PRASANTA S. BANDYOPADHYAY and GORDON G. BRITTAN, JR.

ACCEPTIBILITY, EVIDENCE, AND SEVERITY[1]

ABSTRACT. The notion of a *severe test* has played an important methodological role in the history of science. But it has not until recently been analyzed in any detail. We develop a generally Bayesian analysis of the notion, compare it with Deborah Mayo's error-statistical approach by way of sample diagnostic tests in the medical sciences, and consider various objections to both. At the core of our analysis is a distinction between *evidence* and *confirmation* or *belief*. These notions must be kept separate if mistakes are to be avoided; combined in the right way, they provide an adequate understanding of severity.

> Those who think that the weight of the evidence always enables you to choose between hypotheses "ignore one of the factors (the prior probability) altogether, and treat the other (the likelihood) as though it ... meant something other than it actually does. This is the same mistake as is made by someone who has scruples about measuring the arms of a balance (having only a tape measure at his disposal ...), but is willing to assert that the heavier load will always tilt the balance (thereby implicitly assuming, although without admitting it, that the arms are of equal length!). (Bruno de Finetti, *Theory of Probability*)[2]

0. INTRODUCTION

In a justly famous essay first published more than fifty years ago,"Studies in the Logic of Confirmation",[3] Carl Hempel admitted that it was "an open question to what extent a satisfactory system of rules [of acceptability] can be formulated in purely logical terms".[4] To our knowledge, the question has never been closed. We propose to do so here, with the proviso that "purely logical terms" be understood so as to include the basic concepts of probability theory. In fact, a single rule suffices: *a hypothesis is acceptable (other things being equal)*[5] *just in case it has passed a severe test*.[6] Our main aim in this paper is to characterize "severe test". It is possible to do so, we will argue, using a generally Bayesian framework of ideas.[7] Within this framework, such traditional criteria of acceptability as confirmation, evidence, and simplicity can be understood in an intuitive way and then incorporated quite naturally in a characterization of severity.

Synthese (2006) 148: 259–293
DOI 10.1007/s11229-004-6222-6

The word “acceptable” is, of course, problematic, for it suggests both that a particular hypothesis can be accepted and that it *ought* to be. Moreover, some philosophers think that to “accept a hypothesis” is further ambiguous, its meaning varying as a function of one’s metaphysics. Thus Bas van Fraassen (1980, 12) contends that for realists, to accept a hypothesis is to believe that it is true, while for empiricists, it is to believe that it is *empirically adequate*. There are even philosophers, Elliott Sober is one of them, who want to give up the notion of “acceptance”, and with it “acceptability”, altogether (Sober 1993).

Sober’s argument derives from reflection on the lottery paradox.[8] Suppose a fair lottery with a thousand tickets. Exactly one ticket will win and, since the lottery is fair, each stands an equal chance of doing so. Consider the hypothesis, “ticket #1 will not win”. This hypothesis has a probability of 0.999. Therefore, we have good reason to believe, and in this sense “accept”, the hypothesis. But the same line of reasoning applies to all of the other tickets. In which case, we should never accept the hypothesis that any one of them will win. But we know, given our initial supposition, that one of them will win. This paradoxical result is to be avoided, according to Sober, by denying that we ever “accept” a hypothesis.

Sober uses the lottery paradox to argue for a wholesale rejection of the notion of acceptance. But, of course, this is not the only or, we might add, the most plausible option.[9] For one could also modify the notion of “acceptance”. We do so in two ways in what follows. First, to accept a hypothesis is to have a good reason for believing it to be true (or empirically adequate, or whatever). But the converse does not hold. However good our reason for believing it to be true, we might still reject the hypothesis. Second, we might reject the hypothesis because our evidence for it is not, in a sense to be specified more precisely, “significant”. On our view, evidence is insignificant when it fails to distinguish between competing hypotheses. In the lottery case, the likelihoods of all of the competing hypotheses, that is, the probability of cashing a winning ticket on the hypothesis that it is not a winning ticket, are the same. In which case, the evidence for any one of these hypotheses is not significant. Whatever epistemic attitude is involved in our “accepting” a hypothesis, it has to be something *stronger* than merely having a good reason to believe that it is true.

For us, a hypothesis is acceptable just in case it has passed a severe test, and a severe test in turn is characterized in terms of belief and evidence. We think that this understanding of “acceptable” is neutral as between realists and empiricists. Philosophers who have styled themselves as one or the other have jointly agreed on the importance of severe tests in the history of science, and we do not intend here to favor either. Furthermore, we want to take "acceptable" as primarily normative, that is, as qualifying hypotheses that *should* be accepted. At this point, we follow the lead of John Stuart Mill[10] in thinking that the best evidence for what is "acceptable" is what in fact has been accepted, not by individuals at particular times, but by relatively stable scientific communities. The task then is to determine why.

Newton's laws may be taken as paradigms of "accepted" hypotheses. In large perspective, their acceptance turned more than anything else on the severity of the tests they passed. One test, in particular, was crucial. The English astronomer Edmond Halley first observed a large comet in 1682. Going back through the records, he found reports of comets precise enough to compare to his own. The orbits recorded for two of these, one in 1606–1607 and the other in 1530–1531, were very close to that which he calculated in 1682. Arguing that it was very unlikely that three comets should have such similar orbits, he concluded that three appearances of a single comet had been observed. Then, using data from what he took to be three appearances of a single comet, together with the hypothesis that Newton's laws applied to the phenomenon, he predicted that the comet would appear again in December, 1758. The comet reappeared as predicted on Christmas Day, 1758, fifteen years after Halley's death, and was promptly named in his honor. Generally taken as "severe", this test more than any other led to the widespread acceptance of Newton's theory, particularly on the continent of Europe.[11]

The point has, of course, been very much emphasized by Karl Popper. The degree to which Halley's prediction confirmed[12] the applicability of Newton's laws depended on the high probability that the prediction was false. Everything known at the time, with the single exception of these laws, made it highly likely that a comet would not appear within a thirty-day period fifty-three years after Halley first made the prediction. It was this fact that made the prediction a significant or "severe" test.[13] Hypotheses with improbable consequences are more highly confirmed if these consequences are not falsified.

It is easy to show that of two consequences of a hypothesis, the more improbable (or unexpected) confirms it more strongly. This follows trivially from Bayes Theorem, that the posterior probability of an event is directly proportional to its prior probability multiplied by its likelihood and inversely proportional to the probability of the evidence. We could put this by saying that the more improbable the consequence, the "more severe" the test, and hence the greater the degree of hypothesis confirmation if the consequence is observed.[14] But it is also clear that one test could nominally be "more severe" than another without either test being at all, in a straightforward and intuitive sense, severe. That is, one hypothetical consequence could be more improbable than another, neither being very improbable in the first place. In such cases, we would not want to say that the evidence for the hypothesis was very strong. Moreover, as we shall see, although a high degree of confirmation is a necessary condition on severity, it is not also sufficient.

Popper enjoins us to look for more and more severe tests. Well and good. But in our view this search should be motivated by the frank recognition that many of our present tests, however much they improve on available alternatives, are not in themselves severe. As we shall see shortly, this is especially important in the medical sciences, where advice based on inadequate (but the "best available") tests is often misleading. Only when severity itself has been characterized can it serve as a research goal, and a prod to redouble our efforts. The task, then, is to characterize a more methodologically useful, "absolute" notion.

Recently, Deborah Mayo has developed in complex detail an account of severity based on what she calls the "error-statistical" approach.[15] It differs in several important respects from the Bayesian account. After outlining both accounts, we compare them with respect to some standard diagnostic tests in the medical sciences, show why the Bayesian position gives generally more acceptable results, and provide reasons why this should be so.

