Learning from Error, Severe Testing, and the Growth of Theoretical Knowledge

Deborah G. Mayo

I regard it as an outstanding and pressing problem in the philosophy of the natural sciences to augment the insights of the new experimentalists with a correspondingly updated account of the role or roles of theory in the experimental sciences, substantiated by detailed case studies. (Chalmers, 1999, p. 251)

1 Background to the Discussion

The goal of this chapter is to take up the aforementioned challenge as it is posed by Alan Chalmers (1999, 2002), John Earman (1992), Larry Laudan (1997), and other philosophers of science. It may be seen as a first step in taking up some unfinished business noted a decade ago: "How far experimental knowledge can take us in understanding theoretical entities and processes is not something that should be decided before exploring this approach much further" (Mayo, 1996, p. 13). We begin with a sketch of the resources and limitations of the "new experimentalist" philosophy.

Learning from evidence, in this experimentalist philosophy, depends not on appraising large-scale theories but on local experimental tasks of estimating backgrounds, modeling data, distinguishing experimental effects, and discriminating signals from noise. The growth of knowledge has not to do with replacing or confirming or probabilifying or "rationally accepting" large-scale theories, but with testing specific hypotheses in such a way that there is a good chance of learning something – whatever theory it winds up as part of. This learning, in the particular experimental account we favor, proceeds by testing experimental hypotheses and inferring those that pass probative or *severe* tests – tests that would have unearthed some error in, or discrepancy from, a hypothesis H, were H false. What enables this account of severity to work is that the immediate hypothesis H under test by means

of data is designed to be a specific and local claim (e.g., about parameter values, causes, the reliability of an effect, or experimental assumptions). "*H* is false" is not a disjunction of all possible rival explanations of the data, at all levels of complexity; that is, it is not the so-called catchall hypothesis but refers instead to a specific error being probed.

1.1 What Is the Problem?

These features of piecemeal testing enable one to exhaust the possible answers to a specific question; the price of this localization is that one is not entitled to regard full or large-scale theories as having passed severe tests, so long as they contain hypotheses and predictions that have not been well probed. If scientific progress is thought to turn on appraising high-level theories, then this type of localized account of testing will be regarded as guilty of a serious omission, unless it is supplemented with an account of theory appraisal.

1.2 The Comparativist Rescue

A proposed remedy is to weaken the requirement so that a large-scale theory is allowed to pass severely so long as it is the "best-tested" theory so far, in some sense. Take Laudan:

> [W]hen we ask whether [the General Theory of Relativity] GTR can be rationally accepted, we are not asking whether it has passed tests which it would almost certainly fail if it were false. As Mayo acknowledges, we can rarely if ever make such judgments about most of the general theories of the science. But we can ask "Has GTR passed tests which none of its known rivals have passed, while failing none which those rivals have passed." Answering such a question requires no herculean enumeration of all the possible hypotheses for explaining the events in a domain. (Laudan, 1997, p. 314)

We take up this kind of comparativist appraisal and argue that it is no remedy; rather, it conflicts with essential ingredients of the severity account – both with respect to the "life of experiment" and to the new arena, the "life of theory."

1.3 Is Severity Too Severe?

One of the main reasons some charge that we need an account showing acceptance of high-level theories is that scientists in fact seem to accept them; without such an account, it is said, we could hardly make sense

of scientific practice. After all, these philosophers point out, scientists set about probing and testing theories in areas beyond those in which they have been well tested. While this is obviously true, we question why it is supposed that in doing so scientists are implicitly accepting all of the theory in question. On the contrary, we argue, this behavior of scientists seems to underscore the importance of distinguishing areas that are from those that are not (thus far) well tested; such a distinction would be blurred if a full theory is accepted when only portions have been well probed. Similarly, we can grant Earman's point that "in 1918 and 1919 physicists were in no position to be confident that the vast and then unexplored space of possible gravitational theories denoted by [not-GTR] does not contain alternatives to GTR that yield that same prediction for the bending of light as GTR" (Earman, 1992, p. 117), while asking why this shows our account of severity is too severe rather than being a point in its favor. It seems to us that being prohibited from regarding GTR as having passed severely, at that stage, is just what an account ought to do. At the same time, the existence of what Earman aptly dubs a "zoo of alternatives" to GTR did not prevent scientists from severely probing and passing claims about light-bending and, more generally, extending their knowledge of gravity. We shall return to consider GTR later.

1.4 The Challenge

We welcome the call to provide the "life of experiment" with a corresponding "life of theory": the challenge leads to extending the experimental testing account into that arena in ways that we, admittedly, had not been sufficiently clear about or had not even noticed. In particular, taking up the large-scale theory challenge leads to filling in some gaps regarding the issues of (1) how far a severity assessment can extend beyond the precise experimental domain tested and (2) what can be said regarding hypotheses and claims that *fail* to have passed severe tests. Regarding the first issue, we argue that we can inductively infer the absence of any error that has been well probed and ruled out with severity. Although "*H* is false" refers to a specific error, this may and should encompass erroneous claims about underlying causes and mistaken understandings of any testable aspect of a phenomenon of interest. Concerning the second issue, we wish to explore the value of understanding why evidence may prohibit inferring a full theory severely – how it helps in systematically setting out rivals and partitioning the ways we can be in error regarding the claims that have so far agreed with data.

Thus, we accept the challenge in the epigraph, but in addition wish to "raise the stakes" on what an adequate account of theory appraisal should provide. More than affording an after-the-fact reconstruction of past cases of theory appraisal, an adequate account should give forward-looking methods for making progress in both building and appraising theories. We begin in Section 2 by considering the severity account of evidence; then in Section 3, we consider some implications for high-level theory. In Section 4, we examine and reject the "comparativist rescue" and in Section 5, we take up the case of theory testing of GTR. Our issue – let me be clear at the outset – is not about whether to be a realist about theories; in fact the same criticisms are raised by philosophers on both sides of this divide. Thus, in what follows we try to keep to language used by realists and nonrealists alike.

2 Error-Statistical Account of Evidence

2.1 Severity Requirement

Let us begin with a very informal example. Suppose we are testing whether and how much weight has been gained between now and the time George left for Paris, and we do so by checking if any difference shows up on a series of well-calibrated and stable weighing methods, both before his leaving and upon his return. If no change on any of these scales is registered, even though, say, they easily detect a difference when he lifts a 0.1-pound potato, then this may be regarded as grounds for inferring that George's weight gain is negligible within limits set by the sensitivity of the scales. The hypothesis H here might be that George's 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 George to have gained δ pounds or more (i.e., were H false), then this method would almost certainly have detected this. Clearly H has been subjected to, and has passed, a more stringent test than if, say, H were inferred based solely on his still being able to button elastic-waist pants. The same reasoning abounds in science and statistics (p. 256).

Consider data on light-bending as tests of the deflection effect λ given in Einstein's GTR. It is clear that data based on very long baseline radio interferometry (VLBI) in the 1970s taught us much more about, and provided much better evidence for, the Einsteinian-predicted light deflection (often set these days at 1) than did the passing result from the celebrated 1919 eclipse tests. The interferometry tests are far more capable of uncovering a variety of errors, and discriminating values of the deflection, λ, than

were the crude eclipse tests. Thus, the results set more precise bounds on how far a gravitational theory can differ from the GTR value for λ. Likewise, currently-planned laser interferometry tests would probe discrepancies even more severely than any previous tests.

