Introduction and Background

Deborah G. Mayo and Aris Spanos

I Central Goals, Themes, and Questions

1 Philosophy of Science: Problems and Prospects

Methodological discussions in science have become increasingly common since the 1990s, particularly in fields such as economics, ecology, psychology, epidemiology, and several interdisciplinary domains – indeed in areas most faced with limited data, error, and noise. Contributors to collections on research methods, at least at some point, try to ponder, grapple with, or reflect on general issues of knowledge, inductive inference, or method. To varying degrees, such work may allude to philosophies of theory testing and theory change and philosophies of confirmation and testing (e.g., Popper, Carnap, Kuhn, Lakatos, Mill, Peirce, Fisher, Neyman-Pearson, and Bayesian statistics). However, the different philosophical "schools" tend to be regarded as static systems whose connections to the day-to-day questions about how to obtain reliable knowledge are largely metaphorical. Scientists might "sign up for" some thesis of Popper or Mill or Lakatos or others, but none of these classic philosophical approaches – at least as they are typically presented – provides an appropriate framework to address the numerous questions about the legitimacy of an approach or method.

Methodological discussions in science have also become increasingly sophisticated; and the more sophisticated they have become, the more they have encountered the problems of and challenges to traditional philosophical positions. The unintended consequence is that the influence of philosophy of science on methodological practice has been largely negative. If the philosophy of science – and the History and Philosophy of Science (HPS) – have failed to provide solutions to basic problems of evidence and inference, many practitioners reason, then how can they help scientists to look to philosophy of science to gain perspective? In this spirit, a growing tendency is to question whether anything can be said about what makes an enterprise scientific, or what distinguishes science from politics, art or other endeavors. Some works on methodology by practitioners look instead to sociology of science, perhaps to a variety of post-modernisms, relativisms, rhetoric and the like.

However, for the most part, scientists wish to resist relativistic, fuzzy, or postmodern turns; should they find themselves needing to reflect in a general way on how to distinguish science from pseudoscience, genuine tests from ad hoc methods, or objective from subjective standards in inquiry, they are likely to look to some of the classical philosophical representatives (and never mind if they are members of the list of philosophers at odds with the latest vogue in methodology). Notably, the Popperian requirement that our theories and hypotheses be testable and falsifiable is widely regarded to contain important insights about responsible science and objectivity; indeed, discussions of genuine versus ad hoc methods seem invariably to come back to Popper's requirement, even if his full philosophy is rejected. However, limiting scientific inference to deductive falsification without any positive account for warranting the reliability of data and hypotheses is too distant from day-to-day progress in science. Moreover, if we are to accept the prevalent skepticism about the existence of reliable methods for pinpointing the source of anomalies, then it is hard to see how to warrant falsifications in the first place.

The goal of this volume is to connect the methodological questions scientists raise to philosophical discussions on *Experimental Reasoning, Reliability, Objectivity, and Rationality* (E.R.R.O.R) of science. The aim of the "exchanges" that follow is to show that the real key to progress requires a careful unpacking of the central reasons that philosophy of science has failed to solve problems about evidence and inference. We have not gone far enough, we think, in trying to understand these obstacles to progress.

Achinstein (2001) reasons that, "scientists do not and should not take . . . philosophical accounts of evidence seriously" (p. 9) because they are based on a priori computations; whereas scientists evaluate evidence empirically. We ask: Why should philosophical accounts be a priori rather than empirical? Chalmers, in his popular book *What is This Thing Called Science?* denies that philosophers can say anything general about the character of scientific inquiry, save perhaps "trivial platitudes" such as "take evidence seriously" (Chalmers, 1999, p. 171). We ask: Why not attempt to answer the question of what it means to "take evidence seriously"? Clearly, one is not taking evidence seriously in appraising hypothesis *H* if it is predetermined that a way would be found to either obtain or interpret data as supporting *H*. If a procedure had little or no ability to find flaws in *H*, then finding none scarcely counts in *H*'s favor. One need not go back to the discredited caricature of the objective scientist as "disinterested" to extract an uncontroversial minimal requirement along the following lines:

Minimal Scientific Principle for Evidence. Data $\mathbf{x}_0$ provide poor evidence for *H* if they result from a method or procedure that has little or no ability of finding flaws in *H*, 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.

It seems fairly clear that this principle, or something very much like it, undergirds our intuition to disparage ad hoc rescues of hypotheses from falsification and to require hypotheses to be accepted only after subjecting them to criticism. *Why then has it seemed so difficult to erect an account of evidence that embodies this precept without running aground on philosophical conundrums?* By answering this question, we hope to set the stage for new avenues for progress in philosophy and methodology. Let us review some contemporary movements to understand better where we are today.

2 Current Trends and Impasses

Since breaking from the grip of the logical empiricist orthodoxy in the 1980s, the philosophy of science has been marked by attempts to engage dynamically with scientific practice:

1. Rather than a "white glove" analysis of the logical relations between statements of evidence ***e*** and hypothesis *H*, philosophers of science would explore the complex linkages among data, experiment, and theoretical hypotheses.
2. Rather than hand down pronouncements on ideally rational methodology, philosophers would examine methodologies of science empirically and naturalistically.

Two broad trends may be labeled the "new experimentalism" and the "new modeling." Moving away from an emphasis on high-level theory, the new experimentalists tell us to look to the manifold local tasks of distinguishing real effects from artifacts, checking instruments, and subtracting the effects of background factors (e.g., Chang, Galison, Hacking). Decrying the straightjacket of universal accounts, the new modelers champion the disunified and pluralistic strategies by which models mediate among data, hypotheses, and the world (Cartwright, Morgan, and Morrison). The historical record itself is an important source for attaining relevance to practice in the HPS movement.

Amid these trends is the broad move to tackle the philosophy of methodology empirically by looking to psychology, sociology, biology, cognitive science, or to the scientific record itself. As interesting, invigorating, and right-headed as the new moves have been, the problems of evidence and inference remain unresolved. *By and large, current philosophical work and the conceptions of science it embodies are built on the presupposition that we cannot truly solve the classic conundrums about induction and inference.* To give up on these problems, however, does not make them go away; moreover, the success of naturalistic projects demands addressing them. Appealing to "best-tested" theories of biology or cognitive science calls for critical evaluation of the methodology of appraisal on which these theories rest.