Thus in Section 1, we propose a Bayesian account of a severe test, in Section 2, illustrate this account with respect to diagnostic testing,[16] in Section 3, set out Mayo's very different error-statistical account, in Section 4, provide a quantitative comparison of the two accounts, in Section 5 explain why the Bayesian account is to be preferred, and in Section 6 consider an objection to our version of it. In Section 7, we will discuss variants on the Bayesian position and show how our account combines their best features. Section 8 consists of our conclusions, among which is that careful distinctions must be drawn between the types of questions accounts of confirmation, evidence, and severity are meant to answer.

1. A BAYESIAN ACCOUNT

What is Bayesianism? It is at once a theoretical perspective that accords special importance to Bayes Theorem in the confirmation and acceptance of theories, a particular interpretation of that Theorem, and a description of how rational agents both do and should change their beliefs on the basis of accumulating data.

Bayes Theorem is easily derived from the axioms of probability theory. It states that the posterior probability of a hypothesis conditional on data, Prob(H|D), is equal to the prior probability of the hypothesis, Prob(H), multiplied by the likelihood of the data, Prob(D|H), and divided by the degree to which the data were to be expected when H does not hold, Prob(D|not-H). Bayesianism is, first, the view that confirmation is a function of posterior probabilities; to confirm a theory is to raise its posterior probability.

But, second, Bayesianism is an interpretation of Bayes Theorem. By this we mean that the prior probability of a theory is taken as a measure of the degree of belief of a rational agent in the theory prior to inaugurating tests of it.[17] Two cases should be distinguished. In what is sometimes called the "standard case", the degree of belief in a theory is simply a function of the relative frequency with which the hypothesis (that someone has HIV, for example) is true within the specified population. In the so-called "non-standard" case, the degree of belief in the prior probability of a theory is a function of the agent's background information, past experience, and general expectations, among other things. It is characteristic of Bayesianism to use Bayes Theorem even in "non-standard" cases, where inevitably a certain amount of "subjectivity" enters into the estimation of prior probabilities.

Finally, Bayesianism incorporates the following picture. Rational agents begin with a probability distribution over possible hypotheses. In "standard" cases, this distribution is based on relative frequencies, in "non-standard" cases on the agent's informal expectations and, we will add, the relative simplicity of the hypotheses. As data accumulate, these probabilities are readjusted using Bayes Theorem. Over time, the beliefs of rational agents will, subject to various conditions, converge, however various their points of departure. It is simply a question of progressive conditionalizations on the evidence. This picture is descriptive of the way in which rational agents actually behave and normative with respect to the testing of scientific theories.

For the Bayesian, then, the testing of a theory involves both the invocation of prior probabilities (even in "non-standard" cases) and application of the notion of likelihood, that is, the probability of the datum given the hypothesis. Once the priors are in hand, posterior probabilities can be calculated on the basis of likelihoods.

We are now in position to provide a Bayesian characterization of the notion of a severe test for a hypothesis. On our account, T counts as a *severe test* for a hypothesis H just in case T yields data D such that the following two conditions are met:[18]

(i) Prob(H|D) is high, i.e., the posterior probability of H is high;

(ii) The ratio of the likelihoods, Prob(D|H)/Prob(D|H′), where H and H′ are competing hypotheses, is high.[19]

Note that in (ii), the value of the ratio could range between 0 and ∞ exclusive, whereas in (i) the value could only range between 0 and 1.[20] Note also that the expression "high" in "high probability" and "high ratio value" is context-sensitive. That is, what counts as "high" will vary with the context, as is, in fact, the customary practice of scientists. From one context to the next there will usually be a consensus as to what numbers are to be taken as thresholds. This sort of "subjectivity" in science, to be distinguished from the apparent subjectivity involved in the determination of prior probabilities, is inevitable.

We call the first condition in our characterization of severity the Confirmation Condition (CC) and the second condition, which utilizes the so-called Bayes Factor, the ratio of two likelihoods, the Evidential Significance Condition (ESC). CC is motivated by the consideration that the passage of a severe test must result in a high degree of belief in the hypothesis being tested. ESC is motivated by the consideration that the hypothesis that does the better job of supporting the data is the one "that did the better job of predicting what happened".[21] It rests on the well-known Likelihood Principle (LP). According to the LP, D1 provides as much evidence as D2 does for H1 against H2 if and only if Prob(D1|H1)/Prob(D2|H2) = 1, It is important to note that both "D1" and "D2" stand for actual, observed data and not for possible, unobserved data. The Confirmation Condition implies that if these data are themselves unexpected, we have that much more reason to believe the hypothesis.[22] Both conditions are in an obvious way "Bayesian".

Before proceeding to illustrate our Bayesian account of severity, we might note three points.

First, both conditions are comparative, the first implicitly, the second explicitly, in the sense that they assume a comparison between hypotheses. In the case of CC, Prob(H|D) is high just in case every other hypothesis incompatible with H has a low probability on the same data. But in this respect, the two conditions accord well with the notion of severity which, while itself not comparative, typically presupposes a contrast between hypotheses, in the base case a contrast between a hypothesis and its negation.[23] In this same respect, a severe test is like a "crucial experiment", although "crucial experiments" have often, and unlike our characterization of severity, been understood in deductive rather than probabilistic terms, one of the two hypotheses conclusively "falsified".[24]

Second, the two conditions have built into them a very intuitive notion of simplicity.[25] For in "non-standard" cases, where the prior probability of a hypothesis has not already been determined as a relative frequency, it is very natural to hold that the prior probability measures, among other things, the simplicity of a hypothesis. A hypothesis gets a higher probability than its contenders, other things being equal, if it has fewer adjustable parameters. The likelihood function, on the other hand, measures goodness of fit to the data. A hypothesis with more parameters generally has a higher likelihood than one with fewer parameters; as the order of a polynomial model increases, so too does its maximized likelihood. But the posterior probability of a hypothesis is a function of its prior probability and its likelihood. In requiring that one result of passing a severe test is that the posterior probability of a hypothesis be high, we are at the same time requiring that it trade likelihood off with simplicity in an optimal way.

Third, as will be clear in what follows, our conditions of severity are independent; satisfaction of one does not entail satisfaction of the other.

2. ILLUSTRATIONS OF THE BAYESIAN ACCOUNT

We provide four examples based on diagnostic tests to show how the account works in practice. The first example is called the Pap smear case (PAP), the second one, the HIV case (HIV), the third, the tuberculosis case (TB), and the fourth, the malaria case (MAL).

2.1. *PAP and the Absence of Severity*

Consider a case where we are interested in two mutually exclusive and exhaustive states of health. H is the hypothesis that an individual has a particular disease, and not-H is the hypothesis that she does not have the disease. Let D represent a positive screening test result. We would like to find Prob(H|D), the probability that a person with a positive test result does actually have the disease.

Cervical cancer is a disease for which the chance of containment is high if it is detected early. The Pap smear is a widely used screening technique that can detect a cancer that is as yet asymptomatic. An on-site proficiency test conducted in 1972, 1973, and 1978, assessed the competency of technicians who scan Pap smear slides for abnormalities.[26]

Overall, 16.25% of the tests performed on the women with cancer resulted in false negative outcomes (not-D). A false negative occurs when the test of a woman who has cancer of the cervix incorrectly indicates that she does not. Therefore, in this study, Prob(not-D|H) = 0.1625. The other

100 − 16.25 = 83.75% of the women who had cervical cancer did in fact test positive; as a result, Prob(D|H) = 0.8375.

Not all of the women who were tested actually suffered from cervical cancer. In fact, 18.64% of the tests were false positive outcomes. This implies that Prob(D|not-H) = 0.1864. The probability that the test results will be negative given that the individual tested does not have the disease is Prob(not-D|not-H) = 1 − 0.1864, or 0.8136.

We would like to know Prob(H|D), i.e., the posterior probability that a person with a positive test result actually does have the disease. To apply Bayes Theorem, we need to know the prior probability of H. Prob(H) is the probability that a woman suffers from cervical cancer when randomly selected from the population.