We set out a conception of evidence for a claim or hypothesis *H*:

Severity Principle (SP): Data **x** (produced by process *G*) provides a good indication or evidence for hypothesis *H* if and only if **x** results from a test procedure *T* which, taken as a whole, constitutes *H* having passed a severe test – that is, a procedure which would have, at least with very high probability, uncovered the falsity of, or discrepancies from *H*, and yet no such error is detected.

Instead, the test produces results that are in accord with (or fit) what would be expected under the supposition that *H* is correct, as regards the aspect probed.

While a full explication of severity is developed throughout this volume (e.g., introductory chapter), we try to say enough for current purposes. To begin with, except for formal statistical contexts, "probability" here may serve merely to pay obeisance to the fact that all empirical claims are strictly fallible. Take, for example, the weighing case: if the scales work reliably and to good precision when checked on test objects with known weight, we would ask, rightly, what sort of extraordinary circumstance could cause them to all go systematically astray just when we do not know the weight of the test object (George)? We would infer that his weight gain does not exceed such-and-such amount, without any explicit probability model.[1] Indeed, the most forceful severity arguments usually do not require explicit reference to probability or statistical models. We can retain the probabilistic definition of severity so long as it is kept in mind that it covers this more informal use of the term. Furthermore, the role of probability where it does arise, it is important to see, is not to assign degrees of confirmation or support or belief to hypotheses but to characterize how frequently methods are capable of detecting and discriminating errors, called error frequencies or *error probabilities.* Thus, an account of evidence broadly based on error probabilities may be called an *error-statistical account,* and a philosophy of science based on this account of evidence may be called an error-statistical philosophy of science (see Introduction and Background, Part II).

---

[1] Even in technical areas, such as in engineering, it is not uncommon to work without a well-specified probability model for catastrophic events. In one such variation, *H* is regarded as having passed a severe test if an erroneous inference concerning *H* could result only under extraordinary circumstances. (Ben-Haim, 2001, p. 214)

The severe test reasoning corresponds to a variation of an "argument from error" (p. 24):

Argument from Error: There is evidence that an error is absent when a procedure of inquiry with a high probability of detecting the error's presence nevertheless regularly yields results in accord with no error.

By "detecting" an error, we mean it "signals the presence of" an error; we generally do not know from the observed signal whether it has correctly done so. Since any inductive inference could be written as inferring the absence of an error of some type, the argument from error is entirely general. Formal error-statistical tests provide tools to ensure that errors will be correctly detected (i.e., signaled) with high probabilities.[2]

2.2 Some Further Qualifications

The simple idea underlying the severity principle (SP), once unpacked thoroughly, provides a very robust concept of evidence. We make some quick points of most relevance to theory testing: Since we will use T for theory, let E denote an experimental test.[3] First, although it is convenient to speak of a severe test E, it should be emphasized that E may actually, and usually does, combine individual tests and inferences together; likewise, data $\mathbf{x}$ may combine results of several tests. So long as one is explicit about the test E being referred to, no confusion results. Second, a severity assessment is a function of a particular set of data or evidence $\mathbf{x}$ and a particular hypothesis or claim. More precisely, it has three arguments: a test, an outcome or result, and an inference or a claim. "The severity with which H passes test E with outcome $\mathbf{x}$" may be abbreviated as SEV(Test E, outcome $\mathbf{x}$, claim H). When $\mathbf{x}$ and E are clear, we may write SEV(H). Defining severity in terms of three arguments is in contrast with 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* as test sensitivity

---

[2] Control of error rates, even if repetitions are hypothetical, allows the probativeness of *this* test to be assessed for reliably making *this* inference (see chapter 7). Nevertheless, low long-run error rates at individual stages of a complex inquiry (e.g., the error budgets in astronomic inferences) play an important role in the overall severity evaluation of a primary inference.

[3] Experiments, for us, do not require literal control; it suffices to be able to develop and critique arguments from error, which include the best practices in observational inquiries and model specification and validation. Nor need "thought experiments" be excluded.

increases (not the reverse). For example, we expect our interferometry test to yield some nonzero difference from the GTR prediction ($\lambda = 1$), the null hypothesis of the test, even if $\lambda = 1$. To interpret any observed difference, regardless of how small, as signaling a substantive discrepancy from the GTR prediction would be to infer a hypothesis with very *low* severity. That is because this test would very often purport to have evidence of a genuine discrepancy from $\lambda = 1$, even if the GTR prediction is correct (perhaps within a specified approximation).

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 (sometimes abbreviated as "accepting" h_0) – type II error. The actual inference, H, will generally go beyond the stark formal statistical output. For example, from a statistical rejection of h_0, one might infer:

H: **x** is evidence of a discrepancy δ from h_0.

Then calculating SEV(H) directs one to consider the probability of a type I error.

If h_0 is not rejected, the hypothesis inferred might take the form:

H: **x** is evidence that any discrepancy from h_0 is less than δ.

Now the type II error probability (corresponding to δ) becomes relevant. Severity, as a criterion for evidence, avoids standard statistical fallacies due both to tests that are overly sensitive and to those insufficiently sensitive to particular errors and discrepancies (e.g., statistical vs. substantive differences; see Mayo, 1996; Mayo and Spanos, 2006).

Note that we always construe the question of evidence using testing language, even if it is described as an estimation procedure, because this is our general terminology for evidence, and any such question can be put in these terms. Also, the locution "severely tested" hypothesis H will always mean that H has *passed* the severe or stringent probe, not, for example, merely that H was subjected to one.

2.3 Models of Inquiry

An important ingredient of this account of testing is the insistence on avoiding oversimplifications of accounts that begin with statements of evidence and hypotheses overlooking 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 distinguish the steps involved in any realistic experimental inquiry and locate the necessary background information and the errors being probed – even more so when attempting to relate low-level tests to high-level theories. To organize these interconnected pieces, it helps to view any given inquiry as involving a *primary question* or *problem*, which is then embedded and addressed within one or more other models which we may call "experimental".[4] *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, couched in an appropriate experimental model, may be investigated by means of properly modeled data, not "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 model *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* which (combined 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). We return to this in Section 5.

How to carve out these different models, each sometimes associated with a level in a hierarchy (e.g., Suppes, 1969) is not a cut-and-dried affair, but so long as we have an apparatus to make needed distinctions, this leeway poses no danger. Fortunately, philosophers of science have become increasingly aware of the roles of models in serving as "mediators," to use an apt phrase from Morrison and Morgan (1999), and we can turn to the central issue of this paper.[5]

3 Error-Statistical Account and Large-Scale Theory Testing

This localized, piecemeal testing does have something to say when it comes to probing large-scale theories, even if there is no intention to severely pass the entire theory. Even large-scale theories when we have them (in our account) are applied and probed only by a piecemeal testing of local

[4] This is akin to what Spanos calls the "estimable" model; see Chapter 6, this volume. See also note 3.

[5] Background knowledge, coming in whatever forms available – subject matter, instrumental, simulations, robustness arguments – enters to substantiate the severity argument. We think it is best to delineate such information within the relevant models rather than insert a great big "*B*" for "background" in the SEV relation, especially because these assumptions must be separately probed.

experimental hypotheses. Rival theories T_1 and T_2 of a given phenomenon or domain, even when corresponding to very different primary models (or rather, very different answers to primary questions), need to be applicable to the same data models, particularly if T_2 is to be a possible replacement for T_1. This constraint motivates the development of procedures for rendering rivals applicable to shared data models.