The position of the editors of this volume takes elements from each of these approaches (new experimentalism, empirical modeling, and naturalism). We think the classic philosophical problems about evidence and inference are highly relevant to methodological practice and, furthermore, *that they are solvable.* To be clear, we do not pin this position on any of our contributors! However, the exchanges with our contributors elucidate this stance. Taking naturalism seriously, we think we should appeal to the conglomeration of research methods for collecting, modeling, and learning from data in the face of limitations and threats of error – including modeling strategies and probabilistic and computer methods – all of which we may house under the very general rubric of the methodology of inductive-statistical modeling and inference. For us, statistical science will always have this broad sense covering experimental design, data generation and modeling, statistical inference methods, and their links to scientific questions and models. We also regard these statistical tools as lending themselves to informal analogues in tackling general philosophical problems of evidence and inference. Looking to statistical science would seem a natural, yet still largely untapped, resource for a naturalistic and normative approach to philosophical problems of evidence. Methods of experimentation, simulation, model validation, and data collection have become increasingly subtle and sophisticated, and we propose that philosophers of science revisit traditional problems with these tools in mind. In some contexts, even where literal experimental control is lacking, inquirers have learned how to determine "what it would be like" if we were able to intervene and control – at least with high probability. Indeed "the challenge, the fun, of outwitting and outsmarting drives us to find ways to learn what it would be like to control, manipulate, and change, in situations where we cannot" (Mayo, 1996, p. 458). This perspective lets us broaden the umbrella of what we regard as an "experimental" context. When we need to restore the more usual distinction between experimental and observational research, we may dub the former "manipulative experiment" and the latter "observational experiment."

The tools of statistical science are plagued with their own conceptual and epistemological problems – some new, many very old. It is important to our goals to interrelate themes from philosophy of science and philosophy of statistics.

- The first half of the volume considers issues of error and inference in philosophical problems of induction and theory testing.
- The second half illuminates issues of errors and inference in practice: in formal statistics, econometrics, causal modeling, and legal epistemology.

These twin halves reflect our conception of philosophy and methodology of science as a "two-way street": on the one hand there is an appeal to methods and strategies of local experimental testing to grapple with philosophical problems of evidence and inference; on the other there is an appeal to philosophical analysis to address foundational problems of the methods and models used in practice; see Mayo and Spanos (2004).

3 Relevance for the Methodologist in Practice

An important goal of this work is to lay some groundwork for the methodologist in practice, although it must be admitted that our strategy at first appears circuitous. We do not claim that practitioners' general questions about evidence and method are directly answered once they are linked to what professional philosophers have said under these umbrellas. Rather, we claim that it is by means of such linkages that practitioners may better understand the foundational issues around which their questions revolve. In effect, practitioners themselves may become better "applied philosophers," which seems to be what is needed in light of the current predicament in philosophy of science. Some explanation is necessary.

In the current predicament, methodologists may ask, if each of the philosophies of science have unsolved and perhaps insoluble problems about evidence and inference, then how can they be useful for evidential problems in practice? "If philosophers and others within science theory can't agree about the constitution of the scientific method . . . doesn't it seem a little dubious for economists to continue blithely taking things off the [philosopher's] shelf?" (Hands, 2001, p. 6). Deciding that it does, many methodologists in the social sciences tend to discount the relevance of the principles of scientific legitimacy couched within traditional philosophy of science. The philosophies of science are either kept on their shelves, or perhaps dusted off for cherry-picking from time to time. Nevertheless, practitioners still (implicitly or explicitly) wade into general questions about evidence or principles of inference and by elucidating the philosophical dimensions of such problems we hope to empower practitioners to appreciate and perhaps solve them. In a recent lead article in the journal *Statistical Science*, we read that "professional agreement on statistical philosophy is not on the immediate horizon, but this should not stop us from agreeing on methodology" (Berger, 2003, p. 2). But we think "what is correct methodologically" depends on "what is correct philosophically" (Mayo, 2003). Otherwise, choosing between competing methods and models may be viewed largely as a matter of pragmatics without posing deep philosophical problems or inconsistencies of principle. For the "professional agreement" to have weight, it cannot be merely an agreement to use methods with similar numbers when the meaning and import of such numbers remain up in the air (see Chapter 7). We cannot wave a wand and bring into existence the kind of philosophical literature that we think is needed. What we can do is put the practitioner in a better position to support, or alternatively, question the basis for professional agreement or disagreement.

Another situation wherein practitioners may find themselves wishing to articulate general principles or goals is when faced with the need to modify existing methods and to make a case for the adoption of new tools. Here, practitioners may serve the dual role of both inventing new methods and providing them with a principled justification – possibly by striving to find, or adapt features from, one or another philosophy of science or philosophy of statistics. Existing philosophy of science may not provide off-the-shelf methods for answering methodological problems in practice, but, coupled with the right road map, it may enable understanding, or even better, solving those problems.

An illustration in economics is given by Aris Spanos (Chapter 6). Faced with the lack of literal experimental controls, some economic practitioners attempt to navigate between two extreme positions. One position is the prevailing theory-dominated empirical modeling, largely limited to quantifying theories presupposed to be true. At the other extreme is data-driven modeling, largely limited to describing the data and guided solely by goodness-of-fit criteria. The former stays too close to the particular theory chosen at the start; the second stays too close to the particular data. Those practitioners seeking a "third way" are implicitly thrust into the role of striving to locate a suitable epistemological foundation for a methodology seemingly at odds with the traditional philosophical image of the roles of theory and data in empirical inquiry. In other words, the prescriptions on method in practice have trickled down from (sometimes competing) images of good science in traditional philosophy. We need to ask the question: *What are the threats to reliability and objectivity that lay behind the assumed prescriptions to begin with?* If data-dependent methods are thought to require the assumption of an overarching theory, or else permit too much latitude in constructing theories to fit data, then much of social science appears to be guilty of violating a scientific canon. But in practice, some econometricians work to develop methods whereby the data may be used to provide independent constraints on theory testing by means of intermediate-level statistical models with a "life of their own," as it were. This is the key to evading threats to reliability posed by theory-dominated modeling. By grasping the philosophical issues and principles, such applied work receives a stronger and far less tenuous epistemological foundation.

This brings us to a rather untraditional connection to traditional philosophy of science. In several of the philosophical contributions in this volume, we come across the very conceptions of testing that practitioners may find are in need of tweaking or alteration in order to adequately warrant methods they wish to employ. By extricating the legitimate threats to reliability and objectivity that lie behind the traditional stipulations, practitioners may ascertain where and when violations of established norms are justifiable. The exchange essays relating to the philosophical contributions deliberately try to pry us loose from rigid adherence to some of the standard prescriptions and prohibitions.

In this indirect manner, the methodologists' real-life problems are connected to what might have seemed at first an arcane philosophical debate. Insofar as these connections have not been made, practitioners are dubious that philosophers' debates about evidence and inference have anything to do with, much less help solve, their methodological problems. We think the situation is otherwise – that getting to the underlying philosophical issues not only increases the intellectual depth of methodological discussions but also paves the way for solving problems.

