Error Statistics and Duhem's Problem*

Gregory R. Wheeler†‡
Departments of Philosophy and Computer Science, University of Rochester

No one has a well developed solution to Duhem's problem, the problem of how experimental evidence *warrants* revision of our theories. Deborah Mayo proposes a solution to Duhem's problem in route to her more ambitious program of providing a philosophical account of inductive inference and experimental knowledge. This paper is a response to Mayo's Error Statistics (*ES*) program, paying particular attention to her response to Duhem's problem. It turns out that Mayo's purported solution to Duhem's problem is very significant to her project, for the epistemic license claimed by *ES* and the philosophical underpinnings to her account of experimental knowledge depend on this solution. By introducing the partition problem, I argue that *ES* fails to solve Duhem's problem and therefore fails to provide an adequate account of experimental knowledge.

1. Introduction. Duhem's problem arises when we have experimental evidence that is contrary to a theory's predictions. Given this situation, we have reason to believe that at least one of the statements of the theory plus auxiliaries is false: the conjunction of theoretical statements, the auxiliaries, and the experimental evidence statement is inconsistent. But do we have adequate grounds for determining *which* statement among the set is to blame? Pierre Duhem argued that such grounds are not found in laboratory notebooks per se but rather in the good sense of their authors. The very nature of experimental evidence renders its bearing on theory essentially opaque. So, according to Duhem, treating experimental evidence as if it determined which statement is to blame is mistaken. Most

*Received May 1999; revised April 2000.

†Send requests for reprints to the author, Department of Philosophy, 534 Lattimore Hall, Rochester NY 14627; email: wheeler@philosophy.rochester.edu.

‡Previous versions of this paper were presented at Cornell University, University of Rochester, M.I.T., and University of Lethbridge. The author wishes to thank Prasanta Bandypadhyay, Earl Conee, Heidi Dankosh, Joe Halpern, Deborah Mayo, and especially Henry Kyburg and an anonymous referee for their comments.

Philosophy of Science, 67 (September 2000) pp. 410–420. 0031-8248/2000/6703-0004$2.00

commentators have found this holistic account less than satisfactory. If we assume that decisions of this kind are rational and that experimental evidence plays a significant role in their being rational, then Duhem's problem is just the problem of determining how experimental evidence warrants assigning error to one statement but not another.

Recently Deborah Mayo (Mayo 1996b) has proposed a novel solution to Duhem's problem, a proposal that explicitly rejects the claim of evidential opacity and thus whose burden it is to show "that there are good grounds for localizing the bearing of evidence" (Mayo 1996b, 102). Mayo's proposal stems from her Error Statistics account of experimental knowledge (*ES*). Her idea is that in many experimental situations Duhem's problem can be resolved because not all alternative hypotheses are at once susceptible to revision. We say "many" situations and not "all" since not every occasion of disconfirming evidence does Duhem's problem make: sometimes the reasonable thing to do is to collect more evidence. Mayo accounts for this easily enough by appealing to a fundamental distinction found in classical statistics between two kinds of cases: (i) those in which there are positive grounds for attributing the error to some statement h, and (ii) those in which there are inadequate grounds for attributing the error to statement h. So what we need to know is what it is to have good grounds for attributing error to a statement.

Mayo adopts Karl Popper's slogan that we learn the most about hypotheses which are severely tested. But rather than propose an update to Popper's falsificationism, Mayo instead grounds her notion of severity in classical statistics. Under *ES,* severity is a property of statistical method that ensures that a test of h is a good one. Curiously, however, severity "attaches to a particular hypothesis passed" (Mayo 1997a, 250) thereby granting the hypothesis epistemic warrant. By identifying " 'having good evidence for h (or just having evidence for h)' and 'having a good test of h," ' Mayo then identifies "whether e counts as good evidence for h . . . [with] whether h has passed a good test with e" (Mayo, 1996b, 179).[1] An hypothesis h then is acceptable just to the extent that it passes a severe test. So of the cases that call for a solution to Duhem's, problem we can expect each to satisfy this severity requirement.

