Duhem's Problem, the Bayesian Way, and Error Statistics, or "What's Belief Got to Do with It?"*

Deborah G. Mayo†‡
Department of Philosophy, Virginia Polytechnic Institute and State University

I argue that the Bayesian Way of reconstructing Duhem's problem fails to advance a solution to the problem of which of a group of hypotheses *ought* to be rejected or "blamed" when experiment disagrees with prediction. But scientists do regularly tackle and often enough solve Duhemian problems. When they do, they employ a logic and methodology which may be called *error statistics.* I discuss the key properties of this approach which enable it to *split off* the task of testing auxiliary hypotheses from that of appraising a primary hypothesis. By discriminating patterns of error, this approach can at least block, if not also severely test, attempted explanations of an anomaly. I illustrate how this approach directs progress with Duhemian problems and explains how scientists actually grapple with them.

1. Introduction. Pierre Duhem states his problem as follows:

> The physicist can never submit an isolated hypothesis to the control of experiment, but only a whole group of hypotheses. When experiment is in disagreement with his predictions, it teaches him that one at least of the hypotheses that constitute this group is wrong and must be modified. But experiment does not show him the one that must be changed. (Duhem 1954, 185)

*Received October 1995; revised May 1997.

†Send reprint requests to the author, Department of Philosophy, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0126.

‡I thank Philip Kitcher and an anomymous referee for extremely useful queries on a very early version of this paper. Variations on this paper have been presented at the University of Pittsburgh, Virginia Tech, The London School of Economics and Political Science, the University of Rochester, and the University of Minnesota. I benefited greatly from the questions and criticisms of all of these audiences.

Philosophy of Science, 64 (June 1997) pp. 222–244. 0031-8248/97/6402-0002$2.00

This raises a well-known problem for a simple Hypothetico-deductive model of hypothesis testing. On this simple HD model if an experimental result disagrees with the expected consequence of a hypothesis H, then H is disconfirmed. However, Duhem points out, the deduction of the expected consequence from hypothesis H generally requires auxiliary assumptions and background conditions and these may themselves be open to question. The simple HD method, in and of itself, does not tell us whether to allocate the blame to H or to part of the auxiliary assumptions.

The task that Duhem's problem poses for philosophers of science is to provide a way to determine which of the hypotheses used to derive a predicted consequence should be rejected or disconfirmed when experiment disagrees with that prediction. If the simple HD method will not succeed, as clearly it will not, the question arises as to whether other models of testing or confirmation will. In actual scientific episodes sometimes H is taken to blame and other times H is retained while auxiliary assumptions are said to be responsible for the anomalous result. And an adequate model of testing should account for this. Several defenders of the subjective Bayesian model of confirmation have argued that even if scientists are not conscious or unconscious Bayesians, reconstructing scientific inference in Bayesian terms is of value in solving key problems in philosophy of science. The problem for which the Bayesian Way is most often touted as scoring an impressive success is the Duhem problem.

The Bayesian strategy subjective Bayesians appeal to in their solution to Duhem is that of Jon Dorling (1979), and since Dorling's work is credited as the exemplar for the Bayesian solution to Duhem, I will take it as my example too. Dorling's aim, as he puts it,

> is to point out that the Bayesian personalist approach to scientific inference provides a . . . solution to this [Duhem] puzzle by telling us exactly when [disregarding unsuccessful predictions] can be reconstructed as rational and when it has to be deemed irrational. Rationality here, for the Bayesian, simply means conformity with [Bayes'] theorem. (Dorling 1979, 177)

Several Bayesians feel he has largely succeeded. Michael Redhead (1980) endorses and elaborates on Dorling's analysis. Colin Howson and Peter Urbach, currently among the strongest defenders of the subjective Bayesian Way, believe that Dorling's examples show, not only how Bayes' theorem solves the Duhem problem, but also "that the Bayesian model is essentially correct." They declare, further, that no non-Bayesian methods have the resources to even begin to solve Duhem's problem (p. 101). John Earman, that most refreshingly critical

of Bayesians, calls Dorling's proposed Bayesian solution a "highly qualified success for Bayesianism," but even with this hedge he allows that "the apparatus provides for an illuminating representation of the Quine and Duhem problem" (Earman 1992, 85).

In this paper I shall be arguing that the Bayesian Way out of Duhem's problem is really no way out at all, nor do I think it offers a more illuminating statement of the problem than that of Duhem himself. It fails to solve the problem because a Bayesian reconstruction does not show which hypothesis it is warranted to credit or blame. It does not illuminate the problem because it does not accord with how Duhem's problem is or should be grappled with in science. But my aim is not primarily negative. Scientists do regularly tackle and often enough solve Duhemian problems, and my real aim is to set the stage for the kinds of methods and reasoning they use to adjudicate disagreements about what is at fault.

2. The Subjective Bayesian Model. In a nutshell, the subjective Bayesian model of confirmation says that evidence *e* confirms hypothesis H to the extent that an agent's degree of belief in H is higher given evidence *e* than what it was or would be without evidence *e*. Probability measures subjective degree of belief. The agent's degree of belief in H after evidence *e* is called the posterior probability assignment. The degrees of belief an agent has in a hypothesis H and its alternatives without evidence *e* are the prior degree of belief assignments. Inductive inference from evidence is a matter of updating one's degree of belief to yield a posterior degree of belief so as to cohere with Bayes' theorem. The simplest statement of Bayes' theorem is this:

$$\mathrm{P}(\mathrm{H} \mid e) = \frac{\mathrm{P}(e \mid \mathrm{H})\,\mathrm{P}(\mathrm{H})}{\mathrm{P}(e \mid \mathrm{H})\,\mathrm{P}(\mathrm{H}) + \mathrm{P}(e \mid \text{not-H})\,\mathrm{P}(\text{not-H})}$$

The key quantities are the prior probabilities in H and in not-H, the likelihood, P(*e* | H) and the Bayesian "catchall factor," P(*e* | not-H). If you have these then you can compute Bayes' theorem to get the posterior probability of H.

Bayes' theorem just follows from the probability calculus and is unquestioned by critics. What is questioned by critics is the relevance of a certain use of this theorem; namely for scientific inference. Their question for the subjective Bayesian is whether scientists have prior degrees of belief in the hypotheses they investigate and whether, even if they did, it is desirable to have them figure centrally in learning from data in science. Would not such personal opinions be highly unstable, vary-

ing not just from person to person, but from moment to moment? That they would is expected and accepted by subjectivists.