We find this strategy empowers students of methodology to evaluate critically, and perhaps improve on, methodologies in practice. Rather than approach alternative methodologies in practice as merely a menu of positions from which to choose, they may be grasped as attempted solutions to problems with deep philosophical roots. Conversely, progress in methodology may challenge philosophers of science to reevaluate the assumptions of their own philosophical theories. That is, after all, what a genuinely naturalistic philosophy of method would require. A philosophical problem, once linked to methodology in practice, enjoys solutions from the practical realm. For example, philosophers tend to assume that there are an infinite number of models that fit finite data equally well, and so data underdetermine hypotheses. Replacing "fit" with more rigorous measures of adequacy can show that such underdetermination vanishes (Spanos, 2007). This brings us to the last broad topic we consider throughout the volume.

We place it under the heading of *metaphilosophical themes*. Just as we know that evidence in science may be "theory-laden" – interpreted from the perspective of a background theory or set of assumptions – our philosophical theories (about evidence, inference, science) often color our philosophical arguments and conclusions (Rosenberg, 1992). The contributions in this volume reveal a good deal about these "philosophy-laden" aspects of philosophies of science. These revelations, moreover, are directly relevant to what is needed to construct a sound foundation for methodology in practice. The payoff is that understanding the obstacles to solving philosophical problems (the focus of Chapters 1–5) offers a clear comprehension of how to relate traditional philosophy of science to contemporary methodological and foundational problems of practice (the focus of Chapters 6–9).

4 Exchanges on E.R.R.O.R.

We organize the key themes of the entire volume under two interrelated categories:

(1) *experimental reasoning (empirical inference) and reliability*, and
(2) *objectivity and rationality of science*.

Although we leave these terms ambiguous in this introduction, they will be elucidated as we proceed. Interrelationships between these two categories immediately emerge. Scientific rationality and objectivity, after all, are generally identified by means of scientific methods: one's conception of objectivity and rationality in science leads to a conception of the requirements for an adequate account of empirical inference and reasoning. The perceived ability or inability to arrive at an account satisfying those requirements will in turn direct one's assessment of the possibility of objectivity and rationality in science. Recognizing the intimate relationships between categories 1 and 2 propels us toward both understanding and making progress on recalcitrant foundational problems about scientific inference. If, for example, empirical inference is thought to demand reliable rules of inductive inference, and if it is decided that such rules are unobtainable, then one may either question the rationality of science or instead devise a different notion of rationality for which empirical methods exist. On the other hand, if we are able to show that some methods are more robust than typically assumed, we may be entitled to uphold a more robust conception of science. Under category 1, we consider the nature and justification of experimental reasoning and the relationship of experimental inference to appraising large-scale theories in science.

4.1 Theory Testing and Explanation

Several contributors endorse the view that scientific progress is based on accepting large-scale theories (e.g., Chalmers, Musgrave) as contrasted to a view of progress based on the growth of more localized experimental knowledge (Mayo). Can one operate with a single overarching view of what is required for data to warrant an inference to *H*? Mayo says yes, but most of the other contributors argue for multiple distinct notions of evidence and inference. They do so for very different reasons. Some argue for a distinction between large-scale theory testing and local experimental inference. When it comes to large-scale theory testing, some claim that the most one can argue is that a theory is, comparatively, the best tested so far (Musgrave), or that a theory is justified by an “argument from coincidence” (Chalmers). Others argue that a distinct kind of inference is possible when the data are “not used” in constructing hypotheses or theories (“use-novel” data), as opposed to data-dependent cases where an inference is, at best, conditional on a theory (Worrall). Distinct concepts of evidence might be identified according to different background knowledge (Achinstein). Finally, different standards of evidence may be thought to emerge from the necessity of considering different costs (Laudan). The relations between testing and explanation often hover in the background of the discussion, or they may arise explicitly (Chalmers, Glymour, Musgrave). What are the explanatory virtues? And how do they relate to those of testing? Is there a tension between explanation and testing?

4.2 What Are the Roles of Probability in Uncertain Inference in Science?

These core questions are addressed both in philosophy of science, as well as in statistics and modeling practice. Does probability arise to assign degrees of epistemic support or belief to hypotheses, or to characterize the reliability of rules? A loose analogy exists between Popperian philosophers and frequentist statisticians, on the one hand, and Carnapian philosophers and Bayesian statisticians on the other. The latter hold that probability needs to supply some degree of belief, support, or epistemic assignment to hypotheses (Achinstein), a position that Popperians, or critical rationalists, dub 'justificationism' (Musgrave). Denying that such degrees may be usefully supplied, Popperians, much like frequentists, advocate focusing on the rationality of rules for inferring, accepting, or believing hypotheses. But what properties must these rules have?

In formal statistical realms, the rules for inference are reliable by dint of controlling error probabilities (Spanos, Cox and Mayo, Glymour). Can analogous virtues be applied to informal realms of inductive inference? This is the subject of lively debate in Chapters 1 to 5 in this volume. However, statistical methods and models are subject to their own long-standing foundational problems. Chapters 6 and 7 offer a contemporary update of these problems from the frequentist philosophy perspective. Which methods can be shown to ensure reliability or low long-run error probabilities? Even if we can show they have good long-run properties, how is this relevant for a particular inductive inference in science? These chapters represent exchanges and shared efforts of the authors over the past four years to tackle these problems as they arise in current statistical methodology. Interwoven throughout this volume we consider the relevance of these answers to analogous questions as they arise in philosophy of science.

4.3 Objectivity and Rationality of Science, Statistics, and Modeling

Despite the multiplicity of perspectives that the contributors bring to the table, they all find themselves confronting a cluster of threats to objectivity in observation and inference. Seeing how analogous questions arise in philosophy and methodological practice sets the stage for the meeting ground that creates new synergy.

- Does the fact that observational claims themselves have assumptions introduce circularity into the experimental process?
- Can one objectively test assumptions linking actual data to statistical models, and statistical inferences to substantive questions?

On the one hand, the philosophers' demand to extricate assumptions raises challenges that the practitioner tends to overlook; on the other hand, progress in methodology may point to a more subtle logic that gets around the limits that give rise to philosophical skepticism.

What happens if methodological practice seems in conflict with philosophical principles of objectivity? Some methodologists reason that if it is common, if not necessary, to violate traditional prescriptions of scientific objectivity in practice, then we should renounce objectivity (and perhaps make our subjectivity explicit). That judgment is too quick. If intuitively good scientific practice seems to violate what are thought to be requirements of good science, we need to consider whether in such cases scientists guard against the errors that their violation may permit.

To illustrate, consider one of the most pervasive questions that arises in trying to distinguish genuine tests from ad hoc methods:

Is it legitimate to use the same data in both constructing and testing hypotheses?

This question arises in practice in terms of the legitimacy of data-mining, double counting, data-snooping, and hunting for statistical significance. In philosophy of science, it arises in terms of novelty requirements. Musgrave (1974) was seminal in tackling the problems of how to define, and provide a rationale for, preferring novel predictions in the Popper-Lakatos traditions. However, these issues have never been fully resolved, and they continue to be a source of debate. The question of the rationale for requiring novelty arises explicitly in Chapter 4 (Worrall) and the associated exchange.