What then constitutes a good (i.e., severe) test? According to Mayo, a severe test T is one such that "there is a very low probability that test procedure T would yield such a passing result, if hypothesis h is false" (Mayo, 1997, 248). So, in type (i) cases we need a good test that determines whether or not an auxiliary statement, A, is to blame for the contrary experimental result, e'. Given e', *ES* says we may consider a statement for blame only when a test is run which measures the probability of accepting

1. I substitute 'h' for Mayo's 'H'.

the statement when it is false (i.e., measures the probability of committing a Type-I error). So, under *ES* a statement (hypothesis) *h* is shown to be in error as a result of *e'* only if the alternative hypothesis *A* has been shown to pass a severe test (Mayo 1996b, 108). Since often there is more than one alternative hypothesis we may generalize *ES*'s severity condition for type (i) cases as follows:

> *Severity Condition*: Hypothesis *h* is shown to be in error as a result of *e'* only if: given the set of auxiliary statements Γ, all $A_n \in \Gamma$ have been shown to pass severe tests.

Thus we have a necessary condition for the *ES* solution to Duhem's problem. In the next two sections I argue that *ES* cannot satisfy this condition. Hence, if it is not the case that all A_n have been shown to pass severe tests, then *h* is not shown to be in error as a result of *e'*. And if *h* is not shown to be in error as a result of *e'*, then there are inadequate grounds for attributing the error to *h*. But any case in which there are inadequate grounds for attributing error to some hypothesis is a case of type (ii) and hence is not a solvable Duhem case. The upshot of our argument is that *ES* renders all cases type (ii) cases since it does not include an adequate account of when enough evidence is enough.

2. The Partition Problem. Our interest in this section is to see why the *ES* solution to Duhem's problem is inadequate. We observed that if *ES* treats all cases as cases in which there are inadequate grounds for assigning error to a statement, then we are left wondering how experimental evidence warrants rejecting one statement instead of another. In other words, if *ES* treats all cases as type (ii) cases, then we are left precisely with Duhem's problem. So, what reasons do we have to think that *ES* treats all cases as type (ii) cases?

To begin, one may suspect that the severity condition invites a kind of third-man argument. We'll call this *the testing regress*. One could add auxiliaries indefinitely to a typical set Γ of given auxiliaries thereby introducing an indefinite series of tests.[2] Indeed, without restrictions on the auxiliary statements in Γ there are, in principle, an infinite number of tests. A recipe for inflating Γ in this manner is to randomly pick declarative sentences out of the language, without replacement, construct an auxiliary stating *it* doesn't affect the test hypothesis and then test whether the fac-

2. Mayo, adopting Suppes' (Suppes 1969) notion of a model, presents *ES* "as a series of conceptual representations or models ranging from the primary scientific hypothesis or questions . . . to the nitty gritty details of the generation and analysis of data" (Mayo 1996b, 128). Notice that this maneuver doesn't resolve the testing regress; nothing in Suppes' sketch of data models precludes there being an infinite series of such models, given the task Mayo assigns for them.

tor(s) denoted by *each* such construction is responsible for an error. But, of course, if there are infinite tests then not all $A_{n\to\infty} \in \Gamma$ could be shown to pass a severe test. So there would be inadequate grounds for blaming h for e' and Duhem's problem would remain.

As formulated, the testing regress objection bears a similarity to what Mayo calls *the alternative hypothesis objection*—an objection she contends only bears against Popper's account of severe testing, not hers. For Popper, h passes a severe test with e only if all so-far-considered hypotheses have been tested and each entails $\neg h$. But this objection doesn't apply to *ES*, since for Mayo a severe test of h must, with high power, probe the ways that h can err, and need not test an alternative h'. Severity, then, is a property that is always assessed within some context or theory. As Mayo observes:

> Satisfying the severity requirement demands that we make our questions appropriately small or local. . . . By using simple local contexts in which the assumptions *may be shown to hold sufficiently*, it is possible to ask one question at a time. (Mayo 1997, 254. Italics added and deleted.)
>
> Notice that these methods correspond to the goal of satisfying the experimental assumptions . . . Then there is an array of extraneous factors assumed to be either irrelevant to the effect of interest or satisfactorily controlled. The correctness of this assumption can, in principle, come up for questioning after-the-trial. (Mayo 1996b, 144).

Mayo's proposal then is this: localize test questions by sorting out what is relevant for testing from what is safe to assume is irrelevant. This will reduce the number of factors to a manageable size where *ES* can do its work, that is where we can severely test single hypotheses. Popper's account fails, then, precisely because it does not accommodate the hierarchical structure of experimental inquiry. Presented with an anomaly, the hypothetico-deductive method leaves a disjunction of negated statements—the negation of each $A_{n\to\infty} \in \Gamma$, plus $\neg h$. *ES* works, we're told, because it imposes a structure on the statements in Γ, sorting them into relevant models simple enough for error statistical methods to work out a solution to Duhemian cases. So, the severity condition in play imposes a structure on Γ:

> *Structured Severity Condition*: Hypothesis h is shown to be in error as a result of e' only if: given a finite set of relevant auxiliaries, $\{R_1,\ldots,R_k\} \subset A_{n\to\infty} \in \Gamma$, all $R_k \in \Gamma$ have been shown to pass severe tests.