Another serious problem is the Bayesian catchall factor P(e | not-H). Not-H, the catchall hypothesis, refers to a disjunction of hypotheses other than H, including those not yet thought of. The smaller the assignment to the Bayesian catchall factor, the higher the confirmation to H. For very small assignments to the Bayesian catchall factor the denominator may hardly be greater than the numerator.

3. Dorling's Homework Problem. When Bayesians say they can solve Duhem's problem, what they mean is this: Give me a case in which an anomaly is taken to disconfirm a hypothesis H out of a group of hypotheses used to derive the prediction, and I will show you how certain subjective probability assignments can justify doing so. The "justification" is that H gets a low (or lower) posteriori probability than the other hypotheses. Solving Duhem comes down to a homework assignment—not to say a necessarily easy one—of determining how various assumptions and priors allow the scientific inference reached to be in accord with that reached via Bayes' theorem. Let us look at Dorling's homework problem informally.

Dorling considers a situation where despite the fact that an anomalous result e' occurs, the blame is placed on an auxiliary hypothesis A while the credibility placed on hypothesis H is barely diminished. In Dorling's simplified problem, only one auxiliary hypothesis A is considered. The hypothesis H he considers is "the relevant part of solidly established Newtonian theory which Adams and Laplace used" (Dorling 1979, 178) to compute e, the predicted secular acceleration of the moon, which conflicted with the observed result e'. The auxiliary hypothesis, A, is the following:

> A: the effects of tidal friction are *not* of a sufficient order of magnitude to affect appreciably the lunar acceleration.

Dorling's homework problem is to provide probability assignments so that, in accordance with the episode, an agent's credibility in hypothesis H is little diminished by the anomaly e', while the credibility in auxiliary A is greatly diminished. Although the asymmetric effect of the evidence may seem surprising, a closer look at the Bayesian assignments shows it not to be surprising at all.

We can sidestep the numerical gymnastics to get a feel for one type of context where the agent faults auxiliary A. (The Appendix contains the corresponding numerical assignments.)

Hypothesis H and auxiliary A entail e, but e' is observed. When might e' justifiably be taken to blame A far more than H? Here's one

scenario which I will sketch in terms that I intend to be neutral between accounts of inference. Suppose the following 3 conditions obtain:

(1) there is a great deal of evidence in favor of a theory or hypothesis H, whereas
(2) there is little evidence for the truth of auxiliary A, say hardly more evidence for its truth than for its falsity, and
(3) unless A is false, there is no other plausible way to explain e'.

This would seem to describe, in neutral terms, a situation where e' indicates (or is best explained by) A being in error.

A Bayesian rendering may be effected by inserting "agent x believes that" prior to assertions (1), (2), and (3). We then have a description of a circumstance where the agent believes or decides that A is discredited by e'. Nothing is said about whether the assignments are warranted, or, more importantly, how a scientist should go about determining where the error really lies.

Consider the numbers corresponding to Dorling's Bayesian reconstruction. To begin with, the scientist's degree of belief is such that a high degree of belief is accorded to H initially (e.g., P(H) = 0.9), in any case, H is substantially more probable than A, which is considered only slightly more probable than not (e.g., P(A) = 0.6). That is, the assumed prior probabilities are:

(i) P(H) = 0.9 P(A) = 0.6.

To get the other key assignments, imagine our agent contemplating two different possibilities. First, the agent contemplates the possibility that auxiliary hypothesis A is true.

3.1. The agent contemplates auxiliary A being true. Clearly, H could not also be true (since together they counterpredict e'). But might not some rival to H explain e'? Here is where the key assumption enters. It corresponds to assigning a low value to the Bayesian catchall factor. The agent believes there to be no plausible rival that predicts e'. That is to say, the agent sees no rival which, in his or her opinion, has any plausibility, that would make anomaly e' expected. In subjective probability terms, this becomes:

(ii) The probability of e', given A and not-H, is very small. Let this very small value be ε.

Since the anomaly e' has been observed, it might seem that the agent would assign it a subjective probability of 1. Doing so would have

serious ramifications (i.e., this is the "old-evidence problem"[1]). To avoid assigning degree of belief 1 to e' Bayesian agents need to imagine how strongly they *would have believed* in the occurrence of anomaly e' *before* it was observed—no mean feat. We are to imagine that the scientist considers how strongly he or she would expect e' before knowing of e''s occurrence. Putting aside for now the difficulties in assigning such probabilities, the Bayesian assumes the agent can and does make the key assumption that, on his or her view, the observed anomaly e' is extremely improbable were auxiliary hypothesis A true. It is assumed, that is, that the agent gives a very low assignment to the Bayesian catchall factor.

Now consider the agent's beliefs assuming auxiliary A is false.

3.2. The agent contemplates auxiliary A being false. In contrast, were auxiliary A to be false, the agent finds e' to be much more likely than if A were true. In fact, Dorling imagines that scientists assign a probability to e', given not-A, that is 50 times as high as that in (ii), whether or not H is true. That is, $P(e'|\text{ not-A}) = 50\varepsilon$. We have,

(iii) (a) The probability of e', given H and not-A, is 50ε.
(b) The probability of e', given not-H and not-A, is 50ε.

Of course, (a) and (b) need not be exactly equal, but what they must together yield is a probability of e' given not-A many times the probability assignment to e' given A.

A further assumption, it should be noted, is that H and A are probabilistically independent.

Together, (i)–(iii) describe a situation where the outcome e' is believed to be far more likely if A does *not* hold than if it does. Indeed, the posterior of A becomes very low, dropping from 0.6 to 0.003—now that the anomaly is known. The degree of belief in not-A becomes practically 1! In contrast, the posterior probability of H remains rather high, slipping a little from 0.9 to 0.897.

This gives one algorithm—Dorling's—for how evidence can yield a Bayesian disconfirmation of auxiliary A, even though A was deemed more plausible than not at the start. Non-quantitatively put, the algorithm for solving the homework problem is this: Start with a suitably high degree of belief in H as compared with A, believe no plausible rival to H exists that would make the anomalous result expected, and hold that the falsity of A renders e' many times more probable than does any plausible rival to H.

