Response to Howson and Laudan
Author(s): Deborah G. Mayo
Source: *Philosophy of Science*, Vol. 64, No. 2 (Jun., 1997), pp. 323-333
Published by: The University of Chicago Press on behalf of the Philosophy of Science Association
Stable URL: https://www.jstor.org/stable/188312
Accessed: 18-10-2020 23:02 UTC

Response to Howson and Laudan*

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

A toast is due to one who slays
Misguided followers of Bayes,
And in their heart strikes fear and terror
With probabilities of error![1] (E.L. Lehmann)

1. Response to Howson.

1.1 A Cavalier Treatment of Anomalies for the Bayesian "Solution" to Duhem. I, and several others (Worrall 1993, Earman 1992, Laudan (this volume)), have articulated a number of obstacles or anomalies for the subjective Bayesian attempt to solve Duhem's problem, the problem of which hypothesis ought to be blamed in the face of anomaly. Howson has not answered these challenges. Consistent with the cavalier treatment of anomalies that his favored solution countenances, Howson's response comes down to reiterating, in a variety of forms, his prior and deep commitment to the fundamental rightness of the Bayesian Way.

What do we learn from Howson in response to the charge that a Bayesian reconstruction does not show which hypothesis it is *warranted* to credit or blame? Or to the charge that the Bayesian Way does not accord with how Duhem's problem is actually grappled with in science? Or to the charge that, in practice, it is error statistical methods that are appealed to in checking auxiliaries, distinguishing real effects from ar-

*Received December 1996; revised September 1997.

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

1. This verse, intended to be taken in a jocular vein, was Lehmann's reaction to my recent attempts to do battle with subjective Bayesians in philosophy of science. I thank him for the poem and for allowing me to share it here.

Philosophy of Science, 64 (June 1997) pp. 323–333. 0031-8248/97/6402-0008$2.00

tifacts, estimating backgrounds, discriminating different errors and so on in all the tasks called for in reliably pinpointing blame?

We learn from Howson that scientists have simply not yet discovered that the error statistical "NP criteria are simply fallacious."

But would not a plausible alternative be that scientists actually find error statistical guarantees useful for the very tasks that a reliable solution to Duhem demands? Howson seems not to entertain this hypothesis at all—I guess he gives it a zero prior probability. But surely the hypothesis he discounts is supported by the evidence I have offered (this issue and 1996) that error statistical methods are of value because they supply *severe tests* in the sense I have set out. Faced with this indictment of the Bayesian solution, Howson promotes an alternative hypothesis that saves the Bayesian Way—scientists are just confused by the ambiguities of ordinary language:

> It is not surprising, in view of the difficulty of distinguishing the conditional probabilities [P(H is false|H is accepted) and P(H is accepted|not-H)] . . . that NP ideas have been so tenacious. It *is* surprising that the fallacious nature of NP inferences is taking so long to be recognized. (Howson, this issue)

Identifying, quite incorrectly, my notion of severity with the NP probability of a type 2 error, Howson declares that it too is infected with the fallacy.

What grounds are there for accepting the hypothesis that scientists and others are simply confused by the ambiguities of ordinary language? Howson's answer is that "what everyone really wants to know from the outcome of the test is" the conditional probability that H is false given it is accepted, and error probabilities do not give us this. But is not this "confusion hypothesis" undermined by the numerous articles in which statisticians have warned against confusing an error probability with a posterior probability to a hypothesis? Presumably not enough to seriously decrease Howson's belief in it. How then can we account for the fact that Neyman, to mention just one famous example, spends numerous articles and chapters demonstrating the contradictions that can result if one commits the confusion Howson alleges he commits? Since these demonstrations go unmentioned, perhaps they are not part of Howson's background knowledge.