Lurking in the background of all of the contributions in this volume is the intuition that good tests should avoid double-uses of data, that would result in violating what we called the *minimal scientific principle for evidence.* Using the same data to construct as well as test a hypothesis, it is feared, makes it too easy to find accordance between the data and the hypothesis even if the hypothesis is false. By uncovering how reliable learning may be retained despite double-uses of data, we may be able to distinguish legitimate from illegitimate double counting.

The relevance of this debate for practice is immediately apparent in the second part of the volume where several examples of data-dependent modeling and non-novel evidence arise: in accounting for selection effects, in testing assumptions of statistical models, in empirical modeling in economics, in algorithms for causal model discovery, and in obtaining legal evidence.

This leads to our third cluster of issues that do not readily fit under either category (1) or (2) – the host of "meta-level" issues regarding philosophical assumptions (theory-laden philosophy) and the requirements of a successful two-way street between philosophy of science and methodological practice.

The questions listed in Section 6 identify the central themes to be taken up in this volume. The essays following the contributions are called "exchanges" because they are the result of a back-and-forth discussion over a period of several years. Each exchange begins by listing a small subset of these questions that is especially pertinent for reflecting on the particular contribution.

5 Using This Volume for Teaching

Our own experiences in teaching courses that blend philosophy of science and methodology have influenced the way we arrange the material in this volume. We have found it useful, for the first half of a course, to begin with a core methodological paper in the given field, followed by selections from the philosophical themes of Chapters 1–5, supplemented with 1–2 philosophical articles from the references (e.g., from Lakatos, Kuhn, Popper). Then, one might turn to selections from Chapters 6–9, supplemented with discipline-specific collections of papers.* The set of questions listed in the next section serves as a basis around which one might organize both halves of the course. Because the exchange that follows each chapter elucidates some of the key points of that contribution, readers may find it useful to read or glance at the exchange first and then read the corresponding chapter.

6 Philosophical and Methodological Questions Addressed in This Volume

6.1 Experimental Reasoning and Reliability

Theory Testing and Explanation

- Does theory appraisal demand a kind of reasoning distinct from local experimental inferences?
- Can generalizations and theoretical claims ever be warranted with severity?
- Are there reliable observational methods for discovering or inferring causes?
- How can the gap between statistical and structural (e.g., causal) models be bridged?

* A variety of modules for teaching may be found at the website: http://www.econ.vt.edu/faculty/facultybios/spanos_error_inference.htm.

- Must local experimental tests always be done within an overarching theory or paradigm? If so, in what sense must the theory be assumed or accepted?
- When does *H*'s successful explanation of an effect warrant inferring the truth or correctness of *H*?
- How do logical accounts of explanation link with logics of confirmation and testing?

How to Characterize and Warrant Methods of Experimental Inference

- Can inductive or "ampliative" inference be warranted?
- Do experimental data so underdetermine general claims that warranted inferences are limited to the specific confines in which the data have been collected?
- Can we get beyond inductive skepticism by showing the existence of reliable test rules?
- Can experimental virtues (e.g., reliability) be attained in nonexperimental contexts?
- How should probability enter into experimental inference and testing: by assigning degrees of belief or by characterizing the reliability of test procedures?
- Do distinct uses of data in science require distinct criteria for warranted inferences?
- How can methods for controlling long-run error probabilities be relevant for inductive inference in science?

6.2 Objectivity and Rationality of Science

- Should scientific progress and rationality be framed in terms of large-scale theory change?
- Does a piecemeal account of explanation entail a piecemeal account of testing?
- Does an account of progress framed in terms of local experimental inferences entail a nonrealist role for theories?
- Is it unscientific (ad hoc, degenerating) to use data in both constructing and testing hypotheses?
- Is double counting problematic only when it leads to unreliable methods?
- How can we assign degrees of objective warrant or rational belief to scientific hypotheses?

- How can we assess the probabilities with which tests lead to erroneous inferences (error probabilities)?
- Can an objective account of statistical inference be based on frequentist methods? On Bayesian methods?
- Can assumptions of statistical models and methods be tested objectively?
- Can assumptions linking statistical inferences to substantive questions be tested objectively?
- What role should probabilistic/statistical accounts play in scrutinizing methodological desiderata (e.g., explanatory virtues) and rules (e.g., avoiding irrelevant conjunction, varying evidence)?
- Do explanatory virtues promote truth, or do they conflict with well-testedness?
- Does the latitude in specifying tests and criteria for accepting and rejecting hypotheses preclude objectivity?
- Are the criteria for warranted evidence and inference relative to the varying goals in using evidence?

6.3 Metaphilosophical Themes

Philosophy-Laden Philosophy of Science

- How do assumptions about the nature and justification of evidence and inference influence philosophy of science? In the use of historical episodes?
- How should we evaluate philosophical tools of logical analysis and counterexamples?
- How should probabilistic/statistical accounts enter into solving philosophical problems?

Responsibilities of the "Two-Way Street" between Philosophy and Practice

- What roles can or should philosophers play in methodological problems in practice? (Should they be in the business of improving practice as well as clarifying, reconstructing, or justifying practice?)
- How does studying evidence and methods in practice challenge assumptions that may go unattended in philosophy of science?

II The Error-Statistical Philosophy

The Preface of *Error and the Growth of Experimental Knowledge* (EGEK) opens as follows:

> Despite the challenges to and changes in traditional philosophy of science, one of its primary tasks continues to be to explain if not also to justify, scientific methodologies for learning about the world. To logical empiricist philosophers (Carnap, Reichenbach) the task was to show that science proceeds by objective rules for appraising hypotheses. To that end many attempted to set out formal rules termed inductive logics and confirmation theories. Alongside these stood Popper's methodology of appraisal based on falsification: evidence was to be used to falsify claims deductively rather than to build up inductive support. Both inductivist and falsificationist approaches were plagued with numerous, often identical, philosophical problems and paradoxes. Moreover, the entire view that science follows impartial algorithms or logics was challenged by Kuhn (1962) and others. What methodological rules there are often conflict and are sufficiently vague as to "justify" rival hypotheses. Actual scientific debates often last for several decades and appear to require, for their adjudication, a variety of other factors left out of philosophers' accounts. The challenge, if one is not to abandon the view that science is characterized by rational methods of hypothesis appraisal, is either to develop more adequate models of inductive inference or else to find some new account of scientific rationality. (Mayo, 1996, p. ix)

Work in EGEK sought a more adequate account of induction based on a cluster of tools from statistical science, and this volume continues that program, which we call the error-statistical account.