In addition to accounting for specific episodes, the Bayesian Way can

1. See, for example, Glymour 1980.

be used to derive a set of general statements of the probabilistic relationships that would have to hold for one or another parceling out of the blame. These equations are neat, and the algorithms they offer for solving such homework problems are interesting. What they do not provide, however, is a solution to Duhem's problem. Duhem's problem, as Howson and Urbach themselves say, is to determine "which of the several distinct theories involved in deriving a false prediction should be regarded as the false element" (1989, 94). The possibility of a degree of belief reconstruction does not help to pinpoint which element ought to be regarded as the false one. After all, Dorling's homework problem can be done in *reverse.* Scientists who assign the above degrees of belief, but with A substituted for H, reach the opposite conclusion about H and A. In the reverse case one blames H rather than A.[2]

Bayesians may retort that the probabilities stipulated in their reconstruction are plausible descriptions of the beliefs actually held at the time; and others are not. That may well be. For my own part, I have no idea as to how to assign the odds Dorling asks us to: namely the odds that a typical scientist "would have been willing to place [on] a bet on the correct quantitative value of the effect [e'], in advance even of its qualitative discovery" assuming Newton's theory is false. (Dorling 1979, 182). (Something like this, recall, is the contortion required to get around assigning e' a probability of 1 once it is known.)

Even if one can imagine what its value would be, the question remains, Why is it relevant to the scientist's reasoning once the effect e' is known? And isn't that where the scientist is in grappling with Duhem?

Nor is it easy to justify the prior probability assignments needed to solve the homework problem, in particular, that hypothesis H is given a prior probability of 0.9. The "tempered personalism" of Abner Shimony (e.g., 1970) advises that fairly low prior probabilities be assigned to hypotheses being considered, so as to leave a fairly high probability for their denial—for the "catchall" of other hypotheses not yet considered. The Dorling assignment leaves only 0.1 for the catchall hypothesis.

I do not see how the Bayesian reconstruction illuminates (more than it distorts) the problem for the simple reason that scientists do not succeed in justifying a claim that an anomaly is due not to H but to an auxiliary hypothesis by describing the degrees of belief that would allow them to do this. On the contrary, scientists are one in blocking an attempted explanation of an anomaly until and unless it is provided with positive evidence in its own right. Scientists, in short, are required to go out and muster evidence for their belief that auxiliaries are responsible. And what they would need to show is that this evidence

2. Similar criticisms of the Bayesian solution to Duhem occur in Worrall 1993.

succeeds in circumventing the many ways of erroneously attributing blame. Both in arriving at this evidence and in scrutinizing these errors it is to non-Bayesian methods and reasoning that they turn.

They employ a logic and a methodology that is of a sort that I call *error statistics.* The error statistics approach is based on the non-Bayesian statistical methods used on a daily basis by scientists who employ statistics in their work. In the error statistics approach, the role of probability is not to assign degrees of belief or confirmation to hypotheses but to characterize the reliability or *severity* of experimental test procedures. Whereas the Bayesian confirmation for A′ rested on the comparatively higher prior probability assignment to hypothesis H, determining if A′ passes a *severe test* need have nothing to do with an appraisal of H. In the logic of error statistical testing, the task of finding out whether auxiliary hypotheses are satisfied is split off from that of appraising the primary hypothesis H. A scientist may believe that some auxiliary hypothesis, rather than H, is to be blamed for an anomalous result, but to *warrant* that claim requires it to have passed a reliable test. High prior degrees of belief in H have nothing to do with it.

4. The Error-Statistical Approach to Duhem's Problem. In referring to error statistics I am including the familiar techniques of statistical analysis we read about every day in polls and studies (statistical significance tests and confidence interval estimates). These familiar techniques come from the methodologies of Neyman-Pearson as well as Fisherian statistics, although I adapt them in ways that go beyond what is strictly found in statistics texts. In particular, a statistics text will not say how to use these methods for solving Duhem, but that is the story that I am supplying. Nevertheless, my use of these methods is not really new; it reflects their actual uses in science. But to free them up from the confines of the particular philosophies of statistics often associated with them, and to allow them to be used in more flexible ways, I give them this new name of (standard) *error statistics.*[3] It seems an appropriate name since what is central to this approach is the reliance upon error probabilities of procedures.

I cannot hope to fully lay out the error statistical way of handling Duhem's problem here, but will limit myself to identifying its key differences from the Bayesian approach exemplified in Dorling's treatment, and to why I think the account enjoys the right kinds of prin-

3. For a discussion of how this approach contrasts with the behavioristic philosophy most often associated with these methods, see Mayo 1985 and 1996.

ciples and methods to justify the distinctions we would like to make between plausible and implausible assignments of blame.

There are two key features of the error statistical approach that are of central relevance to grappling with Duhem's problem: First that it is designed to split off the task of testing auxiliaries from that of testing a primary hypothesis, and second that it is designed to distinguish between hypotheses that equally well fit the evidence by considering the *error probabilities* of their respective tests. I will sketch the relevance of these features for our problem.

4.1. A Piecemeal Approach. The first feature we can identify by saying that it is a piecemeal approach. It corresponds to a view of hypothesis testing that sees data and hypotheses as related, not directly, but by a series of linkages—from the experimental design to the data analysis and only then to some primary hypothesis or question. Different tasks relate to the different models in a given inquiry: the *primary scientific model, experimental testing models,* and *data models.* Each is split off and addressed piecemeal. Two main tasks are explicitly directed to the two key parts of Duhemian worries: the first task is to determine if the data itself are reliable, to determine if there is a real effect (a real anomaly) that needs explaining; the second is to determine if the assumptions of an experiment are met sufficiently, that is, to the problem of checking if alternative auxiliary factors are intervening or if the experiment is adequately controlled.[4]

Duhem is fond of analogizing the physicist and the doctor, in contrast to the watchmaker:

> Physics is not a machine that lets itself be taken apart. We cannot test each piece in isolation. . . . Presented with a watch that does not work, the watchmaker takes apart all the little wheels and examines them one by one until the one that is bent or broken is found. Presented with a sick person, the doctor cannot perform a dissection to establish a diagnosis. The doctor must decide the seat of the illness only by inspecting the effects produced on the whole body. The physicist charged with reforming a defective theory resembles the doctor, not the watchmaker. (Duhem 1996, 85)

We can accept the analogy between the physicist and the doctor but reach a different conclusion from Duhem. True, doctors cannot (generally) perform dissections to establish a diagnosis but, fortunately, they do not have to. Although a great many illnesses could explain a

4. A full discussion of the error statistical framework, the series of models, and so on, occurs in Mayo 1996.

given set of symptoms, this does not prevent the doctor from running a given test (e.g., an MRI scan, a strep test) to determine the presence or absence of a specific condition (e.g., a brain tumor, strep throat)—and without having to list, much less assess the probability of, all of the other possible explanations.