Granted, Howson's subjectivism frees him to hold these beliefs, however implausible. But Howson goes further; he argues that since it is obvious that posterior probabilities of hypotheses are what everybody wants, and yet NP error probability criteria do not give them to us, then NP theory is unsound. "Thus NP criteria are simply fallacious,

or in logicians' terminology demonstrably unsound rules of (inductive) inference" (Howson, this issue).

1.2 Unsoundness? No. A Fundamental Difference in Aims? Yes. Fortunately there are some terms which cannot be tinkered with to suit one's preferences—and "soundness" in logic is one. The error probability criteria of NP tests are deductively sound—and Howson cannot make them unsound by insisting that the NP methods are really intending to give us posterior probabilities (but instead they only give us error probabilities)!

The founders of NP methods could not be clearer in explaining that these methods are intended to satisfy error probabilistic criteria quite without any regard to the prior probabilities of hypotheses—probabilities which can only make sense for a frequentist where hypotheses can be regarded as random variables. Since calculating posterior probabilities via Bayes' theorem requires the introduction of prior probabilities (to an exhaustive set of hypotheses) and since these are unavailable in testing scientific hypotheses regarded as true or false, posteriors are *quite deliberately not* made the goal of NP tests. For Howson to charge that these methods are not really supposed to be doing what they clearly were developed to do, and that they are really supposed to be doing what they clearly were developed to avoid, seems bizarre.

If Howson instead provided an argument for thinking that we ought to have a *different* goal and showed that subjective Bayesianism gives us that goal, that could at least be considered. But no unsoundness charge levied at NP statistics would thereby have been shown.

Now Howson does go on to claim that the Bayesian Way, in contrast to error statistics, gives us what we really want. But does it?

Let us see. To pick up on an example already discussed in my paper (this issue), we saw that various degree of belief assignments would entitle a staunch Newtonian (e.g., Lodge) to accept A′: that the deflection effect was due to some factor N that saved the Newtonian law from refutation, say the "shadow lens effect." What everyone wants to know, says Howson, is the conditional probability that A′ is false given it is accepted by this Bayesian rule (with Lodge's degrees of belief). But the Bayesian account has surely not told us how probable the lens hypothesis is given that the Bayesian accepts it, and it surely has not given any assurances that it is unlikely to be false given that the Bayesian accepts it.

So what *has* it given us? It has given us only the subjective degree of belief that A′ is false given that a Bayesian agent has accepted A′. What the Bayesian has given us comes down to a tautology: if the agent

assigns a high posterior degree of belief to the falsity of A′, then the subjective probability that A′ is false is high. For the error statistician, a hypothesis is warranted only if its errors have been well ruled out; this is not required in order for hypotheses to be strongly believed.

Thus, a Bayesian posterior is not at all what I want in evaluating a hypothesis. If Howson reports to me his degree of belief in a hypothesis, I would want to know if his belief were warranted. If I could assess the frequency of errors his assessment commits I may begin at least to scrutinize his beliefs—but if the criterion is only inner coherency, then the assessment is of little use. By liberating itself from the constraints of objective error probabilities, the subjective Bayesian may be said to be freer than the error statistician, but he cannot be said to be giving us what we want in science. Howson's declaration that "personal opinions are ineradicable" carries little weight with scientists who have enjoyed the power of robust tools with intersubjectively checkable assumptions.

Thus, Howson's attacks on the error statistical account fail to provide him with the defense required to counter charges that the Bayesian solution to Duhem is no solution at all. As Kevin Kelly puts it: "[U]sually when a method is called a solution to a problem, it is understood that the method's assignments *reliably* indicate a *correct* answer to the problem" (Kelly, this issue). Not to Howson. Quite confident that the tide will soon turn in the statistical community, Howson concludes his article in the spirit of a true counterinductivist: "Why it is taking the statistics community so long to recognize the essentially fallacious nature of NP logic is difficult to say, but I am reasonably confident in predicting that it will not last much longer."