In this case, the prior probability is simply a measure of the relative frequency of cervical cancer. One source reports that the rate of cases of cervical cancer among women studied in 1983–1984 was 8.3 per 100,000. That is, the data yield Prob(H) = 0.000083. Then Prob(not-H) = 0.999917. Using the likelihoods given in the last two paragraphs and applying Bayes Theorem, we get Prob(H|D) = 0.000373. So, Prob(H|D) is very low; CC is violated. Hence on our account, the available test does not provide a severe test of H.[27] Does D provide good evidence for H? Consider ESC, our evidential significance condition, in this case Prob(D|H)/Prob(D|not-H). Since Prob(D|H) = 0.8375 and Prob(D|not-H) = 0.1864, their ratio is 4.49.[28] What does this value mean? It means that after we know that the individual has a positive result, then her chances of having cervical cancer have increased almost fivefold. While this might seem ominous, this is in fact not a very large increase. It assuredly justifies a woman who receives a positive Pap result in taking more tests to make sure about her present state of health. However, it does not justify her in believing that she does have the disease, nor does it provide much evidence that she does. The Pap smear is simply not a severe test for the presence of cervical cancer. Neither of our conditions is satisfied.

2.2. *HIV and Reasons for Believing*

Consider a case where an HIV specialist wants to know whether a prostitute chosen randomly in a police raid is afflicted with the virus. He administers the only available test. H is the hypothesis that the prostitute in question has the virus. Given his expertise in the field (and without firm relative frequencies to guide him), he has assigned 0.99 in advance to the hypothesis that the prostitute is carrying the virus. D represents a positive test outcome; not-D means that the test turns out negative. The specialist wants to know what the probability is that the prostitute is afflicted with the virus given that the test says she is, i.e. what is Prob(H|D)? But he also wants to know whether the test is severe.

To compute this, he needs to know the likelihoods of the data under H and not-H respectively. We might suppose that the following are the likelihoods derived from his past experience with the test.[29] Prob(D|H) = 0.1625 (i.e., true positive case), Prob (not-D|H) = 0.8375 (false negative case), Prob(D|not-H) = 0.8136 (false positive case), and Prob(not-D|not-H) = 0.1864 (true negative case). These likelihoods at once suggest that the test is not severe. We compute Prob(H|D) = 0.94, which in fact is very high, in large part a result of the value assigned by the specialist to the prior. So CC is satisfied.[30] He has reason to believe on the basis of the evidence that the prostitute has HIV. However, the test at stake does not count as a severe test of H. We find that Prob(D|H)/Prob(D|not-H) = 0.199, which is less than 1. Hence, the ratio provides evidence against H in favor of not-H. ESC is violated.

2.3. *TB and Evidential Significance*

Among 1,820 subjects in a study, 30 suffered from tuberculosis and 1,790 did not. Chest X-rays were given to all of the individuals. 73 had a positive X-ray, indicating that there was evidence of inflammatory disease, whereas the results of the other 1,747 were negative.[31]

Let H represent the hypothesis that an individual is suffering from tuberculosis and not-H the hypothesis that she is not. These two hypotheses are mutually exclusive and jointly exhaustive. Assume that D represents a positive X-ray. We would like to determine whether the test which yields it constitutes a severe test of the hypothesis that a given individual actually has the disease. As always, we need to know Prob(H), Prob(not-H), Prob (D|H), Prob (D|not-H), Prob(not-D|H), and Prob(not-D|not-H).

Prob(H) is the prior probability that an individual in the general population has tuberculosis. In a 1987 survey,[32] there were 9.3 cases of tuberculosis per 100,000 population. Prob(H) = 0.000093. Hence, Prob(not-H) = 0.999907. Prob(D|H) is the probability of a positive X-ray given that an individual has tuberculosis. Prob(D|H) = 0.7333. Prob(D|not-H), the probability of a positive X-ray given that a person does not have tuberculosis, is 1 − Prob(not-D|not-H) = 1 − 0.9715. = 0.0285. Using all of this information, we compute Prob(H|D) = 0.00239. For every 100,000 positive X-rays, only 239 signal true cases of tuberculosis. The posterior probability is very low. Hence, CC is violated. Since CC is violated, we conclude that a chest X-ray, T, should not be counted as a severe test of H.

But there is more to this case. Although 99,761/100,000 individuals with a positive X-ray do not actually have the disease, we have greatly increased the chances of properly diagnosing TB. For the likelihood ratio in this case is 25.72, the probability that an individual with a positive X-ray has TB is 25.7 times greater than the probability that a person randomly selected from the population has the disease. So D is evidentially significant for H; to this extent, the insurance company may charge her a higher premium. However, satisfaction of ESC should not be taken as an indication that the person is really suffering from tuberculosis, although it does put her in a higher risk group. In fact, our Bayesian account captures this intuition; D is evidence for H, but the test that produces it is not a severe test of H.

It is worth emphasizing the point. Since Prob(H|D) is 0.00239, the agent's degree of belief that the patient does not have TB, Prob(not-H|D), is 1 −0.00239, or 0.9976. Yet at the same time the evidence very much points to the presence of the disease, since Prob(D|H)/Prob(D|not-H) is high. In this sort of case as in others, degree of belief or confirmation must be very sharply separated from evidence.

2.4. *MAL and Severity*

On our Bayesian account, to count T as a severe test of H entails that both CC and ESC are met. Consider an example where a general physician is confronted with a case in which a patient feels feverish and shows symptoms of malaria.[33] However, no prior case of malaria has been reported in the locality or elsewhere. So if a person is chosen randomly, the probability that the person is afflicted with malaria must be very low. H is the hypothesis that the person has been afflicted with malaria. The doctor's prior probability for that person having malaria is assumed to be 0.09. Hence, Prob(not-H) = 0.91. However, she has run some blood tests on the patient to make sure if he really has malaria. To her surprise, she has found out that the test is positive. She knows that the test is very reliable.

Let D represent a positive screening test result and assume that Prob(D|H) = 0.95. Not all the people who are tested suffer from malaria. In fact, 1% of the tests are false positive outcomes. So Prob(D|not-H) = 0.01. Now what's the probability that the individual in question has malaria given that the test is positive? To answer that, we have to compute Prob(H|D) = 0.98, which is close to 1. This is a very high posterior probability. Hence, D should be regarded as confirming H. Should the test which yields D also be regarded as a severe test of H? To check whether ESC is satisfied, we compute the ratio of Prob(D|H)/Prob(D|not-H), which is 950. This means that the probability an individual with a positive X-ray has malaria is 950 times greater than the probability that an individual randomly selected from the population does. ESC is satisfied because the ratio is significantly high. Since both CC and ESC are met, then according to our Bayesian account, the blood screening carried out should be regarded as a severe test of H.

3. THE ERROR-STATISTICAL ACCOUNT

There are a variety of objections to Bayesianism. One is that the use of prior probabilities in some diagnostic cases, for example in the HIV and MAL cases, depends on the experience and expectations of the diagnostician, and for that very reason introduces an element of "subjectivity" into the testing process. For different clinicians could have different initial degrees of belief in the same hypothesis. Since the goal of science is to test and eventually confirm or reject hypotheses in an entirely clinician- and belief-invariant way, some philosophers and statisticians have developed alternatives to the Bayesian approach. One such alternative has been worked out in interesting detail by Deborah Mayo.[34] It is known as the error-statistical approach. This alternative way of characterizing severity relies on the notion of error probabilities central to Neyman-Pearson statistics. Of interest here is the way in which Mayo has elaborated it as an account of experimental testing. The basic idea is that a test is severe just in case it would detect an error in the hypothesis (i.e., the chances of its doing so are overwhelming), if in fact an error were present.

There is little point in denying that Bayesian and error-statistical approaches have rather different aims in general,[35] as well as widely varying presuppositions.[36] The former seek to provide an account of the relation between evidence and hypothesis when the one confirms the other, raising or lowering the degree of one's belief in it. The latter offer methods whereby the reliability of statistical inferences can be checked. Indeed, error statistics does not really offer a "theory" of anything, certainly not of the ways in which, ideally, beliefs should change as we accumulate evidence for them, nor is it an "inductive logic" (in the usual sense of the words), perhaps two reasons why philosophers, interested in how premises support conclusions, have not taken it more seriously.