The series of models piecemeal framework is an ideal one for grappling with Duhem's problem. In fact, the central reason for separating out the models relating data and hypotheses is to achieve the aim of correctly apportioning blame (as well as praise). Before experimental results can speak for or against a hypothesis under test, it is necessary to check and estimate the extent of any errors along the way—regarding the data and the auxiliaries. This calls for methods to discern if the experiment was well run, to distinguish real effects from artifacts, estimate backgrounds, and "subtract out" influences of factors other than some intended one. More than striving to check if auxiliaries and assumptions hold, it gives us tools to *distinguish* the effects of given factors. The methods and models from standard error statistics, as well as the logic associated with error statistics, are regularly used to carry out and give structure to these tasks.

Two contrasts with the Bayesian Way may be noted:

(i) Getting beyond a single probability pie: If inference is by way of Bayes' theorem, then pinpointing one hypothesis to blame *does* depend (by the mathematics of the theorem) on the probabilistic assignment given to the alternative hypotheses. One has one probability pie, as it were, and the size of the piece an auxiliary hypothesis A gets depends upon how much the alternatives get. In contrast, a hypothesis about an auxiliary factor, much like the doctor's hypothesis about a given disease, may pass a highly severe test quite apart from an assessment of the primary hypothesis H.

(ii) Getting beyond a white-glove analysis: The Bayesian analysis begins with given data or a given anomaly e':

> The Bayesian theory of support is a theory of how the acceptance as true of some evidential statement affects your belief in some hypothesis. How you came to accept the truth of the evidence, and whether you are correct . . ., are matters which, from the point of view of the theory, are simply irrelevant. (Howson and Urbach 1989, 272)

If anomalies are approached by way of such a "white glove" logical analysis, it is little wonder that they tell us only that there is an error somewhere and that they are silent about its source. By recognizing that the anomaly itself is a highly modelled entity levels away from the

raw experimental data, the error statistician can exploit the nitty gritty details of an experimental context to test different auxiliaries. By becoming shrewd inquisitors of errors, anomalies may be made to speak volumes. This leads to the second key feature of the error statistics approach to Duhem's problem.

4.2. The Fundamental Use of Error Probabilities of Tests. Error probabilities are not assignments of probabilities to hypotheses. No such probabilities are desired or needed in this approach.[5] Probability enters instead as a way of characterizing the experimental or testing process itself, to express how reliably it discriminates between alternative hypotheses and how well it facilitates the detection of errors. Examples of error probabilities are significance levels or p-values, confidence levels, and standard errors of estimates. Distinguishing tests by their associated error properties offers a basis for the discrimination we are after: between warranted and unwarranted assignments of blame.

4.2.1. Error Statistics and Severe Tests. Rather than assign degrees of probability or support to hypotheses, the error statistical approach stipulates when an accordance between evidence and a hypothesis H counts as a good test of or good evidence for hypothesis H. It counts as good evidence for H just to the extent that H passes a *severe test.* Probability enters as a way to assess the severity of the test. I begin with a sketch of the basic logic of error statistical testing.

Minimally, for H to pass a test with evidence *e, e* should agree with or *fit* what is expected or predicted according to H. Some require that H entails *e,* yielding $P(e \mid H) = 1$. A more useful notion of "fit" would require that *e* be within some specified distance from H. Something more, however, is required for the test to be severe. Suppose evidence *e* is found to fit hypothesis H adequately so that H passes a test T. Then H has passed a severe test only if, *in addition,* there is a very high probability that test procedure T would *not* yield such a passing result, if hypothesis H is false. That is,

> H's passing test T (with result *e*) is a *severe test* of H just to the extent that there is a very low probability that test procedure T would yield such a passing result, if hypothesis H is false.

5. Exceptions would be those cases where the truth of a hypothesis can be regarded as the outcome of a random experiment. Except for such cases, the only probabilities that could be assigned to hypotheses are the trivial ones 0 and 1, according to whether it is true 0% of the time or 100% of the time.

If hypothesis H is false, then, with high probability, the experimental result would have been more discordant from H than *e* is.

Arguing from passing a severe test corresponds to an informal pattern of argument that might be called an *argument from error* or *learning from error.* The overarching structure of the argument is guided by the following thesis:

> It is learned that an error is absent when (and only to the extent that) a procedure of inquiry (which may include several tests taken together) that has a very high probability of detecting an error if (and only if) it existed, nevertheless detects no error.

Its failing to detect the error means it produces a result (or set of results) that is in accordance with the absence of the error. Let hypothesis H be an assertion as to the absence of the error in question. That H passes a severe test corresponds to affirming that the error in question fails to be detected by a highly reliable error probe. Such a procedure of inquiry may be called a *reliable or highly severe error probe.* It is this informal argument from error that must take the lead in applying the more formal notion of severity.

Let us see how this logic is exemplified in a standard error statistical tool.

4.2.2. The Statistical Significance Test. To use a familiar type of example, in comparing two groups, say a "treated" and a "control" group, a statistically significant excess in the rate of lung cancer may be observed—perhaps the significance level (or p-value) is .01. This report might be taken to reject the *null-hypothesis* H_0 that there is no increased risk, and infer that there is a genuine increased lung-cancer risk in a given population. Notice that null hypothesis H_0 asserts: any observed excess in risk is not due to a real underlying increase in risk—equivalently, it is *an error* to suppose a real risk is responsible. Thus in rejecting H_0 we are rejecting or denying the error asserted in H_0. But the significance level of .01, with which we reject this error, is *not* an assignment of probability to the null hypothesis that there's no increased risk in the population. Rather, .01 is the probability that such a test procedure would reject the null hypothesis erroneously—thus the term error probability. It asserts, in particular, that the probability of observing such a statistically significant excess in risk, if in fact the null hypothesis H_0 is true, is only .01. Were we observing a case where H_0 is true, so large an observed excess in risk would occur only 1% of the time.