1.3 A Misleading Example. Some comment is owed to an example Howson provides which he regards as showing that it is the posterior probability that provides a guide "to the correctness or reliability of the hypothesis tested" (Howson, this issue), rather than error probabilities of tests. His example follows a common gambit by Bayesian critics. To construct an example where even a frequentist could view a hypothesis as a random variable, it is imagined that we sample randomly from a population in which some proportion has a given property (e.g., a deficiency or disease), and that the hypotheses, H and H′, assert that the sample selected does or does not have the property. However, in all such examples,[2] the hypotheses are forced to be statements about the particular *sample* and are not *statistical hypotheses*.

2. At least, I have not seen any that escape this problem. See also my critique of an analogous example given by Howson in Mayo 1997.

Once one converts these inadmissible hypotheses into legitimate statistical ones, the prior probabilities (derived from the random sampling device) are no longer kosher for an error statistician with respect to the initial sample space. The error statistician can apply the usual error statistical tests which, after all, were expressly designed to avoid any assignment of prior probabilities to hypotheses. Doing so, Howson shows, yields an interpretation of the evidence that conflicts with the one that his Bayesian posterior licenses. But far from vitiating the error statistical test, Howson's own example demonstrates why we would *not* want to regard the Bayesian posterior of H as indicating H's correctness, or as performing the service that statistical tests can provide.

In Howson's example, it is assumed that (i) a given population has a very high proportion, say .999, with a certain disease, and (ii) any randomly chosen test subject from this population has a (prior) probability of having the disease equal to .999. Although I cannot think of a disease possessed by .999 of a population, I can imagine a *deficiency* possessed by a very high proportion of a group, say the lack of readiness for work in a competitive 4-year college. In Howson's example, the null hypothesis H asserts this deficiency is absent, so, using this example, we have

H: the student is ready for college—the student is *not deficient*,

and H′ asserts that the deficiency is present:

H′: the student is not ready for college—the student is *deficient*.

We are allowed to consider only two possible outcomes: e, say a passing grade on an college admissions test, and e', a failing grade on the test.

Now hypotheses H and H′ are not statistical hypotheses but are assertions about particular samples. A statistical hypothesis would have to assign probabilities to the possible outcomes, here e and e', and knowing which of H and H′ is true about my sample does not give me that probabilistic assignment. But we can convert these hypotheses into statistical ones, and in doing so I will use the probability assignments that Howson wants us to consider.

Howson stipulates that non-deficient (or "college-ready") students pass the test with probability 1 (which is too high to be realistic, but I leave that to one side), while deficient students pass only 5% of the time. Hypothesis H, statistically rendered, would assert that the probability of e (a passing grade) is 1, and hypothesis H′ that the probability of e is .05. They are now statistical hypotheses relating to the possible outcomes, e and e'. We observe a student, say Mary, and find she has scored passing grade e on the college admissions test. Does this passing

grade indicate Mary is deficient? Howson says "yes," while the error statistical test says "no."

According to Howson, "correctness is a (i) quality," that is, it is a matter of calculating a posterior probability. Since the posterior probability P(H′ *e*) is high (i.e., .98), the hypothesis whose correctness is indicated, Howson thinks, is H′—Mary is deficient. But notice that whether or not a student passes this admissions test (whether *e* or *e*′ occurs), Howson's Bayesian analysis *always* regards the correct indication to be H′: lack of readiness! This is due to the fact that the student happened to be selected from a population where college readiness is very rare, and to Howson's assumption (ii) above. Rather than providing grounds to favor the Bayesian construal of what *e* indicates, the example is grounds for denying that Howson's diagnostic tool has perfomed the intended job. Howson has chosen H to be the null hypothesis of the error statistical test, and has set a 0 probability of erroneously rejecting H (i.e., a 0 probability of a type 1 error). This signals, for an error tester, that we want college-ready students to have the maximal chance of being detected. We do *not* want to infer H′: the student is deficient, unless the student has been given an excellent chance to show readiness, and yet she fails to do so (she scores a failing grade *e*′). By taking Mary's passing grade as indicating her lack of readiness, Howson's analysis flouts this goal. Scoring the passing grade *e* hardly shows *a lack* of readiness—and on these grounds, the error statistician denies that *e* indicates H′.[3] Yet Howson says that "what everyone really wants to know" in order to assess the correctness of H′ is its Bayesian posterior probability given *e* —which is .95. But reporting that the Bayesian posterior in H′ is .95 does not tell me what I want to know in discerning if Mary's score indicates readiness or not. I would also want to know how reliable this test is. And because the Bayesian test has given Mary *no chance* to show her readiness, even if she is ready, H′ would not be said to have passed a severe or reliable test.