3.1 Implications of the Piecemeal Account for Large-Scale Testing

Several implications or groups of theses emerge fairly directly from our account, and we begin by listing them:

1. *Large-scale theories are not severely tested all at once.* To say that a given experiment E is a test of theory T is an equivocal way of saying that E probes what T says about a particular phenomenon or experimental effect (i.e., E attempts to discriminate the answers to a specific question, H). We abbreviate what theory T_i says about H as $T_i(H)$. This is consistent with the common scientific reports of "testing GTR" when in fact what is meant is that a particular aspect or parameter is going to be probed or delimited to a high precision. Likewise, the theory's passing (sometimes with "flying colors") strictly refers to the one piecemeal question or estimate that has passed severely (e.g., Will, 1993).

2. *A severity assessment is not threatened by alternatives at "higher levels."* If two rival theories, T_1 and T_2, say the same thing with respect to the effect or hypothesis H being tested by experimental test E (i.e., $T_1(H) = T_2(H)$), then T_1 and T_2 *are not rivals* with respect to experiment E. Thus, *a severity assessment can remain stable through changes in "higher level" theories*[6] or answers to different questions. For example, the severity with which a parameter is determined may remain despite changing interpretations about the cause of the effect measured (see Mayo, 1997b).

3. *Severity discriminates between theories that "fit" the data equally well.* T_1 is discriminated from T_2 (whether known, or a "beast lurking in the bush"[7]) by identifying and testing experimental hypotheses on which they disagree (i.e., where $T_1(H) \neq T_2(H)$). Even though *two rival hypotheses might "fit" the data equally well, they will not generally be equally severely tested by experimental test E.*

---

[6] Here we follow Suppes (1969) in placing the models in a vertical hierarchy from the closest to the farthest from data.

[7] We allude here to a phrase in Earman (1992).

The preceding points, as we will see, concern themselves with *reliability*, *stability*, and avoidance of serious *underdetermination*, respectively.

3.2 Contrast with a Bayesian Account of Appraisal

At this point, it is useful to briefly contrast these consequences with an approach, better known among philosophers to the inductive appraisal of hypotheses: the Bayesian approach. Data $\mathbf{x}$ may be regarded as strong evidence for, or as highly confirming of, theory T so long as the posterior probability of T given $\mathbf{x}$ is sufficiently high (or sufficiently higher than the prior probability in T),[8] where probability is generally understood as a measure of degree of belief, and $P(T|\mathbf{x})$ is calculated by means of Bayes's theorem:

$$P(T|\mathbf{x}) = P(\mathbf{x}|T)P(T)/[P(\mathbf{x}|T)P(T) + P(\mathbf{x}|\text{not-}T)P(\text{not-}T)]$$

This calculation requires an exhaustive set of alternatives to T and prior degree-of-belief assignments to each, and an assessment of the term $P(\mathbf{x}|\text{not-}T)$, for "not-$T$," the *catchall hypothesis*. That scientists would disagree in their degree-of-belief probability assignments is something accepted and expected at least by subjectivist Bayesians.[9]

In one sense, it is simplicity itself for a (subjective) Bayesian to confirm a full theory T. For a familiar illustration, suppose that theory T accords with data $\mathbf{x}$ so that $P(\mathbf{x}|T) = 1$, and assume equal prior degrees of belief for T and not-T. If the data are regarded as very improbable given that theory T is false – if a low degree of belief, say e, is accorded to what may be called the *Bayesian catchall factor*, $P(\mathbf{x}|\text{not-}T)$ – then we get a high posterior probability in theory T; that is, $P(T|\mathbf{x}) = 1/(1 + e)$. The central problem is this: What warrants taking data $\mathbf{x}$ as incredible under any theory other than T, when these would include all possible rivals, including those not even thought of? We are faced with the difficulty Earman raised (see 1.3), and it also raises well-known problems for Bayesians.

High Bayesian support does not suffice for well-testedness in the sense of the severity requirement. The severity requirement enjoins us to consider this Bayesian procedure: basically, it is to go from a low degree of belief in the Bayesian catchall factor to inferring T as confirmed. One clearly cannot vouch for the reliability of such a procedure – that it would rarely affirm theory T were T false – in contrast to point 1 above. Similar problems

---

[8] Several related measures of Bayesian confirmation may be given. See, for example, Good (1983).

[9] Some might try to assign priors by appealing to ideas about simplicity or information content, but these have their own problems (e.g., Cox, 2006; Kass and Wasserman, 1996). See Chapter 7, pp. 298–302.

confront the Bayesian dealing with data that are anomalous for a theory *T* (e.g., in confronting Duhemian problems). An anomaly **x**′ warrants Bayesian disconfirmation of an auxiliary hypothesis *A* (used to derive prediction **x**), so long as the prior belief in *T* is sufficiently high and the Bayesian catchall factor is sufficiently low (see, e.g., Dorling, 1979). The correctness of hypothesis *A* need not have been probed in its own right. For example, strictly speaking, believing more strongly in Newton's than in Einstein's gravitational theory in 1919 permits the Bayesian to blame the eclipse anomaly on, say, a faulty telescope, even without evidence for attributing blame to the instrument (see Mayo, 1997a; Worrall, 1993; and Chapters 4 and 8, this volume).

Consider now the assurance about stability in point 2. Operating with a "single probability pie," as it were, the Bayesian has the difficulty of redistributing assignments if a new theory is introduced. Finally, consider the more subtle point 3. For the Bayesian, two theories that "fit" the data **x** equally well (i.e., have identical likelihoods) are differentially supported only if their prior probability assignments differ. This leads to difficulties in capturing methodological strictures that seem important in discriminating two equally well-fitting hypotheses (or even the same hypothesis) based on the manner in which each hypothesis was constructed or selected for testing. We return to this in Section 5. Further difficulties are well known (e.g., the "old evidence problem," Glymour, 1980; Kyburg, 1993) but will not be considered.

I leave it to Bayesians to mitigate these problems, if problems they be for the Bayesian. Of interest to us is that it is precisely to avoid these problems, most especially consideration of the dreaded catchall hypothesis and the associated prior probability assignments, that many are led to a version of a comparativist approach (e.g., in the style of Popper or Lakatos).