Hence, the large observed excess in risk is taken to fail the no-risk hypothesis and pass the hypothesis, call it H, that some increased lung cancer risk exists. Hypothesis H passes a severe test because were H

false and the no-risk (null) hypothesis H_0 true, we would very probably (.99) have obtained a result that accords *less well* with H than the one we got. Equivalently, we infer the error asserted in H_0 is absent because, were it present, we would very probably have gotten a less statistically significant result.

Now let us consider a situation where a different result occurs. Suppose instead of observing an excess in lung cancer rates that the lung cancer rates are observed to be the same. This "no-difference" result accords with the no-risk (null) hypothesis H_0 and is *anomalous* for a hypothesis H that there *is* an increased risk of lung cancer (in the given populations). Were we nevertheless to take this result as support for H, or as passing H, that a risk exists, we would be running a test with a high probability of finding support for H *erroneously.* This high error probability corresponds to saying that H has passed a test with very *low* severity. This would signal the inference to H was unwarranted by this evidence.

But—and this is what I really want to emphasize—a test's high error probabilities, its low severity, alerts us to poor tests even in cases where the hypothesis accords well with the evidence. For example, confronted with the observation of no excess in cancer risk, a researcher might nevertheless hold on to hypothesis H that the treatment in question poses an increased lung cancer risk. The researcher might, for example, search the data for some other factor to explain why the difference did not show up. Perhaps the exposed group had consumed more vitamins and this compensated for the additional risk posed by the exposure to the substance in question. His favored hypothesis H together with this "compensation hypothesis" is made to accord with the (initially anomalous) evidence. Alternatively, he might search the evidence for an excess in a *different* health risk—one that does fit the data. For example, he may find a large enough excess in rates of high cholesterol. That is, had the initial null hypothesis been no increase in cholesterol risk (rather than lung cancer risk), the observed result would have rejected it at a low significance level. In this second strategy there has been a change in the particular health risk being looked for in order to show that the exposure in question produces a health risk.

The main thing to see is that in both of these cases, the hypothesis erected to accord with the evidence fails to pass a severe test. This shows up in the formal error probabilities associated with the test procedures. The probability of erroneously finding some alleged compensating factor *or other,* as in the first case, and the probability of erroneously finding an excess in some risk or other, as in the second case, is no longer the low .01 level as at the start, but is instead higher. Accordingly the hypothesis affirmed has no longer passed a severe test.

Let me be clear that I am not just stating this is so, these gambits for fitting the data are specifically in violation of the logic of the statistical significance test, and this shows up in the fact that the initial low error probability no longer holds.[6]

4.3. Upshot. I propose that Duhemian problems plague accounts of inference where the two features identified in 4.1 and 4.2 are absent: that is, accounts that seek a global measure of evidential support between any data and hypotheses, and accounts that are unable to discriminate hypotheses that "fit" the evidence equally well by appealing to the error probabilities of the overall testing process.

Error probability considerations provide the basis for distinguishing the well-testedness of two hypotheses—despite their both fitting the data equally well. The data may be a better, more severe, test of one than of the other. The reason is that the procedure from which the data arose may have had a good chance of detecting one type of error, but not of a different type of error. What is ostensibly the same piece of evidence is really not the same at all, at least not to the error theorist.

In contrast, any assessment where the import of the evidence on hypotheses reflects only some measure of the fit between them, without consideration of the reliability of the overall test, has no "slot" as it were within which to pick up on the difference we need when hypotheses fit equally well. Bayesians must find the difference in prior probabilities, but the distinction we want needs to be reflected in the reliability of the test itself.

In order to use the error-probability discernments, it is not necessary to compute a precise value of the probability, nor need one identify a specific statistical model that applies. As I see it, the distinctions afforded by error probabilistic criteria provide formal analogues to the kinds of discernments that we need in distinguishing warranted from unwarranted assignments of blame in more informal situations. They may be said to provide standard (or "canonical") models of classic ways of being led to assigning blame unreliably—so much the better to block them. This should become clearer in the examples of Section 6.

5. Some Strategies in the Error Statistical Approach to Duhem's Problem. It is important to see that there are two distinct types of strategies by which the error statistician grapples with Duhemian problems: (1) The first may be called "blocker" strategies, strategies to block or criticize attempts to explain away anomalies. We criticize attempts to explain

6. This may be described as the distinction between the observed (or computed) significance level and the actual significance level. See, for example, Mayo 1996, Ch. 9.

away anomalies (e.g., as due to H-saving factors) on the grounds that (a) they fail to pass severe tests, or, when possible, even more strongly, on grounds that (b) their denials pass severe tests. (2) The second is to show that an anomaly may be legitimately blamed on an auxiliary hypothesis A by showing that the denial of A, A′, passes a severe test. Clearly, we do not always have a warranted way to attribute blame, nor need we always have enough information to scrutinize attempts properly. But even then, these strategies direct progress with Duhemian problems and explain how scientists actually grapple with them.

Let us consider how these ideas apply to Duhemian problems, keeping to the kind of case in Dorling's illustration. In Dorling's illustration, a result e' that is anomalous for H is taken to hardly discredit H at all, but is taken as greatly discrediting the auxiliary hypothesis A needed to derive prediction e. In that reconstruction, anomaly e' is taken to provide positive grounds for discrediting A and confirming its denial A′. The degree of belief in A′ went from 0.4 to 0.99, by dint of anomaly e'. Hypothesis A′ clearly passes the Bayesian test, understood now as assigning high Bayesian support to A′. *But the error statistician wants to know if the test is severe.*[7] To show that A′ has passed a severe test would require positive evidence showing that the alleged extraneous factor is responsible for the anomaly. Strong belief in H together with low enough degree of belief in the Bayesian catchall factor, while sufficing for the high posterior belief in A′, do not suffice for showing A′ has passed a severe test. Indeed, it may be shown that the procedure Dorling endorses—going from satisfying the Bayesian conditions to declaring strong evidence for A′—is a very unreliable one. It makes it too easy to blame auxiliary hypothesis A even if A is true. Such an appeal to A′ would thereby be blocked by an error statistician.