Contributions to this volume reflect some of the "challenges and changes" in philosophy of science in the dozen years since EGEK, and the ensuing dialogues may be seen to move us "Toward an Error-Statistical Philosophy of Science" – as sketchily proposed in EGEK's last chapter. Here we collect for the reader some of its key features and future prospects.

7 What Is Error Statistics?

Error statistics, as we use the term, has a dual dimension involving philosophy and methodology. It refers to a standpoint regarding both (1) a general philosophy of science and the roles probability plays in inductive inference, and (2) a cluster of statistical tools, their interpretation, and their justification. It is unified by a general attitude toward a fundamental pair of questions of interest to philosophers of science and scientists in general:

- *How do we obtain reliable knowledge about the world despite error?*
- *What is the role of probability in making reliable inferences?*

Here we sketch the error-statistical methodology, the statistical philosophy associated with the methods ("error-statistical philosophy"), and a philosophy of science corresponding to the error-statistical philosophy.

7.1 Error-Statistical Philosophy

Under the umbrella of error-statistical methods, one may include all standard methods using error probabilities based on the relative frequencies of errors in repeated sampling – often called *sampling theory*. In contrast to traditional confirmation theories, probability arises not to measure degrees of confirmation or belief in hypotheses but to quantify how frequently methods are capable of discriminating between alternative hypotheses and how reliably they facilitate the detection of error. These probabilistic properties of inference procedures are *error frequencies* or *error probabilities*. The statistical methods of significance tests and confidence-interval estimation are examples of formal error-statistical methods. Questions or problems are addressed by means of hypotheses framed within statistical models.

A statistical model (or family of models) gives the probability distribution (or density) of the sample $\mathbf{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n})$, $f_X(\mathbf{x}; \boldsymbol{\theta})$, which provides an approximate or idealized representation of the underlying data-generating process. Statistical hypotheses are typically couched in terms of an unknown parameter, $\boldsymbol{\theta}$, which governs the probability distribution (or density) of $\mathbf{X}$. Such hypotheses are claims about the data-generating process. In error statistics, statistical inference procedures link special functions of the data, $d(\mathbf{X})$, known as *statistics*, to hypotheses of interest. All error probabilities stem from the distribution of $d(\mathbf{X})$ evaluated under different hypothetical values of parameter $\boldsymbol{\theta}$.

Consider for example the case of a random sample $\mathbf{X}$ of size n from a Normal distribution ($\mathrm{N}(\mu,1)$) where we want to test the hypotheses:

$$H_0\text{: } \mu = \mu_0 \text{ vs. } H_1\text{: } \mu > \mu_0.$$

The test statistic is $d(\mathbf{X}) = (\overline{X} - \mu_0)/\sigma_x$, where $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$ and $\sigma_x = (\sigma/\sqrt{n})$. Suppose the test rule T construes data $\mathbf{x}$ as evidence for a discrepancy from μ_0 whenever $d(\mathbf{x}) > 1.96$. The probability that the test would indicate such evidence when in fact μ_0 is true is $\mathrm{P}(d(\mathbf{X}) > 1.96; H_0) = .025$. This gives us what is called the *statistical significance level.* Objectivity stems from controlling the relevant error probabilities associated with the particular inference procedure. In particular, the claimed error probabilities approximate the actual (long-run) relative frequencies of error. (See Chapters 6 and 7.)

Behavioristic and Evidential Construal. By a “statistical philosophy” we understand a general concept of the aims and epistemological foundations of a statistical methodology. To begin with, two different interpretations of these methods may be given, along with diverging justifications. The first, and most well known, is the *behavioristic construal.* In this case, tests are interpreted as tools for deciding “how to behave” in relation to the phenomena under test and are justified in terms of their ability to ensure low long-run errors. A nonbehavioristic or *evidential construal* must interpret error-statistical tests (and other methods) as tools for achieving inferential and learning goals. How to provide a satisfactory evidential construal has been the locus of the most philosophically interesting controversies and remains the major lacuna in using these methods for philosophy of science. This is what the severity account is intended to supply. However, there are contexts wherein the more behavioristic construal is entirely appropriate, and it is retained within the “error-statistical” umbrella.

Objectivity in Error Statistics. The inferential interpretation forms a central part of what we refer to as *error-statistical philosophy.* Underlying this philosophy is the concept of scientific objectivity: although knowledge gaps leave plenty of room for biases, arbitrariness, and wishful thinking, in fact we regularly come up against experiences that thwart our expectations and disagree with the predictions and theories we try to foist upon the world – this affords objective constraints on which our critical capacity is built. Getting it (at least approximately) right, and not merely ensuring internal consistency or agreed-upon convention, is at the heart of objectively orienting ourselves toward the world. Our ability to recognize when data fail to match anticipations is what affords us the opportunity to systematically improve our orientation in direct response to such disharmony. Failing to falsify hypotheses, while rarely allowing their acceptance as true, warrants the exclusion of various discrepancies, errors, or rivals, provided the test had a high probability of uncovering such flaws, if they were present. In those cases, we may infer that the discrepancies, rivals, or errors are ruled out with *severity*.

We are not stymied by the fact that inferential tools have assumptions but rather seek ways to ensure that the validity of inferences is not much threatened by what is currently unknown. This condition may be secured either because tools are robust against flawed assumptions or that subsequent checks will detect (and often correct) them with high probability. Attributes that go unattended in philosophies of confirmation occupy important places in an account capable of satisfying error-statistical goals. For example, explicit attention needs to be paid to communicating results to set the stage for others to check, debate, and extend the inferences reached. In this view, it must be part of any adequate statistical methodology to provide the means to address critical questions and to give information about which conclusions are likely to stand up to further probing and where weak spots remain.

Error-Statistical Framework of "Active" Inquiry. The error-statistical philosophy conceives of statistics (or statistical science) very broadly to include the conglomeration of systematic tools for collecting, modeling, and drawing inferences from data, including purely "data-analytic" methods that are normally not deemed "inferential." For formal error-statistical tools to link data, or *data models*, to *primary scientific hypotheses*, several different statistical hypotheses may be called upon, each permitting an aspect of the primary problem to be expressed and probed. An auxiliary or "secondary" set of hypotheses is called upon to check the assumptions of other models in the complex network.

The error statistician is concerned with the critical control of scientific inferences by means of stringent probes of conjectured flaws and sources of unreliability. Standard statistical hypotheses, while seeming oversimplified in and of themselves, are highly flexible and effective for the piecemeal probes our error statistician seeks. Statistical hypotheses offer ways to couch canonical flaws in inference. We list six overlapping errors:

1. Mistaking spurious for genuine correlations,
2. Mistaken directions of effects,
3. Mistaken values of parameters,
4. Mistakes about causal factors,
5. Mistaken assumptions of statistical models,
6. Mistakes in linking statistical inferences to substantive scientific hypotheses.