Underlying this proposal, however, is the claim that these structures rest on good grounds. In other words, the error probabilities that underpin

localized experiments must themselves be tested or shown to hold, even if only in principle. This last claim is essential for establishing *ES*'s normative-epistemic credentials and is the target of my criticisms. It is essential because if it turns out that this structure cannot be accounted for within *ES*, then the claim that evidence isn't opaque is undermined: rejecting statements becomes more than a matter of evidence and method, at least as those notions are construed and employed within *ES*.

The key then is the structure of Γ. According to Mayo, by demanding that each test be specific, we are forced to sort out what is relevant and, hence, what are likely sources of error. Each test then has a set of extraneous factors that we ignore or control for, and a smaller "relevant" set that we pay close attention to. This latter set of factors is just what our experiment is about; they are the target properties that are measured, examined, and from which we learn. For example, Adams and Laplace's test of the predicted acceleration δ of the moon involved a set of auxiliary statements, including: A_1: tidal friction is not sufficient to affect measured lunar acceleration more than $\delta^{\pm n}$; A_2: instrument X's margin or error is not sufficient to produce measurements of lunar acceleration more than $\delta^{\pm n}$; A_3: seasonal movements of migratory birds do not affect measured lunar acceleration more than $\delta^{\pm n}$, and so on. These three statements are a subset of the set of auxiliaries Γ for Adams and Laplace's test. The factors in this example are four target properties: Tidal friction (of some magnitude λ), instrument accuracy, lunar acceleration, and collective bird force. To avoid the problems which beset Popper, Mayo's proposal is to assume that most of the auxiliaries in Γ are about properties that are irrelevant to the hypothesis under test (like bird force and, in the original experiment, tidal friction), and so may be ignored. This leaves a few auxiliaries that are controlled (like instrument error) and the test hypothesis involving the factors we are interested in. In the face of disagreeing evidence, we may have a hunch that the tidal friction auxiliary is a better hypothesis to reject than the bird-force auxiliary, but the promise of *ES* is the claim that there is an empirical method for justifying this preference; that is, that our decision is grounded by evidence and method alone.

But notice what is required to fulfill this promise. Each test carves up the auxiliaries $A_{n\to\infty} \in \Gamma$ into auxiliaries to test and auxiliaries to ignore. So, associated with a test is a particular partition of Γ, say the *i*th of infinite possible partitions, that fixes which auxiliaries are to be tested and which to be ignored. The ignored auxiliaries make up that test's ceteris paribus condition. Fixing this partition is crucial to determining which of the $\Gamma_{n\to\infty}$ will be tested and, therefore, which auxiliaries are candidates for "bearing the evidence" e'. For *ES*'s promise to hold it needs to test each of these partitions, or be able to at least in principle. But notice that after-trial testing of the partition's placement (i.e., which of the $\Gamma_{n\to\infty}$ is the right one)

is not possible, even in principle, since to try invites the testing regress. That is, if Γ is infinite in breadth, then so too are there infinite possible divisions of auxiliaries into those to test and those to ignore. So, if Γ is infinite then *ES* cannot get started, even in principle.

Since passing a severe test depends on selecting the right auxiliaries to test and the right ones to ignore, the *ES* version of Duhem's problem is simply this: on what grounds are we justified fixing the partition of a test's auxiliaries one way rather than another? For convenience, let's call this *the partition problem.*

3. When Γ Is Finite. It is important to realize that the partition problem does not depend on Γ being an infinite set of statements. Even if we suppose that Γ is finite, the number of modalities generated by even toy experiments is sufficiently large to introduce the partition problem. To see this we'll look closely at an example Mayo borrows from R. A. Fisher.

Suppose there is a woman who claims to distinguish by taste alone whether tea or milk is added first to a mixture of tea and milk. Suppose we are interested in testing this claim. Let h be: Lady can discriminate order by taste, and let the null hypothesis h_0 be: Lady cannot discriminate order by taste. Fisher reasoned that someone who failed to discriminate the order by taste would do no better than chance at determining the correct order of the mixture. Thus, the question at hand (i.e., whether she has the ability) is reduced to considering two hypotheses h and h_0. The binomial chance model is assumed to accurately model the results of her failing to have the ability. So, h is confirmed or "warranted" to the extent that the experimental record of her correct guesses differs significantly from the results of flipping a fair coin.