Difference of aim is important. So too is level of abstraction. Bayesian approaches are very general. In a typical case, only mutually exclusive and jointly exhaustive hypotheses are considered, as if there were no further alternatives. Statistical approaches, on the other hand, take these alternative hypotheses into consideration. Mayo likes to say, with some justification, that error-statistics is much closer to the "nitty-gritty" of actual scientific practice, where it is often not so much that a woman has breast cancer or not, as it is whether she has a precancerous condition or a benign tumor or . . . or . . .

As a final preliminary, there is no point in denying, as some Bayesians do, that the methods provided by classical statistical approaches are not widely and successfully used across the spectrum of sciences,[37] or that a proper appreciation of them must inform our philosophical understanding of data collection and experimental testing. A blanket charge that these methods are of little help or should not be trusted flies in the face of practice.

At the same time, we cannot leave matters there. Nor does Mayo. There is a suggestion in her writings (Mayo 1997b, S198–99) that, in light of the above considerations, there are two concepts of severity, the Bayesian and the error-statistical, making comparison between them difficult if not impossible.[38]

In the opening pages of *A Theory of Justice*, Rawls (1971) draws a distinction between the concept of justice, on which there is general agreement, and various *conceptions* of justice, which are in dispute. Although he is not very clear on how this distinction is to be drawn, Rawls seems to think that a concept is more general than a conception, the latter making explicit what is no more than implicit in the former. Presumably there would have to be agreement on the analysis of the concept in the first place if rival conceptions were to be possible, an agreement, for example, concerning the set of roughly analytic truths associated with it. One could agree, for example, that justice is rendering every man his due; this would be an analytic truth about justice from which one could proceed. Varying conceptions of justice would then have to do with spelling out what is or should be meant by this phrase, in particular with what is meant by one's "due". Do we have two concepts or two conceptions of severity, and if the latter, what is the analysis of the concept shared in common?

The immense prestige of Carnap is behind the claim that there are two concepts of probability, statistical and inductive.[39] His case rests, at least in part, on a difference in aims and applications: inductive probability measures the degree of support given to a hypothesis by the evidence for it, statistical probability has to do with features of the world, relative frequencies for example. But even if he is right to distinguish two concepts of probability, it does not follow, or so we would argue, that they cannot be used to develop rival conceptions of severity.

First, there is surely an intuitive, largely pre-theoretical understanding of the expression "severe test". Severe tests – of scientific hypotheses, of the durability of automobiles, of one's manhood – have this in common, that hypotheses, automobiles, and men must be tough to weather them.

The varying conceptions of severity have to do, then, with what we mean by "tough" and "weather", and how to determine and measure them in the various contexts in which they are at stake.[40]

Second, Mayo herself provides an account of severity which, she implies, is not simply different from but superior to the Bayesian account, which implication presupposes their comparability. She is ready (as the discussion of the exchange between her and Colin Howson in section 8 of this paper clearly shows) to compare assessments of severity with respect to particular hypotheses and experimental outcomes.

Third, even if they were, at least in some sense, incomparable, practitioners would still be in a position to decide which of the two was in fact the more useful, by their application to specific cases and the generation of some numbers. In this connection, there is something misleading about Mayo's contrast between a Bayesian's presumptive use of severity as an "accept or reject" rule and the error-statistician's reluctance to do so. For our own Bayesian account of severity has more to do, as will be seen eventually, with the adequacy of our tests than it does with our readiness to accept or reject particular hypotheses based on them. Particularly in the sorts of diagnostic examples we have discussed, the primary question is how conclusive the test is. If it is not severe, on our characterization, then efforts to find a better test should be redoubled.

One last point before we turn to the details of Mayo's account of severity. It has to do with the abstract character of Bayesian analysis. In our view, such abstraction, and consequent idealization, is characteristic of scientific theories generally. At least up to a point, in fact, the simpler the theory, the greater its range, the more powerful the explanations that it affords. As our diagnostic cases indicate, what we want to explain is why clinicians should feel the way they do about, e.g., the PAP test. Where there is an incompatibility between our assessment and that which experts usually make, the theory directs us to look for factors so far unconsidered which might explain the incompatibility. It seems to us that in her desire to be as close to the details of scientific practice as possible, Mayo runs the risk of draining her views of explanatory power.

At last to the error-statistical account of severity. Mayo begins by characterizing strong evidence in terms of severity. "Passing a test (with D) counts as a good test of or good evidence for H just to the extent that H fits D and T is a severe test of H". def4041[40] She terms this the severity requirement (SR). First she characterizes fitness: D "fits" H only if D is not improbable on H. [42] Then she proceeds to give criteria of severity (SC). The first of these is SC1a: "There is a very high probability that test procedure T would not yield such a passing result, if H is false". In other words, Prob(not-D|not-H) must be very high.[43] The second criterion, SC1b, says: "There is a very low probability that test procedure T would yield such a passing result, if H is false".[44] In other words, Prob(D|not-H) is very low. Since SC1b is just 1 − SC1a, we can derive the former from the latter. So she does not really need SC1b as a further condition in her account of severity. Consider H and its denial not-H to be mutually exclusive and jointly exhaustive of all possible hypotheses in a domain.[45] Assume that H is a simple hypothesis and that D is data for H. For Mayo, T is a severe test of H just in case it has outcome D and

(i) Prob(D|H) is not very low,
(ii) Prob(D|not-H) is very low, and
(iii) Prob(not-D|not-H) is very high.

The first condition provides a gloss on "fitness" in her Severity Requirement. The second two, as we have already seen, define her Severity Criterion, although as we have shown they are inter-derivable. Note that these conditions make the same sort of appeal to the general consensus in a scientific field that is involved when we say that probabilities or ratios are "very high" or "very low", and in this respect Mayo's account does not differ from our own.[46] The main formal difference, rather, is that she makes no appeal, for reasons already given, to the prior probability of H or, *a fortiori*, to its posterior probability. In her view, it is neither possible nor necessary to do so.

4. A COMPARISON BETWEEN THE ERROR-STATISTICAL AND BAYESIAN ACCOUNTS

Consider first what Mayo's severity conditions entail with respect to the PAP case. Recall that on her account, (i) Prob(D|H) is not very low, (ii) Prob(D|not-H) is very low or (equivalently) Prob(not-D|not-H) is very high. In fact, in the PAP case, Prob(D|H) = 0.8375, which is very high, Prob(D|not-H) = 0.1864, which is very low, and Prob(not-D|not-H) = 0.8136, which is also very high. It follows that the PAP smear is a severe test of the hypothesis that a given woman has cervical cancer. On our Bayesian account, to the contrary, D is not evidentially significant for H nor does it provide us with good reasons for believing H; neither of our conditions on severity is satisfied.

In the HIV case, both Mayo and we Bayesians seem to agree that the test characterized in Section 2.2 is not severe. Prob(D|H) = 0.1625, which is very low. Since one of the conditions of Mayo's account is violated, the test fails to pass H severely relative to E. We agree with her. We take the test to provide a good reason for believing that H is true, but no more than that; the test is not severe.

On our reading of the situation, Mayo must hold that the test for TB described in Section 2.3 is severe since both conditions of her account are satisfied. Here, Prob(D|H) = 0.733; Prob(D|not-H) = 0.285 (or alternatively, Prob(not-D|not-H) = 0.9715). However, on our Bayesian account, the test is not severe. CC has not been met. Although D is evidentially significant, it does not provide us with a good reason for believing H.

The fourth and final example involves the case of malaria. Mayo will take the test of H involved to be severe with respect to D because the two conditions of her account are satisfied. Prob(D|H) = 0.95 and Prob(D|not-H) = 0.001 (or Prob(not-D|not-H) =0.999). For the second time we agree with her; our two conditions, CC and ESC, are both satisfied. The test is severe.