The qualities we look for to express and test hypotheses about such inference errors are generally quite distinct from those traditionally sought in appraising substantive scientific claims and theories. Although the overarching goal is to find out what is (truly) the case about aspects of phenomena, the hypotheses erected in the actual processes of finding things out are generally approximations and may even be deliberately false. Although we cannot fully formalize, we can systematize the manifold steps and interrelated checks that, taken together, constitute a full-bodied experimental inquiry. Background knowledge enters the processes of designing, interpreting, and combining statistical inferences in informal or semiformal ways – not, for example, by prior probability distri-butions.

The picture corresponding to error statistics is one of an activist learner in the midst of an inquiry with the goal of finding something out. We want hypotheses that will allow for stringent testing so that if they pass we have evidence of a genuine experimental effect. The goal of attaining such well-probed hypotheses differs crucially from seeking highly probable ones (however probability is interpreted). This recognition is the key to getting a handle on long-standing Bayesian–frequentist debates.

The error statistical philosophy serves to guide the use and interpretation of frequentist statistical tools so that we can distinguish the genuine foundational differences from a host of familiar fallacies and caricatures that have dominated 75 years of "statistics wars." The time is ripe to get beyond them.

7.2 Error Statistics and Philosophy of Science

The *error-statistical philosophy* alludes to the general methodological principles and foundations associated with frequentist error-statistical methods;

it is the sort of thing that would be possessed by a statistician, when thinking foundationally, or by a philosopher of statistics. By an *error-statistical philosophy of science*, on the other hand, we have in mind the use of those tools, appropriately adapted, to problems of philosophy of science: to model scientific inference (actual or rational), to scrutinize principles of inference (e.g., preferring novel results, varying data), and to frame and tackle philosophical problems about evidence and inference (how to warrant data, pinpoint blame for anomalies, and test models and theories). Nevertheless, each of the features of the error-statistical philosophy has direct consequences for the philosophy of science dimension.

To obtain a philosophical account of inference from the error-statistical perspective, one would require forward-looking tools for finding things out, not for reconstructing inferences as "rational" (in accordance with one or another view of rationality). An adequate philosophy of evidence would have to engage statistical methods for obtaining, debating, rejecting, and affirming data. From this perspective, an account of scientific method that begins its work only once well-defined evidence claims are available forfeits the ability to be relevant to understanding the actual processes behind the success of science. Because the contexts in which statistical methods are most needed are ones that compel us to be most aware of the strategies scientists use to cope with threats to reliability, the study of the nature of statistical method in the collection, modeling, and analysis of data is an effective way to articulate and warrant principles of evidence. In addition to paving the way for richer and more realistic philosophies of science, we think, examining error-statistical methods sets the stage for solving or making progress on long-standing philosophical problems about evidence and inductive inference.

Where the recognition that data are always fallible presents a challenge to traditional empiricist foundations, the cornerstone of statistical induction is the ability to move from less accurate to more accurate data.

Where the best often thought "feasible" means getting it right in some asymptotic long run, error-statistical methods enable specific precision to be ensured in finite samples and supply ways to calculate how large the sample size *n* needs to be for a given level of accuracy.

Where pinpointing blame for anomalies is thought to present insoluble "Duhemian problems" and "underdetermination," a central feature of error-statistical tests is their capacity to evaluate error probabilities that hold regardless of unknown background or "nuisance" parameters.

We now consider a principle that links (1) the error-statistical philosophy and (2) an error-statistical philosophy of science.

7.3 The Severity Principle

A method's error probabilities refer to their performance characteristics in a hypothetical sequence of repetitions. How are we to use error probabilities of tools in warranting particular inferences? This leads to the general question:

When do data $\mathbf{x}_0$ provide good evidence for or a good test of hypothesis H?

Our standpoint begins with the intuition described in the first part of this chapter. We intuitively deny that data $\mathbf{x}_0$ are evidence for *H* if the inferential procedure had very little chance of providing evidence against *H*, even if *H* is false. We can call this the "weak" severity principle:

Severity Principle (Weak)**:** Data $\mathbf{x}_0$ do not provide good evidence for hypothesis *H* if $\mathbf{x}_0$ result from a test procedure with a very low probability or capacity of having uncovered the falsity of *H* (even if *H* is incorrect).

Such a test, we would say, is insufficiently stringent or severe. The onus is on the person claiming to have evidence for *H* to show that the claim is not guilty of at least so egregious a lack of severity. Formal error-statistical tools provide systematic ways to foster this goal and to determine how well it has been met in any specific case. Although one might stop with this negative conception (as perhaps Popperians do), we continue on to the further, positive conception, which will comprise the full severity principle:

Severity Principle (Full)**:** Data $\mathbf{x}_0$ provide a good indication of or evidence for hypothesis *H* (just) to the extent that test *T* has severely passed *H* with $\mathbf{x}_0$.

The severity principle provides the rationale for error-statistical methods. We distinguish the severity rationale from a more prevalent idea for how procedures with low error probabilities become relevant to a particular application; namely, since the procedure is rarely wrong, the probability it is wrong in this case is low. In that view, we are justified in inferring *H* because it was the output of a method that rarely errs. It is as if the long-run error probability "rubs off" on each application. However, this approach still does not quite get at the reasoning for the particular case at hand, at least in nonbehavioristic contexts. The reliability of the rule used to infer *H* is at most a necessary and not a sufficient condition to warrant inferring *H*. All of these ideas will be fleshed out throughout the volume.

Passing a severe test can be encapsulated as follows:

A hypothesis H passes a severe test T with data $\mathbf{x}_0$ *if*

(S-1) $\mathbf{x}_0$ *agrees with H, (for a suitable notion of "agreement") and*
(S-2) *with very high probability, test T would have produced a result that accords less well with H than does* $\mathbf{x}_0$, *if H were false or incorrect.*

Severity, in our conception, somewhat in contrast to how it is often used, is not a characteristic of a test in and of itself, but rather of the test T, a specific test result $\mathbf{x}_0$, and a specific inference being entertained, H. Thereby, the severity function has three arguments. We use SEV(T, $\mathbf{x}_0$, H) to abbreviate "the severity with which H passes test T with data $\mathbf{x}_0$" (Mayo and Spanos, 2006).

The existing formal statistical testing apparatus does not include severity assessments, but there are ways to *use* the error-statistical properties of tests, together with the outcome $\mathbf{x}_0$, to evaluate a test's severity. This is the key for our inferential interpretation of error-statistical tests. The severity principle underwrites this inferential interpretation and addresses chronic fallacies and well-rehearsed criticisms associated with frequentist testing. Among the most familiar of the often repeated criticisms of the use of significance tests is that with large enough sample size, a small significance level can be very probable, even if the underlying discrepancy γ from null hypothesis $\mu = \mu_0$ is substantively trivial. Why suppose that practitioners are incapable of mounting an interpretation of tests that reflects the test's sensitivity? The severity assessment associated with the observed significance level [p-value] directly accomplishes this.