This example does highlight the disparity between the error statistical and the Bayesian aim in such a test. Had Mary been randomly selected from a different population (perhaps a well-to-do suburb rather than a disadvantaged area)—one where, say, only half the students are deficient, then the prior probability that Mary is deficient, according to Howson, is 0.5. The identical score would now yield a very *low* posterior probability hypothesis H′. So the very same test

3. This does not require the error statistician to takes *e* as good grounds for H. The test Howson defines is too coarse-grained to assess the *magnitude* of the deficiency that is or is not indicated. Such an assessment of the test's power to discriminate departures from H is required to interpret a failure to reject a null hypothesis.

score that was taken as a good indication that Mary is deficient becomes a good indication that she is *not* deficient. While this strategy might yield sensible average losses of some sort (in repeatedly laying bets over the given population from which Mary is selected), things are very different if the goal is ruling out an erroneous interpretation of the ability of *this* student, Mary. In appraising what the test score indicates about Mary, the error statistical assessment does not average over the population of *other* students from which she happened to have been sampled. The goal of the error statistical analysis is to determine if the evidence does a good job of ruling out the error in assessing *this* hypothesis (wrongly taking Mary as being deficient). This is what the appropriate error probabilities provide.[4]

2. Remarks on Laudan.

2.1 What is the Question? As always, Laudan goes right to the heart of the matter. At issue is not so much what is the answer to Duhem's problem, but what is the question? Whereas Laudan says I am at risk of "solving not Duhem's problem but quite a different one" (this issue), it seems to me that it is Laudan who is after a different problem—one which is designed to escape but not solve the original problem, and whose solution depends upon already having a theory of testing in place with the resources to solve Duhem's initial problem.

The problem that I am trying to solve *in this particular paper* is the one Duhem initially puts to us. Since a scientist "can never submit an isolated hypothesis to the control of experiment," Duhem reasons, an experimental disagreement with a prediction does not show us the particular hypothesis that is at fault. My response is that (1) we do not need to actually control all factors in order to pinpoint blame correctly, and (2) if we would get beyond a "white glove" analysis of anomalous results, we would find they can be made to speak volumes about their source. To find "what is wrong," and do so correctly, it suffices to be able to:

- distinguish the pattern of effects of different factors (e.g., a mirror distortion from a deflection effect);

4. I am not saying that it is impossible to imagine a legitimate objective prior probability in hypotheses H and H′—I am denying that the assignment derived from randomly sampling from H's and H′'s gives a prior that may legitimately be combined with the admissible statistical rendering of these hypotheses. A legitimate prior would have to refer to the probability that *statistical* hypothesis H is true of the given student, Mary. The indeterminacy; the fact that H does not simply have either probability 0 or 1, could perhaps be due to the various genetic and environmental factors that determine if she is deficient or not.

- learn enough about the extent of an observed effect that could be attributable to a given factor (often by simulations) in order to "subtract it out" (from the anomaly);
- distinguish between hypothesized explanations of the cause of an anomaly by distinguishing the severity of the tests each passes.