We infer whether her guesses are significantly different from flips of a fair coin from probability theory. Suppose we prepare 5 teacups for her to sample. This creates 2^5 possibilities, or 32 possible outcomes. The probability of choosing the correct milk-tea order by chance in all 5 cups is just the probability of picking one of the 32 possible sequences, or .03125. This probability is the probability of committing a Type-I error—i.e., the probability of accepting h (rejecting the null h_0) when h is false.[3]

But notice an assumption that we are making to get this far in the example. Mayo writes that "in order for the comparison offered by the statistical link in the experimental model to go through, the assumptions of the experimental model must hold sufficiently in the actual experiment"

3. In a five-cup case, let c be the number of cups classified correctly. The probability of guessing c correctly is calculated by $\binom{5}{c}\left(\frac{1}{2}\right)^{c}\left(\frac{1}{2}\right)^{5-c}$. We reject the null if $c = 5$, since the probability of $c = 5$ is 1/32, if h_0 is true.

(Mayo 1996b, 136). One assumption that must hold is that the subject isn't tipped off by something other than the taste of the samples. Since we are measuring the lady's ability to discriminate by taste, our confidence that we are only exposed to a 1 in 32 chance of her making the right choice all five times and yet not having the ability to do it by taste turns on this assumption holding. So, the very idea of a severe test is predicated on the assumption that the power of our test is quite high. Yet, on what grounds do we know that it is?

As a first precaution, we might wish to randomize the order of the treatments so that we can avoid giving clues to the subject. We might begin by randomizing the order of the milk-tea mixture for each treatment, altering between milk first and tea first. We may even wish to randomize (or standardize) the presentation of the cups too, in case there is an ordering of the cups' masses, or rim thickness that correlates with the mixture order. Notice that what we are doing is controlling possible factors that may reduce our confidence in our probability assignment for Type-I errors. We've controlled for the possibility that the experimenter knowing the order of the mixture influences the subject's performance, and the possibility that a non-random order of the teacups may give a clue of the order to the subject, respectively.

What we are articulating is the class of auxiliaries to test and those to let pass. Specifically, we are describing the class of controlled factors that have corresponding auxiliaries in some ith partition of Γ: (Γ_i). Hence, implicitly we are fixing a partition. In designing our experiment we make judgments about what to include in the class of tested auxiliaries and what to push into our ceteris paribus condition. For instance, milk-tea mixture order and stirring are to be controlled for, randomizing the cups before presenting them to the experimenter might be a borderline case, and the make of the china most likely isn't considered a serious candidate at all. But how do we make these judgments about what to test and what to regard as extraneous? We might be tempted to cite our "good sense" or previous experience, if not for remembering Duhem's own solution to similar puzzles in the philosophy of experimental physics a century ago. That is, in so far as our previous experience can be codified into an empirical theory, we may ask on what grounds we accept it. The upshot is that even if we treat Γ as finite for toy experiments like tea tasting we can easily inflate it to a size that demands a partition. Yet once Γ is partitioned, we then create a set of n number Γ's and once again are faced with the problem of determining which partition to settle on.

4. A Partial Solution? Mayo's solution to Duhem's problem fails because it depends on a given structure of the set of auxiliary statements that itself can't be justified by *ES* methods. Under *ES* we haven't good grounds to

prefer one partition over another, and so haven't good grounds for considering contrary evidence to count against one hypothesis over another.

But even though we don't have a full solution to Duhem's problem, we might wonder whether *ES* provides a "partial" solution to the problem. Suppose we simply accept a certain partition as a matter of convention. *ES* might provide us with a means to test a limited number of viable alternatives against the current partition thereby giving us some empirical evidence that warrants selecting one over the other. While not solving Duhem's problem, this conventionalist approach might account for inductive practices within some agreed upon domain of inquiry; we might find solace knowing whether we can empirically compare our current theory to at least some others and have an empirical basis for evaluating the merits of alternatives vis-a-vis the current partition.[4]

Let's suppose we are given a particular partition of auxiliaries, Γ_i. What does accepting Γ_i tell us? Γ_i is a set of statements, after all, yet our interest lies in the factors those statements denote. Do we have the resources to compare Γ_i to Γ_j? It turns out that if we're to consider a revision, even when given a partition of Γ, we still must construct a test akin to testing all n-partitions of Γ. Roughly speaking, to compare Γ_i and Γ_j we're forced to consider a test that eliminates any advantage accepting a particular structure of Γ gives. Simply accepting Γ_i doesn't provide us with enough information to effectively use *ES* to evaluate an alternative partition.