Let us return to the example of test T for the hypotheses: H_0: $\mu = 0$ vs. H_1: $\mu > 0$. We see right away that the same value of $d(\mathbf{x}_0)$ (and thus the same p-value) gives different severity assessments for a given inference when n changes.

In particular, suppose one is interested in the discrepancy $\gamma = .2$, so we wish to evaluate the inference $\mu > .2$. Suppose the same $d(\mathbf{x}_0) = 3$ resulted from two different sample sizes, $n = 25$ and $n = 400$. For $n = 25$, the severity associated with $\mu > .2$ is .97, but for $n = 400$ the severity associated with $\mu > .2$ is .16. So the same $d(\mathbf{x}_0)$ gives a strong warrant for $\mu > .2$ when $n = 25$, but provides very poor evidence for $\mu > .2$ when $n = 400$.

More generally, an α-significant difference with n_1 passes $\mu > \mu_1$ less severely than with n_2 where $n_1 > n_2$. With this simple interpretive tool, all of the variations on "large n criticisms" are immediately scotched (Cohen, 1994, Lindley, 1957, Howson and Urbach, 1993, inter alia). (See Mayo and Spanos, 2006, and in this volume, Chapter 7).

Getting around these criticisms and fallacies is essential to provide an adequate philosophy for error statistics as well as to employ these ideas in philosophy of science.

The place to begin, we think, is with general philosophy of science, as we do in this volume.

8 Error-Statistical Philosophy of Science

Issues of statistical philosophy, as we use that term, concern methodological and epistemological issues surrounding statistical science; they are matters likely to engage philosophers of statistics and statistical practitioners interested in the foundations of their methods. Philosophers of science generally find those issues too specialized or too technical for the philosophical problems as they are usually framed. By and large, this leads philosophers of science to forfeit the insights that statistical science and statistical philosophy could offer for the general problems of evidence and inference they care about. To remedy this, we set out the distinct category of an error-statistical philosophy of science. An error-statistical philosophy of science alludes to the various interrelated ways in which error-statistical methods and their interpretation and rationale are relevant for three main projects in philosophy of science: to characterize scientific inference and inquiry, solve problems about evidence and inference, and appraise methodological rules.

The conception of inference and inquiry would be analogous to the piecemeal manner in which error statisticians relate raw data to data models, to statistical hypotheses, and to substantive claims and questions. Even where the collection, modeling, and analysis of data are not explicitly carried out using formal statistics, the limitations and noise of learning from limited data invariably introduce errors and variability, which suggests that formal statistical ideas are more useful than deductive logical accounts often appealed to by philosophers of science. Were we to move toward an error-statistical philosophy of science, statistical theory and its foundations would become a new formal apparatus for the philosophy of science, supplementing the more familiar tools of deductive logic and probability theory.

The indirect and piecemeal nature of this use of statistical methods is what enables it to serve as a forward-looking account of ampliative (or inductive) inference, not an after-the-fact reconstruction of past episodes and completed experiments. Although a single inquiry involves a network of models, an overall "logic" of experimental inference may be identified: *data* $\mathbf{x}_0$ *indicate the correctness of hypothesis H to the extent that H passes a stringent* *or severe test with* $\mathbf{x}_0$. Whether the criterion for warranted inference is put in terms of severity or reliability or degree of corroboration, problems of induction become experimental problems of how to control and assess the error probabilities needed to satisfy this requirement. Unlike the traditional "logical problem of induction," this experimental variant is solvable.

Methodological rules are regarded as claims about strategies for coping with and learning from errors in furthering the overarching goal of severe testing. Equally important is the ability to use *in*severity to learn what is *not* warranted and to pinpoint fruitful experiments to try next. From this perspective, one would revisit philosophical debates surrounding double counting and novelty, randomized studies, the value of varying the data, and replication. As we will see in the chapters that follow, rather than give all-or-nothing pronouncements on the value of methodological prescriptions, we obtain a more nuanced and context-dependent analysis of when and why they work.

8.1 Informal Severity and Arguing from Error

In the quasi-formal and informal settings of scientific inference, the severe test reasoning corresponds to the basic principle that *if a procedure had very low probability of detecting an error if it is present, then failing to signal the presence of the error is poor evidence for its absence.* Failing to signal an error (in some claim or inference *H*) corresponds to the data being in accord with (or "fitting") some hypothesis *H*. This is a variant of the minimal scientific requirement for evidence noted in part I of this chapter. Although one can get surprising mileage from this negative principle alone, we embrace the positive side of the full severity principle, which has the following informal counterpart:

Arguing from Error: An error or fault is absent when (and only to the extent that) a procedure of inquiry with a high probability of detecting the error if and only if it is present, nevertheless detects no error.

We argue that an error is absent if it fails to be signaled by a highly severe error probe.

The strongest severity arguments do not generally require formal statistics. We can retain the probabilistic definition of severity in the general context that arises in philosophical discussions, so long as we keep in mind that it serves as a brief capsule of the much more vivid context-specific arguments that flesh out the severity criterion when it is clearly satisfied or flagrantly violated.

We can inductively infer the absence of any error that has been well probed and ruled out with severity. It is important to emphasize that an "error" is understood as any mistaken claim or inference about the phenomenon being probed – theoretical or non-theoretical (see exchanges with Chalmers and Musgrave). Doubtless, this seems to be a nonstandard use of "error." We introduce this concept of error because it facilitates the assessment of severity appropriate to the particular local inference – it directs one to consider the particular inferential mistake that would have to be ruled out for the data to afford evidence for *H*. Although "*H* is false" refers to a specific error, it is meant to encompass erroneous claims about underlying causes and mistaken understandings of any testable aspect of a phenomenon of interest. Often the parameter in a statistical model directly parallels the theoretical quantity in a substantive theory or proto-theory.

Degrees of severity might be available, but in informal assessments it suffices to consider qualitative classifications (e.g., highly, reasonably well, or poorly probed). This threshold-type construal of severity is all that will be needed in many of the discussions that follow. In our philosophy of inference, if *H* is not reasonably well probed, then it should be regarded as poorly probed. Even where *H* is known to be true, a test that did a poor job in probing its flaws would fail to supply good evidence for *H*.

Note that we choose to couch all claims about evidence and inference in testing language, although one is free to deviate from this. Our idea is to emphasize the need to have done something to check errors before claiming to have evidence; but the reader must not suppose our idea of inference is limited to the familiar view of tests as starting out with hypotheses, nor that it is irrelevant for cases described as estimation. One may start with data and arrive at well-tested hypotheses, and any case of statistical estimation can be put into testing terms.