Let me say a bit more about "blocking strategies," particularly of the stronger type (1)(a), that are available to the error statistician. In the typical presentation of Duhem's problem it is imagined that there are always a number of different factors to which the anomalous result can be ascribed. It seems to be assumed that so long as one puts forward a hypothesis A′ of form:

A′: factor F is responsible for anomaly e'

7. By making distinctions based on error probability considerations, the error statistician can be shown to be incoherent by Bayesian principles. The conflict between error statistical principles and Bayesian ones (i.e., the latter's adherence to the Likelihood Principle) is discussed at length in Mayo 1996. True, there is a relatively new school of Bayesians who attempt to import error probability assessments to Bayesian procedures—so-called "robust Bayesians"—but this error probabilistic brand of Bayesianism is not the Bayesian approach appealed to in solving Duhem.

then A′ entails or otherwise "fits" e' (i.e., that e′ is probable given A′). Thus, any such A′ is (erroneously) thought to have a high likelihood on evidence e'. In fact, a good part of the scientific work is to ascertain whether a hypothesized factor F, even if it *were* operative in the experiment, can actually account for the experimental results e'.

Again, the scientist proceeds like the doctor. The patient's headache may in principle be explained by any number of systems going wrong in the body but, in fact, a closer analysis of the body (blood counts, brain waves, CAT scans) may show that many of the possible factors could not actually account for this patient's headache (because of conflicts with the results of such an analysis). And notice, we may thereby rule out a purported explanation, not because it is less plausible or less frequent than others—it may be a common occurrence. It may be ruled out because it simply could not produce either the *particular extent* or the *particular pattern* of results discerned in these analyses.

A typical strategy is to create or simulate a situation in which the hypothesized factor F is given a very good chance to show that it is capable of bringing about e'—the effect which is anomalous for a hypothesis H. Then, if the effect does not show up, we can argue from error that F was not responsible. An attempt to save H by blaming F is blocked.

It may be objected that this ruling can always be gotten around by inventing an explanation for this "no-show," but this new attempt to save H cannot diminish the force of the blocking strategy just described. We bar a procedure for saving a hypothesis (by blaming an auxiliary factor) unless and until the procedure can be shown to be reliable. The onus is on the proponent of the hypothesized "save" to demonstrate that this is the case.

6. Some Duhemian Problems in the 1919 Eclipse Tests. It would be good to flesh out the main points with an example. Moving up 70 years or so from Dorling's example, pro-Newtonians are faced with another anomaly and the Duhemian question of where to lay the blame was addressed in great detail.[8] The anomaly concerned the deflection of light passing near the sun as was discerned during the 1919 eclipse expeditions undertaken by Eddington and others. The deflection effect, while predicted by Einstein's law of gravitation, was an anomaly for Newton's law of gravitation. If Newton's gravitational law was correct, and assuming light has mass, then the expected deflection effect would be only half what Einstein's law predicted.

The analysis of the results of the 1919 eclipse experiment was split

8. This episode is discussed in more detail in Mayo 1991 and 1996.

into two chief parts. The first part was to assess, using the eclipse results, the extent of the observed deflection effect (on light passing near the sun). This involved showing that the effect was real and not an artifact, as well as estimating the deflection effect using a standard statistical procedure (least squares). The second part was to ascertain whether the effect was attributable to the sun's gravitational field or whether some other factor consistent with a Newtonian account could explain the observed deflection.

Before too long, even staunch defenders of Newton felt compelled to accept the first part of the analysis—that the eclipse evidence showed the anomalous deflection effect was real. But they did not blithely reject Newton or think a modification was called for. Shortly after the deflection effect was affirmed, a joint meeting was held with the key players in this debate. One scientist, Ludwick Silberstein, expressed the views of many who attended. He suggested that the eclipse test result was an "isolated fact" which need not require a new gravitation law. Pointing to a portrait of Newton, Silberstein declared "We owe it to that great man to proceed very carefully in modifying or retouching his law of gravitation" (Silberstein 1919, 389–398).

6.1. Some Blocker Strategies: Alternative N-factors to Accommodate the Deflection Effect. A number of skeptical challenges revolved around the possibility of a mistake about the *cause* of the observed eclipse deflection. The question, in particular, was whether the test discriminated adequately the effect due to the sun's gravitational field from others that might explain the eclipse effect. A "yes" answer boiled down to accepting the following hypothesis:

> A: The observed deflection is due to gravitational effects (as given in Einstein's law), *not* to some other factor N.

A number of specific alternative factors were proposed by which to account for the anomalous deflection effect, without refuting Newton: Ross' lens effect, Newall's corona effect, Anderson's shadow effect, Lodge's ether effect, and several others besides.

If one attempted to express, in terms of degrees of belief, the attitudes of some of the Newtonians at this stage of the debate, one may well attribute to them just the kind of subjective beliefs that suffices for the Bayesian Way to show that 'the refutation' e' should have only a negligible effect on the scientist's degree of belief in [hypothesis H]" (Howson and Urbach 1989, 183)—Newton's gravitation law. Take the famous Newtonian, Sir Oliver Lodge. Lodge, a devout spiritualist made no bones about his passionate commitment to a Newtonian ether. He believed that it was through the ether that one could com-

municate with the souls of the departed. Absent a Newtonian ether, Lodge believed he would not be able to communicate with his dead son Raymond.

Lodge could well be seen as satisfying the subjective probability assignments that would have warranted taking the anomaly as only very slightly decreasing belief in Newton and greatly decreasing belief in auxiliary A that no factor other than gravity was operating. According to the Bayesian Way, Lodge is warranted in being practically certain that A is false and its denial true. In actuality, however, the depth of Lodge's belief did not strengthen, and in fact was no part of assessing, the evidence for the alternative factors that Lodge proposed. The question of alternative factors that could be responsible for the eclipse anomaly was a real and serious one, and it was split off from scientists' attitudes toward either Newton's or Einstein's account. They had, instead, to turn their attention to methods and arguments to test whether proposed alternative factors could be responsible for the observed deflection.

The challenges were conjectures that the effect was due to some factor *other* than the Einstein one (gravity in the sun's field). They were hypotheses of the following form:

> A′: The observed deflection is due to factor N, other than gravitational effects of the sun

where N is a factor that at the same time saved the Newtonian law from refutation.

No one, not even staunch Newtonian defenders, thought the anomalous deflection effect itself was strong evidence for A′, that these other factors were operating. If they had, they would not have gone to the lengths that they did in order to try and provide positive grounds for particular proposed alternative factors. Proposed factors by which to save Newton were evaluated according to whether they stood up to severe scrutiny; when they did not, they were shot down. The concern was that the resulting accommodation of the results failed to constitute a *reliable test* in favor of the hypothesized auxiliary factor or N-factor.

