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 $\mathbf{X} = (X_1, \ldots, X_n)$, $f_X(\mathbf{x}; \boldsymbol{\theta})$, 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:

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

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.