probed vs Highly Proved. In P. Achinstein (ed.), Scientific Evidence, Johns Hopkins University Press, 95-127.

_____. 2010a. Severe Testing, Error Statistics, and the Growth of Theoretical Knowledge. In D. Mayo and A. Spanos (eds.), 28–57. Cambridge: Cambridge University Press.

_____. 2010b. Sins of the Epistemic Probabilist: Exchanges with Peter Achinstein. In D. Mayo and A. Spanos (eds.), 28–57.

_____. 2010c. Can Scientific Theories be Warranted with Severity? Exchanges with Alan Chalmers. In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, edited by D. Mayo and A. Spanos, 73–87. Cambridge: Cambridge University Press.

_____. 2010d. Towards Progressive Critical Rationalism: Exchanges with Alan Musgrave. In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, edited by D. Mayo and A. Spanos 2010, 115–124.

_____. 2011. The Objective Epistemic Probabilist and the Severe Tester. In *Philosophy of Science Matters: The Philosophy of Peter Achinstein*, edited by G. Morgan. Oxford: Oxford University Press.

Mayo, D., & D. Cox. 2010. Frequentist Statistics as a Theory of Inductivist Inference. In D. Mayo and A. Spanos. Cambridge: Cambridge University Press, 247-275.

Mayo, D., & M. Kruse. 2001. Principles of Inference and Their Consequences. In *Foundations of Bayesianism,* edited by D. Cornfield and J. Williamson, 381–403. Dordrecht: Kluwer Academic Publishers.

Mayo, D., & A. Spanos. 2006. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal for the Philosophy of Science* 57(2):23–57.

_____, eds. 2010. *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science.* Cambridge: Cambridge University Press.

_____. 2010. Introduction and Background. In

D. Mayo and A. Spanos (eds.), 1–27. Cambridge: Cambridge University Press.

_____. 2011. Error Statistics. In *Handbook of Philosophy of Science* (7), edited by P. Bandyopadhyay and M. Forster. Amsterdam, The Netherlands: Elsevier.

Musgrave, A. 2010. Critical Rationalism, Explanation, and Severe Tests. In *Error and Inference*: *Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, edited by D. Mayo and A. Spanos, 88–112. Cambridge: Cambridge University Press.

Pierce, C.S. 1931-35. *Collected Papers*. Vols. 1-6, edited by C. Hartshorne and P. Weis. Cambridge: Harvard University Press.

Will, C. 1993. *Theory and Experiment in Gravitational Physics. Cambridge:* Cambridge University Press.

_____. 1996. The Confrontation Between General Relativity and Experiment. A 1995 Update. In *General Relativity: Proceedings of the Forty Sixth Scottish Universities Summer School in Physics*, edited by G. Hall and J. Pulham. Edinburgh: SUSSP Publications. London: Institute of Physics.

_____. 2004. The confrontation between general relativity and experiment. *Living Reviews in Relativity,* http://relativity.livingreviews.org/Articles/1rr-2001-4/title.html.

_____. 2005. Relativity at the centenary. *Physics World* 18:27.

ENDNOTES

[1] A draft of this paper was given as the Henle Lecture at the University of St. Louis, March 2010; I am extremely grateful and honored to have been invited to present this Memorial Lecture, and I acknowledge the useful feedback from that forum.

[2] I have been responding to such challenges over the past decade, and Kent Staley has been a participant in a number of exchanges throughout this period, most recently at an April 29, 2011 forum at Virginia Tech ("Experimental Knowledge and the Deep Structure of the World"). This paper incorporates several of those responses. See also Mayo and Spanos (2010)

[3] Since I will be talking about theory *T* in this paper, I use *E* to stand for an experimental test.

[4] For instance, *E* would need to give us an adequate distance measure, and a way to determine the probability (formal or informal) of different distances being observed under one, or another, hypothesis of interest.

[5] The single notion of severity suffices to direct the interpretation and scrutiny of the two types of errors in statistics: erroneously rejecting a statistical (null) hypothesis h_0—type I error—and erroneously failing to reject h_0—type II error. It lets us immediately avoid often-repeated statistical fallacies due to tests that are overly sensitive, as well as those insufficiently sensitive to particular errors, although that is a topic for a different discussion. (See Mayo 1996; Mayo and Spanos 2006, 2011).

[6] In Mayo 1996, error numbers 4 and 5 were collapsed under the rubric of "experimental assumptions."

[7] This is unsurprising given that a concern with hypothetical error probabilities goes beyond the Bayesian model, at least if it is Bayesian coherent (and thus obeys the likelihood principle). For discussion, see Mayo 1996; Mayo and Kruse 2001.

[8] The usual context-free logical transformations that hold for the probability calculus of events do not hold for a context-dependent construal of well testedness.

[9] These are taken up in Mayo 2010c, 2010d, in exchanges with Chalmers and Musgrave, respectively.

[10] A classic example is to use the data to find a rival to a null hypothesis that makes the data maximally likely, e.g., hunting for statistical significance. It can be shown that such an alternative passes a test with minimal severity , e.g., Mayo and Cox (2010).

[11] To my knowledge, Earman is the only philosopher of science to discuss the PPN framework in some detail. Although the program has extended considerably beyond his 1992 discussion, the current framework continues to serve in much the same manner.

[12] More carefully, we should identify the Einstein Equivalence Principle (EEP) as well as distinguish between weak and strong forms. The EEP states that: (1) the Weak Equivalence Principle (WEP) is valid; (2) The outcome of any local nongravitational experiment is independent of the velocity of the freely falling reference frame in which it is performed (Lorentz invariance); (3) The outcome of any local nongravitational experiment is independent of where and when in the universe it is performed (local position invariance). A subset of metric theories obeys a stronger principle, the Strong Equivalence Principle (SEP). The SEP asserts

that the stipulations of the equivalence principle also hold for self-gravitating bodies, such as the earth-moon system.

[13] In the "secondary" task of scrutinizing the experimental assumptions, they asked whether other factors could mask the η effect. Most, it was argued, can be separated cleanly from the η effect using the multiyear span of data; others are known with sufficient accuracy from previous measurements or from the lunar lasing experiment itself.

[14] An extended formalism was developed by Lightman and Lee (1973) to systematize the search for violations of the Einstein Equivalence Principle (EEP). The class of theories that can be described within the *TH_*μ formalism includes all metric theories, as well as many, but not all, nonmetric theories. The ability to put nonmetric theories into a common framework such that bounds can be put on EEP violations in a systematic way provides a powerful extension of the program of testing within the PPN framework.