3.3 The Holist–Comparativist Rescue

One can see from my first point in Section 3.1 why philosophers who view progress in terms of large-scale theory change are led to advocate a comparative testing account. Because a large-scale theory may, at any given time, contain hypotheses and predictions that have not been probed at all, it would seem impossible to say that such a large-scale theory had severely passed a test as a whole.[10] A comparative testing account, however, would

---

[10] Note how this lets us avoid tacking paradoxes: Even if *H* has passed severely with data **x**, if **x** fails to probe hypothesis *J*, then **x** fails to severely pass *H* and *J* (see Chalmers, 1999). By contrast, Bayesians seem content to show that **x** confirms the irrelevant conjunction less strongly than the conjunct (see Chapter 8, this volume). For a recent discussion and references, see Fitelson (2002).

allow us to say that the theory is best tested so far, or, using Popperian terms, we should "prefer" it so far. Note that their idea is not merely that testing should be comparative – the severe testing account, after all, tests *H* against its denial within a given model or space – but rather that testing, at least testing large-scale theories, may and generally will be a comparison between *nonexhaustive* hypotheses or theories. The comparativist reasoning, in other words, is that since we will not be able to test a theory against its denial (regarded as the "catchall hypothesis"), we should settle for testing it against one or more *existing* rivals. Their position, further, is that one may regard a theory as having been well or severely tested as a whole, so long as it has passed more or better tests than its existing rival(s). To emphasize this we will allude to it as a *comparativist-holist* view:

The comparativist . . . insists on the point, which [Mayo] explicitly denies, that testing or confirming one "part" of a general theory provides, defeasibly, an evaluation of all of it. (Laudan, 1997, p. 315)

Alan Chalmers maintains, in an earlier exchange, that we must already be appealing to something akin to a Popperian comparativist account:

[Mayo's] argument for scientific laws and theories boils down to the claim that they have withstood severe tests better than any available competitor. The only difference between Mayo and the Popperians is that she has a superior version of what counts as a severe test. (Chalmers, 1999, p. 208)

Amalgamating Laudan and Chalmers's suggestions for "comparativist–holism" gives the following:

Comparativist (Holist) Testing: A theory has been well or *severely tested* provided that it has survived (local) severe tests that its known rivals have failed to pass (and not vice versa).

We argue that the comparativist–holist move is no rescue but rather conflicts with the main goals of the severity account, much as the Bayesian attempt does. We proceed by discussing a cluster of issues relating to the points delineated in Section 3.1.

4 Comparing Comparativists with Severe Testers

4.1 Point 1: Best Tested Does Not Entail Well Tested

One cannot say about the comparatively best-tested theory what severity requires – that the ways the theory or claim can be in error have been well-probed and found to be absent (to within the various error margins of the test). It seems disingenuous to say all of theory *T* is well tested (even to a

degree) when it is known there are ways *T* can be wrong that have received no scrutiny or that there are regions of implication not checked at all. Being best tested is relative not only to existing theories but also to existing tests: they may all be poor tests for the inference to *T* as a whole. One is back to a problem that beset Popper's account – namely, being unable to say "What is so good about the theory that (by historical accident) happens to be the best tested so far?" (Mayo, 2006, p. 92).

Whereas we *can* give guarantees about the reliability of the piecemeal experimental test, we *cannot* give guarantees about the reliability of the procedure advocated by the comparativist-holist tester. Their procedure is basically to go from passing hypothesis *H* (perhaps severely in its own right) to passing all of *T* – but this is a highly *un*reliable method; anyway, it is unclear how one could assess its reliability. By contrast, we can apply the severity idea because the condition "given *H* is false" (even within a larger theory) always means "given *H* is false with respect to what *T* says about *this particular* effect or phenomenon" (i.e., *T*(*H*)).[11] If a hypothesis *T*(*H*) passes a severe test we can infer something positive: that *T* gets it right about the specific claim *H*, or that given errors have been reliably ruled out. This also counts as evidence against any rival theory that conflicts with *T*(*H*).

Granted, it may often be shown that ruling out a given error is connected to, and hence provides evidence for, ruling out others. The ability to do so is a very valuable and powerful way of cross-checking and building on results. Sometimes establishing these connections is achieved by using theoretical background knowledge; other times sufficient experimental knowledge will do. But whether these connections are warranted is an empirical issue that has to be looked into on a case-by-case basis – whereas the comparativist-holist would seem to be free of such an obligation, so long as theory *T* is the best tested so far. Impressive "arguments from coincidence" from a few successful hypotheses to the entire theory must be scrutinized for the case in hand. We return to this in Chapter 2.

Rational Acceptability. It is not that we are barred from finding a theory *T* "rationally acceptable," preferred, or worthy of pursuit – locutions often used by comparativists – upon reaching a point where *T*'s key experimental predictions have been severely probed and found to pass. One could infer that *T* had solved a set of key experimental problems and take this as

---

[11] It is important to see that the severity computation is not a conditional probability, which would implicitly assume prior probability assignments to hypotheses which severity does not assume. Rather, severity should be understood as the probability of so good an agreement (between *H* and **x**) *calculated under the assumption that H is false.*

grounds for "deciding to pursue" it further. But these decisions are distinct from testing and would call for a supplement to what we are offering.[12]

As we see it, theories (i.e., theoretical models) serve a role analogous to experimental models in the tasks of learning from data. Just as experimental models serve to describe and analyze the relevance of any of the experimental data for the experimental phenomenon, theoretical models serve *to analyze the relevance of any of the experimental inferences (estimates and tests) for the theoretical phenomenon.* If a theory T_2 is a viable candidate to take the place of rival T_1, then it must be able to *describe and analyze the significance of the experimental outcomes that T_1 can.* We come back to this in considering GTR. We should be concerned, too, by the threat to the *stability* of severity assessments that the comparativist account would yield – the second point in Section 3.1.

4.2 Point 2: Stability

Suppose an experimental test E is probing answers to the question: What is the value of a given parameter λ? Then, if a particular answer or hypothesis severely passes, this assessment is not altered by the existence of a theory that gives the same answer to this question. More generally, our account lets us say that severely passing $T(H)$ (i.e., what T says about H) gives us experimental knowledge about this aspect of T, and this assessment remains even through improvements, revisions, and reinterpretations of that knowledge. By contrast, the entrance of a rival that passes all the tests T *does* would seem to force the comparativist to change the assessment of how well theory T had been tested.

On the severity account, if a rival theory T_2 agrees with T_1 with respect to the effect or prediction under test, then the two theories are not rivals *so far as this experimental test is concerned* – no matter how much they may differ from each other in their full theoretical frameworks or in prediction ranges not probed by the experimental test E. It is very important to qualify this claim. Our claim is not that two theories fail to be rivals just because the test is insufficiently sensitive to discriminate what they say about the phenomenon under test; our claim is that they fail to be rivals when the two say exactly the same thing with respect to the effect or hypothesis under test.[13] The severity assessment reflects this. If theory T_1 says exactly the

---

[12] Larry Laudan (1977) himself has always stressed that we should distinguish theory pursuit from other stances one might take toward theories.

[13] Of course, determining this might be highly equivocal, but that is a distinct matter.

same thing about H as T_2 – that is, $(T_1(H) = T_2(H))$ – then T_2 cannot alter the severity with which the test passes H.[14] Note, though, that this differs from saying $T_1(H)$ and $T_2(H)$ pass with equal severity. We consider this argument in Section 4.3.

4.3 Point 3: Underdetermination

Point 3 refers to a key principle of error statistics, which is also the basis for solving a number of philosophical problems. It is often argued that data underdetermine hypotheses because data may equally well warrant conflicting hypotheses according to one or another base measure of evidential relationship. However, we can distinguish, on grounds of severity, the well-testedness of two hypotheses and thereby get around underdetermination charges. We take this up elsewhere (e.g., Mayo, 1997b). Here our interest is in how the feature in point 3 bears on our question of moving from low-level experimental tests to higher level theories. In particular, two hypotheses may be nonrivals (relative to a primary question) and yet be tested differently by a given test procedure – indeed the same hypothesis may be better- or less-severely tested by means of (what is apparently) the "same" data because of aspects of either the data generation or the hypothesis construction procedure.