As I see it, I have not changed Duhem's problem but rather, that problem has been changed by others, notably Laudan himself, on account of the presumption that Duhem's problem—in its original guise—cannot be, or at least has not been, solved. The reason the problem is given up on is that the available appraisals of experimental support or testing seem always to underdetermine the correct way to pinpoint blame. For any account of experimental testing, it is alleged, there are always several different combinations of theories and auxiliary hypotheses that equally well accommodate the anomaly in question. But scientists do actually adjudicate between these conflicting ways to account for anomalies; thus, it is concluded that the adjudication must involve considerations beyond the mere well-testedness of the hypotheses at hand. The additional considerations, it is supposed, involve criteria for appraising large scale theories, such as problem-solving ability, explanatory power, and scope; and Laudan berates me for leaving out such considerations in my approach to Duhem's problem. What I am alleging, however, is that the move from Duhem's initial problem to the problem of which large-scale theory to accept or prefer, stems from giving up on (or at least looking away from) the original challenge. Laudan developed his problem solving account to be deliberately "immune from criticism of a Duhemian type" (1977, 42), suggesting that "a way out of the Duhemian conundrum may emerge if, far from localizing blame or credit in one place, we simply spread it evenly among the members of the complex" (1977, 43). The problem then *becomes* that of setting out criteria for preferring a better complex of theories, such as explanatory power.

As popular as this move to large-scale appraisals has become, I have the chutzpah to suggest that the initial challenge was given up on too easily, or, at any rate, I want to suggest we go back to it. Perhaps we have learned something about testing over the years, and perhaps a better account of testing can now disambiguate anomalies in ways that older accounts of testing could not.

It seems to me that the reason for the initial capitulation, in broad terms, is that whatever *base relation* that was tried in erecting an account relating evidence to hypothesis—whatever relation of "fit" that was put forward—there were still too many conflicting ways to fit the available evidence. Now the Bayesian has a way to distinguish different

hypotheses that fit the evidence equally well—namely, through differences in their prior probability assignments, generally construed as subjective degrees of belief. The error statistician rejects this way, and argues instead for an account of testing that distinguishes different hypotheses that fit data equally well by considering the error probabilities of the overall testing processes. (All of this I discuss in Sections 4.1–3.)

Thus I can agree with Laudan that "we were initially seeking a way to avoid the apparent ambiguity of falsification as it affects the principal theories of the physical sciences," though I would extend this to include tests of lower level theories where no large-scale or"principal theory" even exists. What I deny is that resolving these ambiguities calls for an account of accepting or appraising large-scale theories, as Laudan sees it. More specifically, I deny that the Duhemian ambiguities—where they can be resolved—require discriminations other than those that are amenable to severe testing in my sense. Laudan has not produced an example that shows otherwise.

Thus, where Laudan wants to help steer me towards what he regards as the central task, I in turn wish to steer philosophers of science back to the initial Duhemian challenge.

2.2 An Account of Local Testing is Already Needed. While making obeisance to error statistical tools for the more local tasks of arriving at data and determining the presence of an anomaly—two very important parts of the full Duhem problem—when it comes to determining the cause of the anomaly or to discriminating hypothesized ways to account for it, Laudan sees no further role for such local testing. The task with which Laudan is concerned gets started only *after* all the messiness of data generation, modeling and analysis is over and a neat and tidy package—"the anomaly"—is in front of us. This "theory-dominated" perspective on hypothesis appraisal is the very thing that I, following other "new experimentalists," want to question. It is not that I think those concerned with large-scale theory change should suffer through the messy practices by which raw data are turned into a reliable data report, it is that by overlooking these practices, "theory-firsters" drastically shortchange themselves. For it is here that the tools for the disambiguation they seek may be uncovered.