The following table summarizes the results reached when the two approaches are applied:[47]

Approaches	*PAP*	*HIV*	*TB*	*MAL*
Bayesianism	No severe test. Both CC & ESC are unsatisfied	No severe test. CC is satisfied, but not ESC	No severe test. CC is unsatisfied, but ESC	Severe test. Both CC & ESC are satisfied.
Error-Statistics	Severe test	No severe test	Severe test	Severe test

5. AN EXPLANATION OF THE DISAGREEMENT BETWEEN THE TWO APPROACHES

Both our Bayesian account and Error-statistics agree that the test at stake in the HIV case is not severe, while in the MAL case it is. In these cases, it does not matter whether the agent adopts one approach rather than the other.

Consider the cases on which we disagree. With respect to the PAP and TB tests, the error-statistical and Bayesian accounts provide different assessments. Mayo must think that the Pap smear is a severe test of the presence of cervical cancer; we don't. Similarly, she must think that the TB test described above is a strong indicator that an individual has tuberculosis; we do not. It is true that the probability of a positive outcome from a Pap smear given that the person suffers from cancer is reasonably high. It is also the case that there is a high probability that a person who has the disease will test positive for TB. But in neither case should it follow that the test is severe.

The likelihood of the data given the presence of cancer or the likelihood of the data when an individual has tuberculosis is not adequate to determine the posterior probabilities of the hypotheses in question. We must also take into account information about the probability of an individual being afflicted with either disease when selected randomly from the population. In the PAP case, this prior probability of an individual's being afflicted is just 8.3 per 100,000 population, whereas in the TB case it is only slightly higher, 9.3 per 100,000. In both cases, the probabilities are extremely low. At the same time, they play a crucial role not simply in determining the posterior probability of an individual's having a disease, but also in saving us from generally counter-intuitive results.[48]

Our insistence on the importance of prior probabilities in determining severity might seem little more than question-begging were it not also the case that our assessments match those of clinicians working in the field.[49] Thus in a very interesting article in the September 13, 1999, issue of *The New Yorker* by the Recanati Professor of Medicine at Harvard, Jerome Groopman, the deficiencies in the Pap test are carefully pointed out. It is generally acknowledged to be misleading ("notoriously inaccurate" is Groopman's phrase) and should be used with great caution. That it is not a severe test has prompted a rather intensive search for alternatives, at least one of which is about to enter practice. Even then, scientists should not rest content with a better ("more severe") test, but seek tests that are in and of themselves severe. At the same time, while a positive result on a Pap smear does not justify a patient in believing that she has cervical cancer, on our account it does have some evidential significance. It should prompt a careful individual to seek more and better tests, particularly since the utilities involved in taking further action when cervical cancer is suspected are so high.[50]

We believe that Mayo has dismissed the importance of prior probabilities because she is worried about the way in which they import "subjective" considerations into science. But this is to throw a clearly helpful baby out with dubiously dirty bathwater, and will inevitably lead her, at least on occasion, to draw misleading conclusions. This is one of the points made by De Finetti in the passage quoted at the outset.

Two further comments need to be made in this connection. One is that with respect to the two cases on which we disagree with Mayo, the prior probabilities are determined by relative frequencies. There is here no trace of the dreaded “subjectivity”.[51]

The other comment is that some subjectivity in the assessment of severity seems necessary. We have already remarked that Mayo’s account and our own appeal to “very high” and “very low” probabilities. But “height” and “depth” are context-sensitive terms; in diagnostic cases such as the ones we have discussed they are relative to the judgment of informed clinicians. This is why, when the result of a test comes back from the lab, it is often “interpreted” by the doctor, e.g., “this is a good result”, as almost anyone who has had a biopsy can attest.

But we also need to put our disagreement with Mayo in a larger perspective. The biostatistician Richard Royall (1997, 2000, 2001) provides it. Our diagnostic examples are concerned with patients whose test results come out positive. Given positive results, Royall suggests that, depending on which statistical school one belongs to, one asks a different type of question.

1. Given the datum, what should you do?
2. Given the datum, what should you believe?
3. Given the datum, what does it say regarding the *evidence* for *H* against its counterparts?

In Royall’s view, classical error-statisticians ask question 1, Bayesians ask 2, and likelihoodists (who take the Likelihood Principle as fundamental) ask 3.[52]

According to Royall, error-statistics is decision-theoretic in character. It has to do with computing error probabilities that are either tied to the data and particular experimental set-ups or generated by human intervention in these set-ups. Classical Neyman-Pearson statistics neatly summarizes these two types of errors as Type I and Type II.

Suppose the clinician is interested in knowing the effect of the drug AZT among patients suffering from HIV after it has been administered. To learn the effect of the drug, and to understand how the situation is related to the two types of errors, consider two hypotheses: Ha and Ho. Ho is the null hypothesis; it is that there is no difference in patients before and after treatment. Ha is the alternative hypothesis that there is a difference. Type I errors occurs *when the decision is to reject the null hypothesis*, when it is actually true. A Type II error occurs *when the decision is to not reject the null hypothesis*, when it is actually false. The task of error-statistics is thus twofold: to reject the null hypothesis as a function of some pre-assigned significance level and to minimize these two errors.

Now whether or not Mayo would accept this characterization of error-statistics as decision-theoretic, it is true that her position turns in part on the utilities involved in acting on various hypotheses, as we will show in the next section of this paper. Our own view is that questions concerning *evidence* should be separated from the utilities in acting on the hypotheses which such evidence supports. Apart from this, and more generally, on any plausible account of (genuine) "evidence", evidence should always be able to discriminate between incompatible hypotheses.[53] The point may be illustrated briefly in connection with the Raven Paradox.[54] Since "All ravens are black" and "All ravens are not black" are, given the existence of a raven, contraries and hence incompatible (since both cannot be true). The difficulty on Hempel's account of confirmation[55] is that such non-raven non-black items as white shoes are instances of, hence evidence for, each, a result which at the very least seems counter-intuitive. If we were to maintain that *genuine* evidence must be able to discriminate between contrary hypotheses, that is to say, cannot be evidence for both, then we could reject white shoes as evidence for the hypothesis that all ravens are black, and with them the paradox. Our use of the Likelihood Principle is intended to incorporate this intuition. The likelihood of a white shoe on the hypothesis that all ravens are black is the same as the likelihood of a white shoe on the hypothesis that all ravens are not black, hence a white shoe does not provide genuine evidence for either of them.

The basic difficulty, in our view, is that one cannot define (genuine) evidence in terms of the passage of (severe) tests, as Mayo does.[56] As we have just argued, any reasonable conception of "evidence" must make use of something like LP; this we have done in our ESC. But our account of severity, like Mayo's, rejects LP. That is to say, statisticians must ask a *fourth* question, when is a test *severe*? and this question must be distinguished from Royall's third, what does the evidence tell us about the hypotheses one is considering?

Mayo thinks that since we are Bayesians we must accept LP across the board, so to speak. But this is a mistake. We need only accept it where it is plausible, as a measure of evidentiary significance. Indeed, although Savage (1962) claims that LP "flows from" Bayes Theorem and certain subjectivist assumptions, contemporary Bayesians have come more and more to question it.[57] In a similar sort of way, Royall thinks that since we are Bayesians we must restrict our attention to the "belief question". But this, too, is a mistake. We can, and have, in a way that is compatible with his own, given an answer to the "evidence question" (and shown, in the tuberculosis case, why the two questions must be distinguished). Moreover, we can answer the "severity question", in terms of confirmation and evidence,[58] in a way that provides a more plausible account of sample diagnostic cases, among others, than does Mayo's. There are simply no *a* *priori* grounds on which to rule out a Bayesian account of either evidence or severity.

Three more points concerning our differences with Mayo might be made. All of them center around the fact that we accept, while she rejects, the Likelihood Principle, so far as our account of "evidence" is concerned.