We can grant, for example, that a rival theory could always be erected to accommodate the data, but a key asset of the error-statistical account is its ability to distinguish the well-testedness of hypotheses and theories by the reliability or severity of the accommodation method. Not all fits are the same. Thus, we may be able to show, by building on individual hypotheses, that one theory *at some level* (in the series or models) or a close variant to this theory, severely passes. In so doing, we can show that no rival to this theory can also severely pass.

Admittedly, all of this demands an examination of the detailed features of the recorded data (the data models), not just the inferred experimental effect or phenomenon. It sounds plausible to say there can always be some rival, when that rival merely has to "fit" already-known experimental effects. The situation is very different if one takes seriously the constraints imposed

---

[14] Mistakes in regarding H as severely passed can obviously occur. A key set of challenges comes from those we group under "experimental assumptions." Violated assumptions may occur because the actual experimental data do not satisfy the assumptions of the experimental model or because the experimental test was not sufficiently accurate or precise to reliably inform about the primary hypothesis or question. Of course, "higher-lower" is just to distinguish primary questions; they could be arranged horizontally.

by the information in the detailed data coupled with the need to satisfy the severity requirement.

Finally, nothing precludes the possibility that so-called low-level hypotheses *could* warrant inferring a high-level theory with severity. Even GTR, everyone's favorite example, is thought to predict a unique type of gravitational radiation, such that affirming that particular "signature" with severity would rule out all but GTR (in its domain). With this tantalizing remark, let us look more specifically at the patterns of progress in experimental GTR.

5 Experimental Gravitation

This example is apt for two reasons. First, it is an example to which each of the philosophers we have mentioned allude in connection with the problem of using local experimental tests for large-scale theories. Second, the fact that robust or severe experiments on gravitational effects are so hard to come by led physicists to be especially deliberate about developing a theoretical framework in which to discuss and analyze rivals to GTR and to compare the variety of experiments that might enable their discrimination. To this end, they 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 only philosopher of science to discuss the PPN framework in some detail, to my knowledge, is John Earman; although the program has been updated and extended since his 1992 discussion, the framework continues to serve in much the same manner. What is especially interesting about the PPN framework is its role in *inventing* new classes of rivals to GTR, beyond those that are known. It points to an activity that any adequate account of theories should be able to motivate, if it is to give forward-looking methods for making theoretical progress rather than merely after-the-fact reconstructions of episodes. Popperians point out that Popper had always advocated looking for rivals as part of his falsification mandate. Granted, but neither he nor the current-day critical rationalists supply guidance for developing the rivals or for warranting claims about where hypotheses are likely to fail if false – eschewing as they do all such inductivist claims about reliable methods (see Mayo, 2006).[15]

Experimental testing of GTR nowadays is divided into four periods: 1887–1919, 1920–1960, 1960–1980, and 1980 onward. Following Clifford

---

[15] Popper's purely deductive account is incapable, by his own admission, of showing the reliability of a method.

Will, the first is the period of *genesis*, which encompasses experiments on (1) the foundations of relativistic physics (Michelson-Morley and the Eötvös experiments) and the GTR tests on (2) the deflection of light and perihelion of Mercury (for excellent discussions, see Will, 1980, 1986, 1996, 2004). From the comparativist's perspective, 1920–1960 would plainly be an era in which GTR enjoyed the title of "best-tested" theory of gravity: it had passed the "classical" tests to which it had been put and no rival existed with a superior testing record to knock it off its pedestal. By contrast, from 1960 to 1980, a veritable "zoo" of rivals to GTR had been erected, all of which could be constrained to fit these classical tests. So in this later period, GTR, from the comparativist's perspective, would have fallen from its pedestal, and the period might be regarded as one of crisis, threatening progress or the like. But in fact, the earlier period is widely regarded (by experimental gravitation physicists) as the period of "stagnation," or at least "hibernation," due to the inadequate link up between the highly mathematical GTR and experiment. The later period, by contrast, although marked by the zoo of alternatives, is widely hailed as the "golden era" or "renaissance" of GTR.

The golden era came about thanks to events of 1959–1960 that set the stage for new confrontations between GTR's predictions and experiments. Nevertheless, the goals of this testing were not to decide if GTR was correct in all its implications, but rather, in the first place, to learn more about GTR (i.e., what does it really imply about experiments we can perform?) and, in the second place, to build models for phenomena that involve relativistic gravity (e.g., quasars, pulsars, gravity waves, and such). The goal was *to learn more about gravitational phenomena.*

Comparativist testing accounts, eager as they are to license the entire theory, ignore what for our severe tester is the central engine for making progress, for getting ideas for fruitful things to try next to learn more. This progress turned on distinguishing those portions of GTR that were and were not well tested. Far from arguing for GTR on the grounds that it had survived tests that existing alternatives could not, as our comparativist recommends, our severe tester would set about exploring just *why* we are *not* allowed to say that GTR is severely probed as a whole – in all the arenas in which gravitational effects may occur. Even without having 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?* Certainly we could be wrong with respect to predictions and domains that were not probed at all. But how could we be wrong even with respect to what GTR says about the probed regions, in particular, solar system tests? One must begin where one is.

Table 1.1. *The PPN parameters and their significance*

Parameter	What it measures relative to GTR	Values in GTR
λ	How much space-curvature produced by unit rest mass?	1
β	How much "nonlinearity" 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

Source: Adapted from Will (2005).

To this end, experimental relativists deliberately designed the PPN framework to prevent them from being biased toward accepting GTR prematurely (Will, 1993, p. 10), while allowing them to describe violations of GTR's hypotheses – discrepancies from what it said about specific gravitational phenomena in the solar system. The PPN framework set out a list of parameters that allowed for a systematic way of describing violations of GTR's hypotheses. These alternatives, by the physicists' own admissions, were set up largely as straw men with which to set firmer constraints on these parameters. The PPN formalism is used to get *relativistic* predictions rather than those from Newtonian theory – but in a way that is not biased toward GTR. It gets all the relativistic theories of gravity talking about the same things and to connect to the same data models (Mayo, 2002).

The PPN framework is limited to probing a portion or variant of GTR (see Table 1.1):

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, p. 10)

The PPN framework permitted researchers to compare the relative merits of various experiments ahead of time in probing the solar system approximation, or solar system variant, of GTR. Appropriately modeled astronomical data supply 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. This permitted the same PPN models of experiments to serve as intermediate links between the data and several alternative primary hypotheses based on GTR and its rival theories.

This mediation was a matter of measuring, or more correctly *inferring*, the values of PPN parameters by means of complex, statistical least-square fits to parameters in models of data. Although clearly much more would need to be said to explain how even one of the astrometric models is developed to design what are described as "high-precision null experiments," it is interesting to note that, even as the technology has advanced, the overarching reasoning shares much with the classic interferometry tests (e.g., those of Michelson and Morley). The GTR value for the PPN parameter under test serves as the null hypothesis from which discrepancies are sought. By identifying the null with the prediction from GTR, any discrepancies are given a very good chance to be detected; so, if no significant departure is found, this constitutes evidence for the GTR prediction with respect to the effect under test. 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 very high probative ability. For example, λ, the deflection-of-light parameter, measures "spatial curvature;" by setting the GTR predicted value to 1, modern tests infer upper bounds to violations (i.e., $|1 - \lambda|$). (See "Substantive Nulls," this volume, p. 264.)

Some elements of the series of models for the case of λ are sketched in Table 1.2.