To see the problem let's return to the tea tasting example. To compare the ith and jth partition of Γ we first need to see what we know from starting with Γ_i. Suppose that the ith partition includes in the class of untested auxiliaries A_4: Using city tap water is not correlated with the subject recording correct responses better than chance. In considering an alternative, Γ_j, how do we evaluate Γ_j if $\bar{A}_4 \in \Gamma_j$?

Since the city-water factor is under consideration we might wish to subject A_4 to test. Suppose we do and there are inadequate grounds for rejecting A_4. May we then infer that the city-water *factor* is not statistically relevant to the lady's ability to respond correctly better than chance? No, not as *ES* stands. The reason is that to do so would be to conclude that the city-water factor is *independent* of all other factors and, hence, not a constituent in a multi-factor effect. By accepting A_4 we accept that the city-water factor alone isn't correlated with the subject performing better than chance, not that it has no effect at all. But suppose circumstances are such that it does affect the subject's performance, but only if the experiment uses unpasteurized milk and the tea kettle is brought to a rolling boil before pouring. In such circumstances it isn't the case that the city-water

4. Larry Laudan sketches a similar approach for *ES* in (Laudan 1997). See also (Kyburg 1990).

factor is irrelevant to the subject's performance. Yet, without keen theoretical knowledge that extends beyond merely accepting Γ_i, *ES* methods alone couldn't help us with detecting this multi-factor effect. So long as we follow Mayo's recommendation of equating "having good evidence" with "having a good test," we are forced to test each auxiliary in Γ that pairs the city-water factor with all permutations of other factors. But this is unreasonable, for among the set of auxiliaries to test is a superstatistic that treats all factors as controlled and whose set of untested auxiliaries is the empty set. Leaving aside the computational expense and incomprehensible size of such a statement, such a test renders *ES* epistemologically vacuous: for, again, it strips *ES* of the structure it needs to target evidence.

One might suggest that we group the auxiliaries into "stable factor sets" in an attempt to represent previous experience. Suppose we determine that our *n*-factor auxiliary is not statistically relevant, but we're curious about an $n+1$ factor chain. Couldn't we cut down on the number of total statements by grouping together auxiliaries denoting factors that have proven steady losers? No; not as *ES* stands. Even if the auxiliary testing the set of factors $\{f_1, f_2, \ldots, f_n\}$ is not statistically relevant to the subject reporting correct guesses, we are without "good grounds" to infer that $\{f_1, f_2, \ldots, f_n, f_{n+1}\}$ is also not statistically relevant without testing the auxiliary denoting *that* set. Notice, too, that "good grounds" isn't monotonic either. In other words, even if $\{f_1, f_2, \ldots, f_n,\}$ is found not statistically relevant we couldn't infer $\{f_2, f_3, \ldots, f_n\}$ isn't statistically relevant too without testing that factor chain. So long as good grounds is akin to a good test, we haven't a viable way to navigate the factor space and actually learn from error. The upshot is that accepting Γ_i doesn't amount to the kind of knowledge we need to guide our use of *ES* methods.

What is important to see is that we are forced not only to accept a ceteris paribus condition to test anything under ES, but must also rely on a robust knowledge base to direct those tests. The very idea of a severe test is predicated on having a very rich body of empirical knowledge that itself is warranted by means other than those provided by *ES*. To propose *ES* as account of experimental knowledge, then, is to have things turned around.

To the extent that *ES* solves Duhem's problem, even partially, it does so by relying heavily on a rich body of knowledge that can't be accounted for by *ES* methods. It is the great experimentalist who knows how to use her limited resources and theoretical knowledge to probe the factor space to maximize her chances of learning about the system under study. *ES* simply fails to give a philosophical account of how this is done.

5. The Problem for *ES*. What is philosophically attractive about *ES* is its epistemic promise. Mayo's account is proposed as an account of how

experimental knowledge claims are *warranted.* The crux of *ES* is Mayo's notion of a severe test. But the *ES* notion of a severe test fails to do its own epistemic work required for solving Duhem's problem. We direct the tools of *ES* in precisely the manner that Duhem's problem concerns. In the end, *ES* describes such inferences and fails to explain the grounds for our preferences. That we in fact reduce the size of the factor space and in fact seem to target evidence is not in dispute: what we want, and *ES* leaves wanting, is an account of how this is done.