The debates over conjectured N-factors went on for about 3 years (scattered through the relevant journals from 1919 to around 1921). What made the debate possible, and finally resolvable, was that all who would enter the debate were held to shared standards for what could count as evidence against auxiliary hypothesis A.

Each such hypothesis was criticized by means of a two-pronged attack: (i) the effect of the conjectured N-factor is too small to account for the eclipse effect; and (ii) if the N-factor were large enough to account for the eclipse effect, it would have other false or contradictory

implications. Each prong was justified by an argument based on severity, and the arguments bolstered one another.

Anderson's shadow effect, for example, proposed that the cooling from the moon's shadow would act as a lens deflecting light rays passing through the shadow. Its critique was typical: It was shown that under the actual conditions of the eclipse test, the shadow lens explanation would require a serious drop of temperature that was not observed. Moreover, even under differences in the assumed conditions, the shadow lens effect was negligibly small compared to the observed effects. (See Moyer 1979, 84.) The reasoning goes like this: If one grants the story that the defender of A′ invokes about the phenomenon in question, e.g., if we accept that the moon's shadow functions in the way described in Anderson's shadow effect hypothesis, then A′ cannot account for the results in the actual experiment conducted.

In other words, A′ asserts two things: factor N, operating in such and such ways, causes the deflection effect e'. The key question is whether, in the actual circumstances of the experiment, the factor (e.g., shadow effect) hypothesized in A′ could account for the effects observed. Although hypothesis A′ asserts that alternative factor F is the reason for the observed anomaly e'—it is a mistake to suppose that A′ can thereby account for e'. A severe test of A′ requires ruling out the ways A′ might erroneously be thought to explain e'. Looking at the specific data points, and other features of the experimental context (e.g., temperature differences) it may be discerned that A′ really does not account for observed effects.[9]

To nevertheless uphold A′ as the way to accommodate the anomaly, it was shown, was to commit a classic case of what is *dis*allowed in saving a threatened theory. It would make it easy for a hypothesis of form A′ to pass, even if it is false and auxiliary hypothesis A is true (i.e., high error probabilities). The test it passes fails to be severe.

6.2. An Example of Discounting an Alleged Anomaly. The eclipse episode also includes a much-discussed[10] instance where an apparent anomaly was explained away successfully by invoking an alternative hypothesis A′. I am thinking of how Eddington was able to explain away an apparent anomaly—this time for the Einstein hypothesis. The

9. For a crude analogy, a hypothesis about a given type of bomb in the cargo hold might be said to give an account for the "anomaly"—an explosion of a jet—but particular features of the plane, residues, damage "signatures," etc. may show that the cargo bomb explanation is in error.

10. See, for example, Glymour and Earman 1980 and Mayo 1991.

apparent anomaly stemmed from one of the sets of eclipse results (from Sobral) which pointed, not to Einstein's prediction, but, as Eddington declares, "with all too good agreement to the 'half-deflection', that is to say, the Newtonian value . . ." (1920, 117). The debate over where to lay the blame was engaged in by scientists with very different opinions about Einstein's theory. Such attitudes were no part of the arguments deemed relevant for the question at hand. The relevant argument, put forth by Eddington (and others), turned on a rather esoteric piece of data analysis showing (holdouts notwithstanding) that the alleged anomaly was not a genuine one, but was caused by a distortion of a mirror in recording the star positions. It was based on analyzing the patterns of errors on the two different days during which star photos were taken. Consider the actual notes penned by Sobral researchers:

> May 30, 3 a.m., four of the astrographic plates were developed. . . . It was found that there had been a serious change of focus . . . This change of focus can only be attributed to the unequal expansion of the mirror through the sun's heat. (Dyson, et al. 1923, 309)

Being affirmed, in short, was the following:

> A′: The results of these (Sobral) plates are due to a systematic distortion (of the mirror) by the sun, and not to the deflection of light.

As the journals of the period make plain, the numerous staunch Newtonian defenders would hardly have overlooked the discounting of an apparently pro-Newtonian result, if they could have mustered any grounds for deeming it biased. The reason they could not fault the explanation in A′ is that it involved well understood principles for using this type of data to test and, in this case, reject, a key assumption of the experiment. Results were deemed usable for estimating the deflection effect only if a common error-statistical method (i.e., least squares regression) was applicable. This demanded sufficiently precise knowledge of the change of focus (scale effect) between the eclipse and night plates (within 0.03 mm)—precisely what was *absent* from the suspect Sobral results.[11]

These examples from the eclipse tests instantiate the two distinct strategies by which the present approach grapples with Duhemian problems. The first is to criticize and bar attempts to explain away anomalies (e.g., as due to the Newton-saving factors) on the grounds

11. Even small systematic errors of focus are of crucial importance because the "scale effect" that results from this alteration of focus quickly becomes as large as the Einsteinian predicted deflection effect of interest. See von Klüber 1960, 50.

that they fail to pass severe tests (or, even more strongly, that their denials pass severe tests). The second is to show that an anomaly may be legitimately blamed on an auxiliary factor F (e.g., a mirror distortion) by showing that "F is responsible" passes a severe test. Clearly, we do not always have a warranted way to attribute blame; we can not always satisfy the requirement of the second strategy. But this requirement directs progress with Duhemian problems, and it explains the lengths to which scientists work to test auxiliaries.

I do not claim that there is not much more to be said to develop a full-blown error statistical solution to Duhem's problem. What I do claim is that it provides the right kind of methods and principles for tackling this and other problems about evidence—the very principles which stand in marked contrast to the Bayesian approach. A major virtue of the error statistics approach is that the issue of whether a primary hypothesis or an auxiliary is discredited is not based on the relative credence accorded to each. The experiment is supposed to find out about these hypotheses, it would only bias things to make interpreting the evidence depend on antecedent opinions. After all, in Dorling's example, and I agree the assumption is plausible, hypothesis H is said to be *independent* of auxiliary A. There is no reason to suppose that assessing auxiliary A should depend at all on one's opinion about H. What are called for are separate tools to detect whether specific auxiliaries are responsible for observed anomalies, tools for discriminating signals from noise, ruling out artifacts, distinguishing backgrounds, and so on. And these tools should be applicable with the kind of information scientists actually tend to have or can obtain. The conglomeration of methods and models from standard (non-Bayesian) error statistics provides such tools. Scientists are free to hypothesize that an extraneous factor, and not H, is to be blamed for an anomalous result, no matter how personal or passionate their reasons for doing so. But to *warrant* that hypothesis requires it to have passed a severe test. Degrees of belief have nothing to do with it.