The PPN framework is more than a bunch of parameters; it provides a general way to interpret the significance of the piecemeal tests for primary gravitational questions, including deciding to which questions a given test discriminates answers. Notably, its analysis revealed that one of the classic tests of GTR (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 null hypotheses (e.g., Eötvös experiments) that assert a zero difference in the accelerations of two differently composed bodies. The high precision with which these null hypotheses passed gave warrant to the inference that "gravity is a phenomenon of curved spacetime, that is, it must be described by a metric theory of gravity" (Will, 1993, p. 10).

For the comparativist, the corroboration of a part of GTR, such as the equivalence principle, is regarded as corroborating, defeasibly, GTR as a whole. In fact, however, corroborating the equivalence principle is recognized only as discriminating between so-called metric versus nonmetric gravitational theories, e.g., those gravity theories that do, versus those that

Table 1.2. *Elements of the series of models for the case of* λ

PRIMARY: Testing the post-Newtonian approximation of GTR
Parameterized 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
Example: For 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
Example: 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, one of which is a function of the observed statistic and the PPN parameter of interest (the function having known distribution)
Example: least-squares estimates of λ from "raw" data in eclipse and radio interferometry experiments.

DATA GENERATION AND ANALYSIS, EXPERIMENTAL DESIGN
Many details which a full account should include.

do not, satisfy this fundamental principle. This recognition only emerged once it was realized that all metric theories say the same thing with respect to the equivalence principle. Following point 2 above, they were not rivals with respect to this principle. More generally, an important task was to distinguish classes of experiments according to the specific aspects each probed and thus tested. An adequate account of the role and testing of theories must account for this, and the comparativist–holist view does not. The equivalence principle itself, more correctly called the Einstein equivalence principle, admitted of new partitions (e.g., into strong and weak, see later discussion), leading to further progress.[16]

---

[16] More carefully, we should identify the Einstein equivalence principle (EEP) as well as distinguish 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); and (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 stipulation of the equivalence principle also hold for self-gravitating bodies, such as the earth–moon system.

5.1 Criteria for a Viable Gravity Theory (during the "Golden Era")

From the outset, the PPN framework included not all logically possible gravity theories but those that passed the criteria for *viable* gravity theories.

(i) *It must be complete*, i.e., it must be capable of analyzing from "first principles" the outcome of any experiment of interest. It is not enough for the theory to *postulate* that bodies made of different material fall with the same acceleration... [This does not preclude "arbitrary parameters" being required for gravitational theories to accord with experimental results.]

(ii) *It must be self-consistent*, i.e., its prediction for the outcome of every experiment must be unique, so that when one calculates the predictions by two different, though equivalent methods, one always gets the same results...

(iii) *It must be relativistic*, i.e., in the limit as gravity is 'turned off' ... the nongravitational laws of physics must reduce to the laws of special relativity...

(iv) *It must have the correct Newtonian limit*, i.e., in the limit of weak gravitational fields and slow motions, it must reproduce Newton's laws... (Will, 1993, pp. 18–21).

From our perspective, viable theories must (1) account for experimental results already severely passed and (2) show the significance of the experimental data for gravitational phenomena.[17] Viable theories would have to be able to analyze and explore experiments about as well as GTR; there is a comparison here but remember that what makes a view "comparativist" is that it regards the full theory as well tested by dint of being "best tested so far." In our view, viable theories are required to pass muster for the goals to which they are put at this stage of advancing the knowledge of gravitational effects. One may regard these criteria as intertwined with the "pursuit" goals – that a theory should be useful for testing and learning more.

The experimental knowledge gained permits us to infer that we have a correct parameter value – but in our view it does more. It also indicates we have a correct understanding of how gravity behaves in a given domain. Different values for the parameters correspond to different mechanisms,

---

[17] Under consistency, it is required that the phenomenon it predicts be detectable via different but equivalent procedures. Otherwise they would be idiosyncratic to a given procedure and would not give us genuine, repeatable phenomena.

however abstract, at least in viable theories. For example, in 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 theoretical background is clearly what provides the interpretation of the relevance of the experimental effects for gravity, no one particular theory needs to be accepted to employ the PPN framework – which is at the heart of its robustness. Even later when this framework was extended to include nonmetric theories (in the fourth period, labeled "the search for strong gravitational effects"), those effects that had been vouchsafed with severity remain (although they may well demand reinterpretations).

5.2 Severity Logic and Some Paradoxes regarding Adjustable Constants

Under the completeness requirement for viable theories there is an explicit caveat that this does not preclude "arbitrary parameters" from being necessary for gravitational theories to obtain correct predictions, even though these are deliberately set to fit the observed effects and are not the outgrowth of "first principles." For example, the addition of a scalar field in Brans–Dicke theory went hand-in-hand with an adjustable constant w: the smaller its value the larger the effect of the scalar field and thus the bigger the difference with GTR, but as w gets larger the two become indistinguishable. (An interesting difference would have been with evidence that w is small, such as 40; its latest lower bound is pushing 20,000!) What should we make of the general status of the GTR rivals, given that their agreement with the GTR predictions and experiment required adjustable constants? This leads us to the general and much debated question of when and why data-dependent adjustments of theories and hypotheses are permissible.

The debate about whether to require or at least prefer (and even how to define) "novel" evidence is a fascinating topic in its own right, both in philosophy of science and statistics (Mayo, 1991, 1996), and it comes up again in several places in this volume (e.g., Chapters 4, 6, and 7); here, we consider a specific puzzle that arises with respect to experimental GTR. In particular, we consider how the consequences of severity logic disentangle apparently conflicting attitudes toward such "data-dependent constructions." Since all rivals were deliberately assured of fitting the effects thanks to their adjustable parameters, whereas GTR required no such adjustments, intuitively we tend to think that GTR was better tested by dint of its agreement with the experimental effects (e.g., Worrall, 1989). This leads the comparativist to reject such parameter adjustments. How then to explain the permissive attitude

toward the adjustments in experimental GTR? The comparativist cannot have it both ways.

By contrast, Bayesians seem to think they can. Those who wish to justify differential support look for it to show up in the prior probabilities, since all rivals fit the observed effects. Several Bayesians (e.g., Berger, Rosenkrantz) postulate that a theory that is free of adjustable parameters is "simpler" and therefore enjoys a higher prior probability; this would explain giving GTR higher marks for getting the predictions right than the Brans–Dicke theory or other rivals relying on adjustments (Jeffreys and Berger, 1992). But to explain why researchers countenance the parameter-fixing in GTR alternatives, other Bayesians maintain (as they must) that GTR should *not* be given a higher prior probability. Take Earman: "On the Bayesian analysis," this countenancing of parameter fixing "is not surprising, since it is not at all clear that GTR deserves a higher prior than the constrained Brans and Dicke theory" (Earman, 1992, p. 115). So Earman denies differential support is warranted in cases of parameter fixing ("why should the prior likelihood of the evidence depend upon whether it was used in constructing *T*?"; Earman, 1992, p. 116), putting him at odds with the Bayesian strategy for registering differential support (by assigning lower priors to theories with adjustable constants).

The Bayesian, like the comparativist, seems to lack a means to reflect, with respect to the *same* example, both (a) the intuition to give less credit to passing results that require adjustable parameters and (b) the accepted role, in practice, of deliberately constrained alternatives that are supported by the *same data* doing the constraining. Doubtless ways may be found, but would they avoid "ad hoc-ness" and capture what is actually going on?

