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 $\mathbf{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_n)$, $f_X(\mathbf{x}; \boldsymbol{\theta})$, 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 (N(μ,1)) where we want to test the hypotheses:

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

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 $P(d(\mathbf{X}) > 1.96; H_0) = .025$. 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 $\boldsymbol{x}_0$ 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.