The first point is that the LP allows us to compare hypotheses with respect to the evidence for and against them. We have already argued that the concept of severity is at least implicitly comparative in the sense that severe tests are intended to adjudicate between, perhaps also among, hypotheses. Mayo seems to think, wrongly in our view, that severity does not always presuppose an implicit comparison between hypotheses, although such a comparison does not preclude saying of a particular test that it is or is not, *tout court*, "severe".

The second point is that our differences with respect to the LP have at least in large part to do with the fact that for Mayo "evidence" has to do not only with what is or has been the case but with what might have been the case as well, that is, has to do with possible as well as actual outcomes. So much is required by Neymann–Pearson theory. We think that it is difficult to argue on *a priori* grounds for what Sober (1993, 42–43) terms the "principle of actualism", that we should (always) form our judgments about hypotheses based on the evidence we actually possess, without begging at least some questions.[59] Analysis of the notion of "evidence" will not itself settle the issue, and there are circumstances in which testing long sequences of hypotheses, and hence an eventual appeal to possible cases, seems desirable. But such testing does not very well fit paradigm cases in the history of science such as Halley's prediction, and even in the sorts of medical diagnostic cases discussed, practitioners will have *per force* to limit themselves to actual rather than possible evidence.[60]

The third point is that there are very general reasons the Likelihood Principle cannot simply be rejected, as Mayo does, out of hand. One of these reasons has to do with the notion of a *sufficient* statistic. In simple terms, a statistic is sufficient if it contains all of the information in the data that is needed to make an inference. This statistic is considered to be a desirable property for a good estimator, and as such is embraced by both classical and Bayesian statisticians. But there is a close tie between the notion of a sufficient statistic and acceptance of the LP. Hence one is ill-advised to simply reject the latter.

Suppose the uncertain quantity of interest is Θ. In addition, assume that sample information concerning Θ can be summarized by the sample statistic y. If y contains all of the information from the sample that is pertinent to the uncertainty about Θ, then y is called a sufficient statistic. To consider a schematic example, suppose that the investigator is interested in making an inference about the population proportion. She needs to know the number of successes, *s*, and the number of failures, *f*, in the sample, so that she can compute the posterior probability of the parameter. We say that *s* and *f* are together sufficient because only their values are needed to find the posterior distribution. If the investigator were presented with a breakdown of the data, she would have some additional information, for example, concerning the order in which the successes and failures occurred. The original data tell us the number of successes, the number of failures, and the order in which they occurred, but to calculate the population proportion we do not need all of that information (and still less information concerning the order in which they *might* have occurred). Only the numbers of *s*'s and *f*'s are required; they provide a sufficient statistic.

6. AN OBJECTION TO OUR ACCOUNT OF EVIDENCE

Another objection to our use of the Bayes Factor to characterize the significance of evidence has been advanced by Teddy Seidenfeld.[61] It takes the form of a counter-example.

First, we assume the apparently plausible principle that if two hypotheses are logically equivalent with respect to a particular piece of evidence, then (intuitively) that piece of evidence will have the same significance or "weight" for both. To say that two hypotheses are "logically equivalent with respect to a piece of evidence" in this sense means simply that with respect to that evidence, one hypothesis is true/false if and only if the other hypothesis is true/false.

Thus consider two hypotheses, that two flips of a coin will both come up heads and that the second flip will come up heads. Now suppose that the first flip comes up heads. Then the first hypothesis, that both flips will come up heads, is true if and only if the second hypothesis, that the second flip comes up heads, is true.

Second, on our account the significance or "weight" of the evidence is to be understood in terms of the Bayes Factor. If data have the same evidential significance for two hypotheses H and H′, then prob(D|H)/prob(D|H′) = 1.

In the sort of case sketched, where two hypotheses are "logically equivalent" with respect to a particular piece of evidence, prob(D|H)/prob(D|H′) = 1. But the simple coin-flipping case sketched gives us a likelihood of D given H of 1 and a likelihood of D given H′ of 1. Since on the principle at stake in the first premise, the evidence should have the same significance for both hypotheses, and since on our account it does not, our account should be rejected.[62]

It is difficult to know what to say in reply to this objection, however ingenious it is, for the notion of "logical equivalence" that Seidenfeld employs is problematic. On the more usual notion, on which logically equivalent hypotheses have all and only the same entailments, the Evidential Significance Condition will not be violated. In fact, the two hypotheses mentioned are not in this sense logically equivalent; H entails H', but the converse does not hold. This fact, in turn, suggests a more general conclusion: severe tests cannot be applied to pairs of hypotheses when one of them entails the other. This is, of course, a condition of employing a Bayesian methodology. But it is not simply *ad hoc*. For the notion of a "severe test" is applied to competing hypotheses (even, as earlier indicated, when one of the hypotheses is at best implicit), and at the very least "competing hypotheses" should issue in incompatible predictions.

Working scientists don't consider two theories as serious contenders when one is nested inside the other, although such a possibility could be entertained as a logico-mathematical possibility. Rather, physical considerations play the crucial role. A case in point is Kepler's search for his first law of planetary motion. Kepler began the search by taking the circle as the correct and simplest trajectory of planets around the sun. But, as is famously known, he had trouble reconciling a circular trajectory with Tycho's data for Mars' orbit. Eventually, trading off simplicity with correctness or goodness-of-fit, Kepler hit upon the first law, that planetary orbits are elliptical.

Circles are, of course, ellipses and in this sense "nested" within them. But since the physics of central forces in the case of the hypotheses is radically different, Kepler treated them as though they were incompatible, and did not simply proceed from the one to the other by way of some mathematical ingenuity. This sort of case is typical; whether two hypotheses are incompatible is determined perhaps even more on the basis of physical than on mathematical considerations.

It is also the case that Seidenfeld's counter-example does not invoke a statistical model. By a "statistical model" we mean a model that represents the description of a tentative chance mechanism; an investigator uses this to capture the systematic information in the data. A statistical model differs from other types of models in so far as it describes a statistical mechanism or process in terms of a certain probabilistic structure. Mathematical concepts such as a probability distribution, independence, and identical distribution constitute forms of probabilistic structure.[63] Whether one is a frequentist or a likelihoodist or a Bayesian, one needs to invoke a statistical model to provide a joint probability distribution of two events, for example, two flips both coming up heads and the second flip coming up heads. But Seidenfeld has not given us a joint probability distribution of two events, just a joint prior probability distribution, when the latter could exist independent of a statistical model.[64]

7. ALTERNATIVE BAYESIAN ACCOUNTS

It remains to consider accounts of severity which, although derived from the same Bayesian perspective, differ from our own. One such account is provided by Horwich (1982). On it, very briefly, a test of a hypothesis is severe just in case the event(s) which it predicts are unexpected. "...the virtue of surprising predictions is a special case of the virtue of surviving severe tests, which in turn consists simply in the virtue of successfully predicting improbable facts." [65] This is, of course, just the intuition described at the outset of this paper, emphasized by Popper, and exemplified in the case of Halley's Comet. But it is captured by our account. When D is a consequence of H, as it is in the sorts of cases we have been considering, then on Bayes Theorem, P(H|D) = Prob(H)/Prob(D). Horwich's formal condition on severity is that Prob(D) be low. But if Prob(D) is low, then (assuming that Prob(H) is not small) Prob(H|D) will be high (and conversely), which is just our first condition, CC. Indeed, CC is intended, at least in part, to capture this intuition about testing.

Although CC is a necessary condition of severity, it is not, as against Horwich, also sufficient. Recall the HIV case. In this case, Prob(H|D) is very high. However, Prob(D|H)/Prob(D|not-H) is very low. A specialist has some reason to believe that the prostitute in question has the virus. But the test is not severe. The ESC condition is violated. And indeed, the number of false positives in this case is so high, that intuitively we would place little faith in the test.

Another Bayesian, Geoffrey Hellman, has put forward a rather different account of severity.[66] In his view, although we put forward an "absolute" characterization of severity, a comparative notion is embedded within it. We talk about a hypothesis having a high or low probability of being true. But this can only mean that the probability is high or low relative to not-H or to the prior probability of H. However well-intentioned our aim to supply a non-comparative account of severity, it cannot be done within a Bayesian perspective.[67] One might as well settle for a comparative notion of a test's severity from the outset.