To correctly diagnose the differential merit, the severe testing approach instructs us to consider the particular inference and the ways it can be in error in relation to the corresponding test procedure. There are two distinct analyses in the GTR case. First consider λ. The value for λ is fixed in GTR, and the data could be found to violate this fixed prediction by the procedure used for estimating λ (within its error margins). By contrast, in adjusting *w*, thereby constraining Brans–Dicke theory to fit the estimated λ, what is being learned regarding the Brans–Dicke theory is *how large would w need to be* to agree with the estimated λ? In this second case, inferences that pass with high severity are of the form "*w* must be at least 500." The questions, hence the possible errors, hence the severity differs.

But the data-dependent GTR alternatives play a second role, namely to show that GTR has not passed severely as a whole: They show that were a rival account of the mechanism of gravity correct, existing tests would not have detected this. In our view, this was the major contribution provided

by the rivals articulated in the PPN framework (of viable rivals to GTR). Even without being fully articulated, they effectively block GTR from having passed with severity as a whole (while pinpointing why). Each GTR rival gives different underlying accounts of the behavior of gravity (whether one wishes to call them distinct "mechanisms" or to use some other term). This space of rival explanations may be pictured as located at a higher level than the space of values of this parameter (Table 1.2). Considering the λ effect, the constrained GTR rivals succeed in showing that the existing experimental tests did not rule out, with severity, alternative explanations for the λ effect given in the viable rivals.[18] But the fact that a rival, say Brans–Dicke theory, served to block a high-severity assignment to GTR, given an experiment *E*, is not to say that *E* accords the rival high severity; it does not.

5.3 Nordvedt Effect η

To push the distinctions further, the fact that the rival Brans–Dicke theory is not severely tested (with *E*) is not the same as evidence against it (the severity logic has all sorts of interesting consequences, which need to be drawn out elsewhere). Evidence against it came later. Most notably, a surprise discovery in the 1960s (by Nordvedt) showed 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 16). This recognition was welcomed (apparently, even by Dicke) as a new way to test GTR as well as to learn more about gravity experiments.

Correspondingly, a new parameter to describe this effect, the Nordvedt effect, was introduced into the PPN framework (i.e., η). The parameter η would be 0 for GTR, so the null hypothesis tested is that $\eta = 0$ as against $\eta \neq 0$ for rivals. Measurements of the round-trip travel times between the Earth and the 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, p. 157) is credited with enabling the conflicting prediction to be identified. Because the tests were sufficiently sensitive, these measurements provided good evidence that the Nordvedt effect is absent, set upper bounds to the possible violations, and provided evidence

---

[18] Another way to see this is that the Brans–Dicke effect blocks high severity to the hypothesis about the specific nature of the gravitational cause of curvature – even without its own mechanism passing severely. For this task, they do not pay a penalty for accommodation; indeed, some view their role as estimating cosmological constants, thus estimating violations that would be expected in strong gravity domains.

for the correctness of what GTR says with respect to this effect – once again instantiating the familiar logic.[19]

5.4 Another Charge We Need to Tackle

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? Mayo's strictures about the limited scope of severity seem to preclude a positive answer to that question. (Laudan, 1997, p. 313)

In our view, there will not be a single answer, positive or negative. Whether *T*'s success in one part or range indicates it is likely to succeed (and to what extent) in another is an empirical question that must be answered on a case-by-case basis. Moreover, because this question seems to us to be the motivation for a good part of what scientists do in exploring theories, a single context-free answer would not even be desirable. But consider GTR: although one splits off the piecemeal tests, we do not face a disconnected array of results; indeed 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 Nordvedt parameter η gives estimates of several others. Because it is now recognized that highly precise estimates of λ constrain other parameters, λ is described as the fundamental parameter in some current discussions.

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 as depicted in Figure 1.1. In this way, entire chunks of theories can be 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).

---

[19] In the "secondary" task of scrutinizing the validity of the experiment, they asked, can other factors 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.
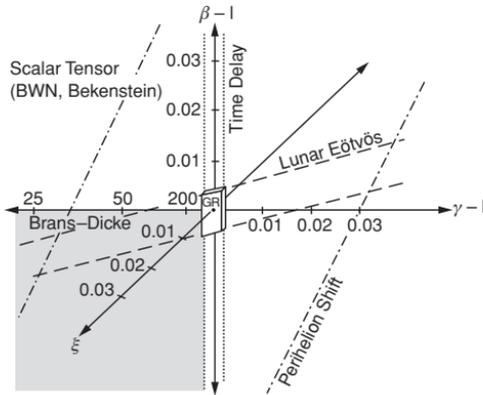
Figure 1.1.

5.5 Going beyond Solar System Tests

We can also motivate what happens next in this episode, although here I must be very brief. Progress is again made by recognizing the errors that are still not ruled out.

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, p. 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. Finding the decrease in the orbital period of this (Hulse-Taylor) binary pulsar at a rate in accordance with the GTR prediction of gravity wave energy loss is often regarded as the last event of the golden age. This example is fascinating in its own right, but we cannot take up a discussion here[20] (see Damour and Taylor, 1991; Lobo, 1996, pp. 212–15; Will, 1996).

There is clearly an interplay between theoretical and experimental considerations driving the program. For example, in the fourth and contemporary period, that of "strong gravity," a number of theoretical grounds indicate that GTR would require an extension or modification for strong gravitational fields – regions beyond the domains for which effects have been probed with severity. Although experimental claims (at a given level, as it

---

[20] For a brief discussion of how the hierarchy of models applies to the binary pulsar analysis, see Mayo (2000).

were) can remain stable through change of theory (at "higher" levels), it does not follow that experimental testing is unable to reach those theoretical levels. An error, as we see it, can concern any aspect of a model or hypothesis or mistaken understandings of an aspect of the phenomenon in question. For example, the severely tested results can remain while researchers consider alternative gravitational mechanisms in regimes not probed. Despite the latitude in these extended gravity models, by assuming only some general aspects on which all the extended models agree, they are able to design what are sometimes called "clean tests" of GTR; others, found sullied by uncertainties of the background physics, are entered in the logbooks for perhaps tackling with the next space shuttle![21] These analyses motivate new searches for very small deviations of relativistic gravity in the solar system that are currently present in the range of approximately 10^{-5}. Thus, probing new domains is designed to be played out in the solar system, with its stable and known results. This stability, however, does not go hand-in-hand with the kind of conservative attitude one tends to see in philosophies of theory testing: rather than hanker to adhere to well-tested theories, there seems to be a yen to find flaws potentially leading to new physics (perhaps a quantum theory of gravity).[22]

General relativity is now the "standard model" of gravity. But as in particle physics, there may be a world beyond the standard model. Quantum gravity, strings and branes may lead to testable effects beyond general relativity. Experimentalists will continue to search for such effects using laboratory experiments, particle accelerators, instruments in space and cosmological observations. At the centenary of relativity it could well be said that experimentalists have joined the theorists in relativistic paradise (Will, 2005, p. 27).

6 Concluding Remarks

Were one to pursue the error-statistical account of experiment at the level of large-scale theories, one would be interested to ask not "How can we severely pass high-level theories?" but rather, "How do scientists break

---

[21] Even "unclean" tests can rule out rivals that differ qualitatively from estimated effects. For example, Rosen's bimetric theory failed a "killing test" by predicting the reverse change in orbital period. "In fact we conjecture that for a wide class of metric theories of gravity, the binary pulsar provides the *ultimate* test of relativistic gravity" (Will, 1993, p. 287).