In closing, note that this failure presents a pressing problem for *ES* in general. For Mayo appeals to a version of C. S. Peirce's thesis that inductive methods are "self-correcting" in order to justify *ES* methods.

> By developing my view of Pierce's error-correcting justification of induction I will . . . be developing the justification I need for error statistical methods in science. The justification for these methods lies in their ability to control error probabilities, hence sustain learning from error, hence provide for the growth of experimental knowledge. (Mayo 1996b, 413)

Yet, necessary for Mayo's version of Pierce's thesis is that "the [test] method should be able to detect its own errors in the sense of checking its own assumptions . . . and it should be able to correct violations or 'subtract them out' in the analysis" (Mayo 1996b, 421). But this, of course, is precisely what *ES* cannot do. The assumptions that distinguish good tests from bad are precisely those that cannot be checked by severe tests. In so far as an *ES* test identifies a statement to reject, it does so because of the wits of its designer, not the features of her test method.

REFERENCES

Ariew, Roger (1984), "The Duhem Thesis", *British Journal for the Philosophy of Science* 35: 313–25.

Ariew, Roger and P. Barker (eds.), (1996), *Pierre Duhem: Essays in the History and Philosophy of Science.* Indianapolis: Hackett Press.

Duhem, Pierre ([1906] 1954), *The Aim and Structure of Physical Theory.* Reprint. Translated by P. P. Wiener. Originally published as *La Théorie Physique: Son Objet, et sa Structure* (Paris: Marcel Rivière & Cie). Princeton: Princeton University Press.

Fisher, R. A. (1956), *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd.

Grünbaum, Adolf (1960), "The Duhemian Argument", *Philosophy of Science* 27: 75–87.

Howson, Colin (1997), "A Logic of Induction", *Philosophy of Science* 64: 268–290.

Kyburg, Henry E., Jr. (1983), *Epistemology and Inference.* Minneapolis: University of Minnesota Press.

Kyburg, Henry E., Jr. (1990), "Theories as Mere Conventions", in Wade Savage, (ed.), *Scientific Theories: Minnesota Studies in the Philosophy of Science, vol. 14.* Minneapolis: University of Minnesota Press, 158–74.

Kyburg, Henry E., Jr. (1997), "Combinatory Semantics", *Computational Intelligence* 13: 215–257.

Lakatos, Imré (1978), "The Methodology of Scientific Research Programmes", in John Wor-

rall and G. Currie, (eds.), *Philosophical Papers: Vol. 1.* Cambridge: Cambridge University Press.
Lakatos, Imre and A. Musgrave, (eds.), (1970), *Criticism and the Growth of Knowledge.* Cambridge: Cambridge University Press.
Laudan, Larry (1997), "How about Bust? Factoring Explanatory Power Back into Theory Evaluation", *Philosophy of Science* 64: 306–316.
Mayo, Deborah G. (1996a), "Ducks, Rabbits, and Normal Science: Recasting the Kuhn's-eye view of Popper's Demarcation of Science", *The British Journal for the Philosophy of Science* 47: 271–90.
Mayo, Deborah G. (1996b), *Error Statistics and the Growth of Experimental Knowledge.* Chicago: University of Chicago Press.
Mayo, Deborah G. (1997a), "Severe Tests, Arguing from Error, and Methodological Underdetermination", *Philosophical Studies* 86: 243–66.
Mayo, Deborah G. (1997b), "Error Statistics and Learning From Error: Making a Virtue of Necessity", *Philosophy of Science* 64 (Proceedings) : S195-S212.
Mayo, Deborah G. (1997c), "Duhem's Problem, The Bayesian Way, and Error Statistics, or 'What's Belief Got to Do With It'?", *Philosophy of Science* 64: 222–244.
Mayo, Deborah G. (1997d), "Response to Howson and Laudan", *Philosophy of Science* 64: 323–333.
Quine, Willard V. O. (1969), *Ontological Relativity and Other Essays.* New York: Columbia University Press.
Suppes, Patrick (1969), "Models of Data", in Patrick Suppes, (ed.), *Studies in the Methodology and Foundations of Science.* Dordrecht: D. Reidel, 24–35.
Worrall, John (1993), "Falsification, Rationality, and the Duhem Problem", in John Earman, A. Janis, G. Massey, and N. Rescher, (eds.), *Philosophical Problems of the Internal and External Worlds: Essays on the Philosophy of Adolf Grünbaum.* Pittsburgh: University of Pittsburgh Press.