Here is Hellman's comparative account:

T is a more severe test of H relative to K than T′ just in case Prob(D|H and K) > Prob(D′|H and K), where "K" represents the "background knowledge" of the agent and "D" are the data yielded by T.

Except for the inclusion of background knowledge, explicit in Horwich's account as well and implicit in ours, Hellman's criterion resembles our second condition, ESC, and like it rests on the LP. But once again, although the condition is necessary it is not also sufficient. This time, recall the TB case. The probability of an individual with a positive X-ray having tuberculosis is many times greater than the probability that a person randomly selected from the population has the disease. ESC is satisfied. Yet, since for every 100,000 positive X-rays only 239 signal true cases of tuberculosis, we cannot take a positive test as a good reason for believing that the individual actually has TB. The test is, intuitively, not severe, an intuition captured by the failure here of our CC condition.

In sum, Horwich (at least implicitly) adopts CC, Hellman ESC. Each is subject to criticism. It is only by combining both conditions that an adequate account of severity can be developed. There is no doubt that the fundamental idea in Bayesian accounts of confirmation is that positive data raise the probability of their attendant hypotheses, that is, make them more probable. This idea is quite clearly comparative. What we have added to this idea is the notion of thresholds, points at which the probabilities in question are "high" or "low". The determination of these points will vary from context to context and, we have maintained, is relative to particular communities of expert opinion. This sort of relativity has nothing to do with the comparison of probabilities intrinsic to Bayesianism. We have tried to show how, upon the assignment of intuitive "highs" and "lows" in real-world cases, the resulting assessments of severity match those given by clinicians in the field.

Two more points need to be made. One is the reminder that one test may be "more severe" than another, while neither is in fact "severe". The Pap smear is presently the most severe test we have for cervical cancer, and, unfortunately, it is not very severe. A merely comparative account of severity is not adequate, and it gives only very limited methodological advice.

The other point is that actual scientific practice very much involves the invocation of thresholds. Despite the impression sometimes given by Bayesians, scientific progress does not consist simply in raising or lowering probabilities. The fact is that at some point in the confirmation of a specific hypothesis, further experimental or observational support of the same kind does not add new reasons for believing the hypothesis; if it did, lab reports from freshman physics classes around the country would be significant, as further evidence for the truth of, say, Galileo's laws. Typically, there is some point at which a theory has passed or failed to pass a severe test, at which point we typically move on to the development of new tests or other theories, although perhaps never more than provisionally. In providing the ingredients for an account of what a severe test is, our own Bayesian approach sets out a much more plausible picture of theory acceptance than the old "pile up the evidence" image otherwise associated with the adjustment of posterior probabilities.

8. FINAL REMARKS

Before we summarize the results of our discussion, one last point needs to be made. It can perhaps best be made by way of reflection on an argument of Howson (1997a) that in certain ways resembles one of our own against Mayo. He argues that there are cases in which an agent making an inductive inference satisfying Mayo's severity requirement generates counter-intuitive results. He therefore concludes that her severity requirement is faulty.

Recall that Mayo's two conditions are that (i) Prob(D|H) must be very high, and (ii) Prob(D|not-H) must be very low. Howson asks us to consider a case in which D (a sentence describing D) is a consequence of H, and thus Prob(D|H) equals 1, and Prob(D|not-H) equals 0.05. Since her two conditions are met, Mayo must say that D should be taken as good grounds for H. But, Howson continues, let's assume that the prior probability, Prob(H), of the disease is very low, say one is a thousand. Given priors of the rival hypotheses and likelihoods of the data with respect to these hypotheses, the posterior probability of H, Prob(H|D), becomes 0.0196, a very poor indicator of the disease. If an agent were to infer from the positive test to the presence of the disease, she would be wrong almost all of the time.

In her reply to Howson, Mayo makes two points. One is that if Prob(H|D) is very low, then Prob(not-H|D), that the individual does not have the disease, must be very high. It follows, Mayo thinks, that if, as in Howson's example, the incidence of the disease is very low (in the group from which the subject was randomly taken), then no matter what the test result, the posterior probability that an individual does not have the disease will always be high. As she puts it in terms of a parallel real-world example (Mayo 1997a, S209), "In the Bayesian screening, no woman could ever have breast disease indicated by this test" (i.e., by an abnormal mammogram). But, she concludes, an abnormal mammogram could not possibly be a severe test for the absence of breast cancer.

Two points about this argument should be noted. One is that it relies on a faulty premise. It doesn't follow from the fact that the incidence of a disease is very low that no matter what the test result the posterior probability will always be high that the randomly selected individual does not have the disease. If Prob(D|not-H) is sufficiently small, then the posterior probability that the individual does not have the disease can also be small, so long as Prob(H) is not zero.[68]

The other point is that on our analysis (and perhaps also Howson's) a positive result in this sort of case is not a severe test for the absence of the disease quite independent of its rate of incidence or the posterior probabilities involved. Since Prob(D|H) is 1 and Prob(D|not H) is 0.05, the evidential significance of the data yielded for not-H is extremely low. Our second condition is clearly violated. Any woman would be ill-advised to conclude from the fact that she has an abnormal mammogram that she does not have breast cancer.

But there is another, more difficult issue at stake here. Mayo suggests that for Bayesians like Howson and us decisions to accept or reject hypotheses are entailed by successes or failures of individual tests. She writes as follows:

> *Newsweek* (Feb. 24, 1997, p. 56) recently reported that only 2.5% of women in their 40's who obtain mammograms are found to have breast cancer. So P(breast cancer/abnormal mammogram) = 0.025 – quite like Howson's made-up example. Using Howson's construal of the evidence, such an abnormal mammogram gives confidence of the *absence* of breast cancer. *So the follow-up that discovered these cancers would not have been warranted.* (Mayo, 1997b, S209, note 11, our italics).

The message is clear: Bayesians in general, and Howson in particular, are to be held (morally) accountable in cases like these if the women involved do not have additional tests. Error-statisticians, on the other hand, who would not have such confidence in the link between abnormal mammograms and the absence of breast cancer (i.e., an abnormal mammogram would *not* constitute a *severe* test of the absence of breast cancer) would not, in this case wrongly, advise women not to have additional tests.

Whether Howson or Mayo or we are right in their assessment of the severity of this sort of test is beside the point here. The point is that whether or not one decides to have additional tests made depends on utilities, the personal costs of developing breast cancer, for example, as much as it does on probabilities, and the utilities will vary from case to case, [69] even from individual to individual. Presumably women who are thirty not only will but should behave differently from women who are ninety when their mammograms are abnormal. But neither Bayesians nor error-statisticians should take these utilities into account when framing their characterizations of severity.

The issues here are very complex and we intend only to clarify, and not resolve, them. The Bayesian might seem committed to a close tie between confirmation and action by way of a particular view about belief: to confirm a hypothesis is to raise one's degree of belief in it, and to believe in a proposition is, other things being equal, to be ready to act on it. This view about belief is behavioristic; beliefs are dispositions to behave. The error-statistician might seem committed to a close tie between "confirmation" and action by way of the sort of principle central to Neyman–Pearson statistical inference: the degree of confirmation that a hypothesis H must have before one is warranted in choosing to act on H relative to an objective P is a function of the seriousness of the error relative to P resulting from basing the action on the wrong hypothesis.[70] Indeed, the very concept of a severe test would seem to make the connection explicit. As we noted at the outset, one is warranted in acting on a hypothesis just in case the hypothesis has passed a severe test.

The problem from our point of view is that, in Isaac Levi's words, "there is always the possibility that some objectives exist relative to which mistakes are so serious as to demand enormously high degrees of confirmation" (Levi 1960, 352, note 15). If we want a conception of severity that is generally usable, then it makes little sense to adjust it for high utilities across the board. Better to keep utilities and severities, and degrees of confirmation, separate.