[22] According to Will, however, even achieving superunification would not overthrow the standard, macroscopic, or low-energy version of general relativity. Instead, any modifications are expected to occur at the Planck energy appropriate to the very early universe, or at singularities inside black holes.

down their questions about high-level theories into piecemeal questions that permit severe testing?" And how do the answers to these questions enable squeezing (if not exhausting) the space of predictions of a theory or of a restricted variant of a theory? We are not inductively eliminating one theory at a time, as in the typical "eliminative inductivism," but rather classes of theories, defined by giving a specified answer to a specific (experimental) question.

Note, too, that what is sought is not some way to talk about a measure of the degree of support or confirmation of one theory compared with another, but rather ways to measure how far off what a given theory says about a phenomenon can be from what a "correct" theory would need to say about it by setting *bounds on the possible violations*. Although we may not have a clue what the final correct theory of the domain in question would look like, the value of the experimental knowledge we can obtain now might be seen as giving us a glimpse of what a correct theory would say regarding the question of current interest, no matter how different the full theory might otherwise be.

References

Ben Haim, Y. (2001), *Information-Gap Decision Theory: Decisions Under Severe Uncertainty*, Academic Press, San Diego, CA.

Chalmers, A. (1999), *What Is This Thing Called Science?* 3rd ed., Open University Press, and University of Queensland Press.

Chalmers, A. (2002), "Experiment and the Growth of Experimental Knowledge," pp. 157–70 in *Proceedings of the International Congress for Logic, Methodology, and Philosophy of Science* (Vol. 1 of the 11th International Congress of Logic, Methodology, and Philosophy of Science, Cracow, August 1999), P. Gardenfors, J. Wolenski, and K. Kijania-Placek (eds.). Kluwer, Dordrecht, The Netherlands.

Cox, D.R. (2006), *Principles of Statistical Inference*, Cambridge University Press, Cambridge.

Damour, T., and Taylor, T.H. (1991), "On the Orbital Period Change of the Binary Pulsar PSR 1913 + 16," *Astrophysical Journal*, 366: 501–11.

Dorling, J. (1979), "Bayesian Personalism, the Methodology of Scientific Research Programmes, and Duhem's Problem," *Studies in History and Philosophy of Science*, 10: 177–87.

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

Fitelson, B. (2002), "Putting the Irrelevance Back into the Problem of Irrelevant Conjunction," *Philosophy of Science*, 69: 611–22.

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

Good, I.J. (1983), *Good Thinking*, University of Minnesota Press, Minneapolis.

Jeffreys, W., and Berger, J. (1992), "Ockham's Razor and Bayesian Analysis," *American Scientist*, 80: 64–72.

Kass, R.E., and Wasserman, L. (1996), "Formal Rules of Selecting Prior Distributions: A Review and Annotated Bibliography," *Journal of the American Statistical Association*, 91: 1343–70.

Kyburg, H.E., Jr. (1993), "The Scope of Bayesian Reasoning," in D. Hull, M. Forbes, and K. Okruhlik (eds.), *PSA 1992*, Vol. II, East Lansing, MI.

Laudan, L. (1977), *Progress and Its Problems*, University of California Press, Berkeley.

Laudan, L. (1997), "How About Bust? Factoring Explanatory Power Back into Theory Evaluation," *Philosophy of Science*, 64:303–16.

Lobo, J. (1996), "Sources of Gravitational Waves," pp. 203–22 in G.S. Hall and J.R. Pulham (eds.), *General Relativity: Proceedings of the Forty-Sixth Scottish Universities Summer School in Physics*, SUSSP Publications, Edinburgh, and Institute of Physics, London.

Mayo, D.G. (1991), "Novel Evidence and Severe Tests." *Philosophy of Science*, 58(4): 523–52.

Mayo, D.G. (1996), *Error and the Growth of Experimental Knowledge*, University of Chicago Press, Chicago.

Mayo, D.G. (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.G. (1997b), "Severe Tests, Arguing from Error, and Methodological Underdetermination," *Philosophical Studies*, 86: 243–66.

Mayo, D.G. (2000), "Experimental Practice and an Error Statistical Account of Evidence." *Philosophy of Science* 67, (Proceedings). Edited by D. Howard. Pages S193–S207.

Mayo, D.G. (2002), "Theory Testing, Statistical Methodology, and the Growth of Experimental Knowledge," pp. 171–90 in *Proceedings of the International Congress for Logic, Methodology, and Philosophy of Science* (Vol. 1 of the 11th International Congress of Logic, Methodology, and Philosophy of Science, Cracow, August 1999), P. Gardenfors, J. Wolenski, and K. Kijania-Placek (eds.). Kluwer, Dordrecht, The Netherlands.

Mayo, D.G. (2006), "Critical Rationalism and Its Failure to Withstand Critical Scrutiny," pp. 63–96 in C. Cheyne and J. Worrall (eds.), *Rationality and Reality: Conversations with Alan Musgrave*, Springer, Dordrecht.

Mayo, D.G., and Spanos, A. (2006), "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction," *British Journal of Philosophy of Science*, 57(2): 323–57.

Morrison, M., and Morgan, M. (eds.) (1999), *Models as Mediators: Perspectives on Natural and Social Science*, Cambridge University Press, Cambridge.

Suppes, P. (1969), "Models of Data," pp. 24–35 in *Studies in the Methodology and Foundations of Science*, D. Reidel, Dordrecht.

Will, C.M. (1980), "General Relativity," pp. 309–21 in J. Ehlers, J.J. Perry, and M. Walker (eds.), *Ninth Texas Symposium on Relativistic Astrophysics*, New York Academy of Sciences, New York.

Will, C.M. (1986), *Was Einstein Right?* Basic Books, New York (reprinted 1993).

Will, C.M. (1993), *Theory and Experiment in Gravitational Physics*, Cambridge University Press, Cambridge.

Will, C.M. (1996), "The Confrontation Between General Relativity and Experiment. A 1995 Update," pp. 239–81 in G.S. Hall and J.R. Pulham, *General Relativity: Proceedings*

of the Forty Sixth Scottish Universities Summer School in Physics, SUSSP Publications, Edinburgh, and Institute of Physics, London.

Will, C.M. (2004), "The Confrontation Between General Relativity and Experiment," *Living Reviews in Relativity*, http://relativity.livingreviews.org/Articles/lrr-2001-4/title.html.

Will, C.M. (2005), "Relativity at the Centenary," *Physics World*, 18: 27.

Worrall, J. (1989), "Fresnel, Poisson and the White Spot: The Role of Successful Predictions in the Acceptance of Scientific Theories," pp. 135–57 in D. Gooding, T. Pinch, and S. Schaffer (eds.), *The Uses of Experiment: Studies in the Natural Sciences*, Cambridge University Press, Cambridge.

Worrall, J. (1993), "Falsification, Rationality and the Duhem Problem: Grünbaum vs Bayes," pp. 329–70 in J. Earman, A.I. Janis, G.J. Massey, and N. Rescher (eds.), *Philosophical Problems of the Internal and External Worlds*, University of Pittsburgh Press, Pittsburgh.