Moreover, one cannot properly begin to attack Duhem's problem at the stage that Laudan recommends, because the crucial tasks in his Stages I-V already require an account of testing adequate to resolve Duhemian worries. For instance, Laudan's need to ask "whether there is any independent evidence for A′" (a denial of auxiliary hypothesis A), *already* assumes one has a way to determine whether A′ has passed

a good test—and that is just to solve Duhem (for A′). It is likewise for determining "which complex, among (H&A′), (H′&A), and (H′&A′) is both better tested and more complete." Laudan says we must take "stages II, III and IV seriously *before* deciding the epistemic impact of the failed prediction," and his idea of severe testing only gets started *after* all this is done and one is ready to tally up the score earned by the different conjunctions of theories and hypotheses. But by that time, in my view, all of the difficult work must have already been done. We must already have been able to discriminate which hypotheses have passed tests that are genuine, independent, reliable, and non-ad hoc, and which have not. And if one does not have a means to adjudicate the different hypotheses reliably, it does not much matter how nifty one's scheme is for tallying and comparing the number of tests rival theories pass.

I am certainly not denying that there are severe tests of higher level theoretical hypotheses—though I deny this is demanded by Duhem's problem. The results from piecemeal studies may be accumulated so as to allow us to say, at some point, that several related hypotheses are correct, or that a theory solves a set of experimental problems correctly. But this stage already depends upon multiple uses of local tests, each of which must solve its Duhemian problems.

2.3 Avoiding the Bayesian Catchall Factor. Despite Laudan's worries, calculating severity does not confront us with "a probability calculation involving an indefinitely long disjunction of rival hypotheses to H." Our approach to Duhem's problem is to recognize that we learn by ruling out specific errors and by making modifications based on errors. Using simple contexts in which the assumptions may be shown to hold sufficiently, it is possible to ask one question at a time. Setting out all possible answers to this one question becomes manageable, and that is all that has to be "caught" by our not-H.

Within an experimental testing model, the falsity of a primary hypothesis H takes on a very specific meaning. If H states a parameter is greater than some value c, not-H states it is less than c; if H states that factor F is responsible for at least p% of an effect, not-H states it is responsible for less than p%; if H states the effect is systematic—of the sort brought about more often than by chance—then not-H states it is due to chance. Each "not-H" may be tested by constructing an appropriate statistical null hypothesis. Scientists may explore, quite apart from testing some underlying, large-scale theory (which may not even be in place), whether neutral currents exist, whether dense bodies are real or merely artifacts of the electron microscope, whether the rate of HIV replication is reduced by protease inhibitors, whether F4 and F5

chromosomes play any part in certain types of Alzheimer's disease, and so on. In setting sail on such explorations, anomalies are distinguished and are learned from; one is not at sea, worrying about all of the ways one could ever be wrong in theorizing about the given domain.

Consider the handling of anomalies for Newton posed by the deflection experiments (Section 6 of my paper). I do not see that *that* problem is grappled with by asking whether Einstein's theory can pass all the same tests that Newton's can. The question, rather, is always: can this hypothesized N-factor (e.g., the shadow-lens effect) account for *these particular* anomalous results? If each such hypothesis were not shot down in the way I maintain, then the Newtonians would not have gone on to immediately propose yet other explanations. Laudan's "wait and see" attitude to grappling with Duhemian worries simply does not accord with the way scientists actually reasoned. If Laudan concedes that of course scientists had to test these various N-saving factors individually, then he must concede that solving Duhem needs to be accomplished prior to any comparative assessment of the explanatory track record of the respective large-scale theories. (And anyone who looks at these tests will see that they posed enormous tasks of disambiguating anomalies.)

Finally, Laudan's comparative account of severity, "a theory has been severely tested provided that it has survived tests its known rivals have failed to pass (and not vice versa)," redolent of Popper's or Lakatos', is neither necessary nor sufficient for severity in my sense (Mayo 1996).

REFERENCES

Howson, C. (1997), "Error Probabilities in Error", *Philosophy of Science* 64 (Proceedings): forthcoming.

Laudan, L. (1977), *Progress and Its Problems.* Berkeley: University of California Press.

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

———. (1997), "Error Statistics and Learning From Error: Making a Virtue of Necessity", *Philosophy of Science* 64 (Proceedings): forthcoming.