Combining Tests in an Inquiry. Although it is convenient to continue to speak of a severe test *T* in the realm of substantive scientific inference (as do several of the contributors), it should be emphasized that reference to "test *T*" may actually, and usually does, combine individual tests and inferences together; likewise, the data may combine results of several tests. To avoid confusion, it may be necessary to distinguish whether we have in mind several tests or a given test – a single data set or all information relevant to a given problem.

Severity, Corroboration, and Belief. Is the degree of severity accorded *H* with $\mathbf{x}_0$ any different from a degree of confirmation or belief? While a hypothesis that passes with high severity may well warrant the belief that it is correct, the entire logic is importantly different from a "logic of belief" or confirmation. For one thing, I may be warranted in strongly believing *H* and yet deny that this particular test and data warrant inferring *H*. For another, the logic of probability does not hold. For example, that *H* is poorly tested does not mean "not *H*" is well tested. There is no objection to substituting "*H* passes severely with $\mathbf{x}_0$ from test *T*" with the simpler form of "data $\mathbf{x}_0$ from test *T* corroborate *H*" (as Popper suggested), so long as it is correctly understood. A logic of severity (or corroboration) could be developed – a futuristic project that would offer a rich agenda of tantalizing philosophical issues.

8.2 Local Tests and Theory Appraisal

We have sketched key features of the error statistical philosophy to set the stage for the exchanges to follow. It will be clear at once that our contributors take issue with some or all of its core elements. True to the error-statistical principle of learning from stringent probes and stress tests, the contributors to this volume serve directly or indirectly to raise points of challenge. Notably, while granting the emphasis on local experimental testing provides "a useful corrective to some of the excesses of the theory-dominated approach" (Chalmers 1999, p. 206), there is also a (healthy) skepticism as to whether the account can make good on some of its promises, at least without compromising on the demands of severe testing. The tendency toward "theory domination" in contemporary philosophy of science stems not just from a passion with high-level physics (we like physics too) but is interestingly linked to the felt shortcomings in philosophical attempts to solve problems of evidence and inference. If we have come up short in justifying inductive inferences in science, many conclude, we must recognize that such inferences depend on accepting or assuming various theories or generalizations and laws. It is only thanks to already accepting a background theory or paradigm *T* that inductive inferences can get off the ground. How then to warrant theory *T*? If the need for an empirical account to warrant *T* appears to take one full circle, *T*'s acceptance may be based on appeals to explanatory, pragmatic, metaphysical, or other criteria. One popular view is that a theory is to be accepted if it is the "best explanation" among existing rivals, for a given account of explanation, of which there are many. The error-statistical account of local testing, some may claim, cannot escape the circle: it will invariably require a separate account of theory appraisal if it is to capture and explain the success of science. This takes us to the question asked in Chapter 1 of this volume: What would an adequate error-statistical account of large-scale theory testing be?

References

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

Berger, J. O. (2003), "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" *Statistical Science*, 18: 1–12.

Cartwright, N. (1983), *How the Laws of Physics Lie*, Oxford University Press, Oxford.

Chalmers, A. (1999), *What is This Thing Called Science? 3rd edition*, University of Queensland Press.

Chang, H. (2004), *Inventing Temperature: Measurement and Scientific Progress*, Oxford University Press, Oxford.

Cohen, J. (1994), "The Earth Is Round ($p < .05$)," *American Psychologist*, 49: 997–1003.

Galison, P. L. (1987), *How Experiments End*, The University of Chicago Press, Chicago.

Hacking, I. (1983), *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*, Cambridge University Press, Cambridge.

Hands, W. D. (2001), *Reflection Without Rules: Economic Methodology and Contemporary Science Theory*, Cambridge University Press, Cambridge.

Howson, C. and Urbach, P. (1993), *Scientific Reasoning: A Bayesian Approach*, 2nd ed., Open Court, Chicago.

Kuhn, T. S. (1962), *The Structure of Scientific Revolutions*, Chicago University Press, Chicago.

Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44:187–92.

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

Mayo, D. G. (2003), "Could Fisher, Jeffreys and Neyman Have Agreed on Testing? Commentary on J. Berger's Fisher Address," *Statistical Science*, 18: 19–24.

Mayo, D. G. and Spanos, A. (2004), "Methodology in Practice: Statistical Misspecification Testing," *Philosophy of Science*, 71: 1007–25.

Mayo, D. G. and Spanos, A. (2006), "Severe testing as a basic concept in a Neyman–Pearson philosophy of induction," *British Journal for the Philosophy of Science*, 57: 323–57.

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

Morrison, M. (2000), *Unifying Scientific Theories: Physical Concepts and Mathematical Structures*, Cambridge University Press, Cambridge.

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

Rosenberg, A. (1992), *Economics – Mathematical Politics or Science of Diminishing Returns?* (Science and Its Conceptual Foundations series) University of Chicago Press, Chicago.

Spanos, A. (2007), "Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach," *Philosophy of Science*, 74: 1046–66.

ONE

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: $\mathbf{x}$ 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: $\mathbf{x}$ 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 **x** may be regarded as strong evidence for, or as highly confirming of, theory *T* so long as the posterior probability of *T* given **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 $\mathrm{P}(T|\mathbf{x})$ is calculated by means of Bayes's theorem:

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

This calculation requires an exhaustive set of alternatives to *T* and prior degree-of-belief assignments to each, and an assessment of the term $\mathrm{P}(\mathbf{x}|\text{not-}T)$, 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 **x** so that $\mathrm{P}(\mathbf{x}|T) = 1$, 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*, $\mathrm{P}(\mathbf{x}|\text{not-}T)$ – then we get a high posterior probability in theory *T*; that is, $\mathrm{P}(T|\mathbf{x}) = 1/(1 + e)$. The central problem is this: What warrants taking data **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 $\mathbf{x}'$ warrants Bayesian disconfirmation of an auxiliary hypothesis A (used to derive prediction $\mathbf{x}$), 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 $\mathbf{x}$ 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 $\mathbf{x}$, if $\mathbf{x}$ fails to probe hypothesis J, then $\mathbf{x}$ fails to severely pass H and J (see Chalmers, 1999). By contrast, Bayesians seem content to show that $\mathbf{x}$ 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.

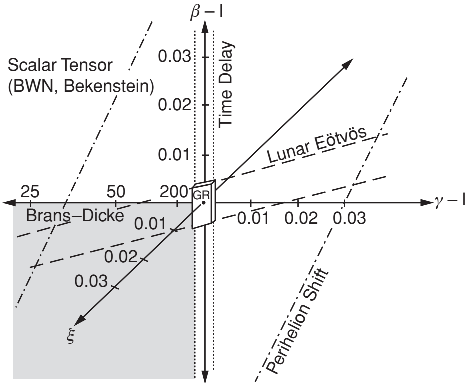

Figure 1.1.

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 Journa*l, 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.