APPENDIX
Calculations for the Homework Problem

It is given that A and H entail e, but e' is observed: $P(e'|A \text{ and } H) = 0$

The assumed prior probabilities:

$$P(H) = .9,\ P(A) = .6,$$

(i) Hypotheses A and H are statistically independent

The assumed likelihoods:

(ii) $P(e'|A \text{ and } \sim H) = \varepsilon$ (very small number, e.g., 0.001)

(iii) (a) $P(e' \mid \sim A \text{ and } \sim H) = 50\varepsilon$
(b) $P(e' \mid \sim A \text{ and } H) = 50\varepsilon$

Bayes's theorem: $P(H|e') = \dfrac{P(e'|H)\,P(H)}{P(e')}$

From the above we get the following:

$$P(e') = P(e'|H)P(H) + P(e'|\sim H)P(\sim H).$$

$$\begin{aligned} P(e' \mid H) &= P(e' \mid A \text{ and } H)\,P(A) + P(e' \mid \sim A \text{ and } H)P(\sim A) \\ &= 0 + 50\varepsilon(.4) \\ &= 20\varepsilon. \end{aligned}$$

$$\begin{aligned} P(e' \mid \sim H) &= P(e' \mid A \text{ and } \sim H)P(A) + P(e' \mid \sim A \text{ and } \sim H)P(\sim A) \\ &= \varepsilon\,(.6) + 50\varepsilon(.4) \\ &= 20.6\,\varepsilon. \end{aligned}$$

So, $P(e') = 20\varepsilon(.9) + 2.06\varepsilon = 20.06\varepsilon$.
The posterior of H can now be calculated:

$$P(H \mid e') = \frac{20\varepsilon(.9)}{20.06\varepsilon} = .897$$

To calculate the posterior probability $P(A|e') = \dfrac{P(e'|A)P(A)}{P(e')}$

$$\begin{aligned} P(e' \mid A) &= P(e'|A \text{ and } H)P(H) + P(e' \mid A \text{ and } \sim H)P(\sim H) \\ &= 0 + \varepsilon(.1) = .1\varepsilon \end{aligned}$$

So,

$$P(A|e') = \frac{.06\varepsilon}{20.06\varepsilon} = .003$$

REFERENCES

Dorling, J. (1979), "Bayesian Personalism, the Methodology of Scientific Research Programmes, and Duhem's Problem", *Studies in History and Philosophy of Science* 10:177–187.

Duhem, P. (1954), *The Aim and Structure of Physical Theory.* Translated by P. Wiener. New York: Atheneum.

———. (1996), *Essays in the History and Philosophy of Science.* Translated and edited by R. Ariew and P. Barker. Indianapolis: Hackett.

Dyson, E. W., A. S. Eddington, and C. Davidson (1923), "A Determination of the Deflection of Light by the Sun's Gravitational Field, from Observations made at the Total Eclipse of May 29, 1919", *Memoirs of the Royal Astronomical Society,* Vol. LXII (1917–1923): 291–333.

Earman, J. and C. Glymour (1980), "Relativity and Eclipses: The British Eclipse Expeditions of 1919 and Their Predecessors", *Historical Studies in the Physical Sciences* 11: 49–85.

Eddington, A. (1918), "Gravitation and the Principle of Relativity", *Nature* 101 (March 14, 1918): 34–36.

———. (1919), "Joint Eclipse Meeting of the Royal Astronomical Society", *Observatory* 42: 389–398.

———. (1920), *Space, Time and Gravitation: An Outline of the General Relativity Theory.* Cambridge: Cambridge University Press (as reprinted in the Cambridge Science Classics series, 1987).

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

Howson, C. and P. Urbach (1989), *Scientific Reasoning: The Bayesian Approach.* La Salle: Open Court.

Lindemann, F. A. (1919), (contribution to) "Discussion on the Theory of Relativity", in *Monthly Notices of the Royal Astronomical Society* LXXX (Dec.): 96–118, p. 114.

Lodge, O. (1919), (contribution to) "Discussion on the Theory of Relativity", in Monthly Notices of the Royal Astronomical Society, LXXX (Dec.): 96–118, pp. 106–109.

Mayo, D. (1985), "Behavioristic, Evidentialist, and Learning Models of Statistical Testing", *Philosophy of Science* 52: 493–516.

———. (1991), "Novel Evidence and Severe Tests", *Philosophy of Science* 58: 523–552.

———. (1996), *Error and the Growth of Experimental Knowledge.* Chicago: the University of Chicago Press.

Moyer, D. (1979), "Revolution in Science: The 1919 Eclipse Test of General Relativity Theory". Edited by A. Perlmutter and L. Scott, 55–72. New York: Plenum Press.

Nature 106: 781–820 (February 1921).

Newall, H. F. (1919), (contribution to) "Joint Eclipse Meeting of the Royal Society and the Royal Astronomical Society", *The Observatory* 42 (Nov.): 389–398, pp. 395–396.

Redhead, M. (1980), "A Bayesian Reconstruction of the Methodology of Scientific Research Programmes", *Studies in History and Philosophy of Science* 11: 341–347.

Shimony, A. (1970), "Scientific Inference", in R. Colodny (ed.), *The Nature and Function of Scientific Theories: Essays in Contemporary Science and Philosophy.* Pittsburgh: University of Pittsburgh Press.

Silberstein, L. (1919), (contribution to) "Joint Eclipse Meeting of the Royal Society and the Royal Astronomical Society", *The Observatory* 42 (Nov.): 389–398.

von Klüber, H. (1960), "The Determination of Einstein's Light-Deflection in the Gravitational Field of the Sun", in A. Beer (ed.), *Vistas on Astronomy,* Vol. 3, pp. 47–77.

Worrall, J. (1993), "Falsification, Rationality, and the Duhem Problem", in J. Earman, A. Janis, G. Massey, and N. Rescher (eds.), *Philosophical Problems of the Internal and External Worlds. Essays on the Philosophy of Adolf Grunbaum.* Pittsburgh: University of Pittsburgh Press.