We have quoted De Finetti before. It is helpful to do so again.

> I do not deem the usual expression 'to accept hypothesis H' to be proper. The 'decision' does not really consist of this 'acceptance' but in *the choice of a definite action A*. The connection between the action *A* and the hypothesis *H* may be very strong, say 'the action *A* is the action that we would choose if we knew that *H* was the true hypothesis'. Nevertheless, this connection cannot turn into an identification. (De Finetti 1951, 217).

The most a scientist can do is to devise severe tests of hypotheses, and then report on their success or failure. Although as a rational agent she will undoubtedly make use of successful hypotheses in future work, and shun failed ones, she cannot also advise us on the long-term costs and benefits of doing so,[71] however frustrating this is to patients who want to be told how they should proceed. For the long-term costs and benefits have to do with particular objectives, and scientists cannot determine what these will be in advance, even with regard to the "acceptance" or rejection of otherwise unproblematic hypotheses.

Now to summarize our discussion. We have argued that despite the conventional wisdom that runs them together, the notions of belief and evidence must be distinguished, and have provided criteria on the basis of which to do so. These criteria are generally Bayesian in character. We combine both in our characterization of a severe test. The notion of a severe test, in turn, is bound up with one traditional understanding of what it is to "accept" a hypothesis. Such a characterization better captures our intuitions concerning at least certain kinds of scientific tests than do competing alternatives like Deborah Mayo's, hence in a rough sort of way and on its own grounds is "acceptable". Accounts of severe tests should themselves be put to severe tests. We have tried to do so here.

No two philosophers have done more to deepen our understanding of the notions of evidence and confirmation than Hempel and Carnap. But in our view they made the fundamental mistake of running these two notions together. In the same essay we quoted at the outset, Hempel says that "an empirical finding is relevant for a hypothesis if and only if it constitutes either favorable or unfavorable evidence for it; in other words, if it either confirms or disconfirms the hypothesis".[72] The "in other words" leads to a variety of problems. In particular, it precludes an adequate analysis of severe testing. Or so we believe.

NOTES

[1] Previous versions of this paper were read at the American Philosophical Association Central Division Meetings, the Rabindra Bharati University, the University of Manitoba, and the University of Western Ontario. We benefitted from the discussions at each. We thank Douglas Allchin, John Bennett, Jim Berger, Robert Boik, Somnath Chakraborty, Martin Curd, William Donaldson, Warren Esty, Travis Ganji, I. J. Good, William Harper, Geoffrey Hellman, Colin Howson, Noretta Koertge, John Roberts, Teddy Seidenfeld, Kent Staley, and Mark Taper for their very helpful comments. We also thank participants in the National Endowment for the Humanities Summer Seminar on "Philosophy of Experimental Inference: Induction, Reliability, and Error" held at Virginia Polytechnic in 1999. Special thanks are due to Deborah Mayo, who directed the Seminar, for her criticisms at the APA Central Division Meeting and for her subsequent correspondence on the issues raised. Finally, we would like to thank the five anonymous referees for *Synthese* who suggested numerous improvements and corrected several errors in the original version of this paper.

[2] Quoted by Good (1983, 132–133).

[3] Reprinted in Hempel (1965, 3–46).

[4] Ibid., p. 42.

[5] More precisely, just in case the utilities involved in acting on the hypotheses are roughly equal. In the final section of this paper, we set out reasons for making this qualification.

[6] Again, we follow Hempel: "As is well known, empirical science decides upon the acceptability of a proposed hypothesis by means of suitable tests" (ibid., p. 83). On our reading (although not on Hempel's) "suitable tests" are *severe*. The passage of "suitable tests" does not imply, any more for us than it does for him, that the hypothesis in question might not be overturned at some later date by the discovery of new, recalcitrant evidence

or alternative, more explanatory hypotheses; as Hempel notes, many hypotheses "once accepted" are no longer. Passing a severe test does not imply that the hypothesis in question is true.

[7] In this and certain other ways we part company with Karl Popper, who also emphasized the importance of severity in the acceptance of scientific hypotheses. See Popper (1959, *passim*).

[8] First propounded by Kyburg (1970). We follow Sober's informal statement of the paradox.

[9] Kyburg himself avoids the paradox by giving up what he terms the "conjunction principle", that if a set of hypotheses is individually accepted, then their conjunction must be as well.

[10] In Utilitarianism, Chapter 4, where he makes the same point in connection with "desirable".

[11] Thus the astronomer Lalande, who with Madame Lepaute completed the numerical calculations based on Clairaut's methods which made possible the precise prediction of the comet's return in 1758–59: "The universe beholds this year the most satisfactory phenomenon ever presented to us by astronomy: an event which unique until this day [very low probability] changes our doubts to certainty [since it was likely only on the Newtonian theory] and our hypotheses to demonstrations". For more on the "severe test" character of Halley's prediction, see the now-standard work by Alan Cook (1998), esp 8.2.

[12] For a variety of reasons not relevant here, Popper prefers to talk about "corroboration" instead.

[13] In fact, two intuitive conditions on severity are mingled together here. One is that the probability of the predicted return is low. The other is that the range of tolerable error is relatively narrow. In general, the more accurate the instruments employed, and hence the range of tolerable error, the more severe the test. See Horwich (1982, 104–105). We will simply assume the second of these conditions in what follows; for our purposes here, it is unimportant.

[14] Thus Popper's discussion is entirely in terms of *degrees* of testability, without any attempt to characterize an (absolute) standard of severity.

[15] Mayo (1996, 1997a, b).

[16] Applications of the Bayesian approach in the medical sciences are quite common. For recent applications of this approach see, *inter alia*, Demissie et al. (1998) and Epstein et al. (1996). For the use of Bayesian statistics in epidemiological cases, see a recent collection of articles edited by Berry et al. (1996). Richard Jeffrey is one of the very few philosophers who has applied a Bayesian technique in analyzing diagnostic test results. See Jeffrey et al. (1988). See also Jeffrey (1992), where Bayes Factor is used for a proper diagnosis of cancer among patients having tumors. Among philosophers, see also Achinstein (1983, 1984, 2001) for his work on evidence as it relates to explanation. Achinstein (1983) is a good and handy collection of articles on evidence written by philosophers. See especially the articles by Salmon and Glymour in that collection. Good (1992) has written extensively on Bayes Factor. A. Raftery is a well-known statistician who has devoted several of his papers to using Bayes Factor in analysing diagnostic test results (Raftery et al. 1997; Raftery 1995).

[17] The connection between confirmation and belief is emphasized by non-Bayesians as well. Thus Hempel (1965, 8): "It is now clear that an analysis of confirmation is of fundamental importance also for the study of a central problem of epistemology, namely, the elaboration of standards of rational belief ..."

[18] In what follows we distinguish between the relevant "data", *D*, for particular hypotheses and the "evidence", *E*, for them. If, for example, the "data" are taken to be series of "observations", these "observations" do not constitute "evidence" unless our second condition, in terms of which this notion is characterized, holds. In order to simplify our account, we have not made mention of "background information" in the statement of our conditions; it may be assumed.

[19] If the ratio is 1, then D has the same evidential significance for the two hypotheses. If the ratio is less than 1, then D provides evidence against H in favor of H′. In the cases which follow, we take H′ = not-H since we want to focus on single tests and particular hypotheses.

[20] Since we characterize acceptability in terms of severity, it follows that for us acceptability is not a probability. In this respect, as in some others, we follow Popper (1959, 394): "…I regard the doctrine that *degree of corroboration or acceptability cannot be a probability* as one of the more interesting findings of the philosophy of knowledge. It can be put very simply like this. A report of the result of testing a theory can be summed up by an appraisal. This can take the form of assigning some degree of corroboration to the theory. But it can never take the form of assigning to it a degree of probability; for *the probability of a statement (given some test statements) simply does not express an appraisal of the severity of the tests a theory has passed, or of the manner in which it has passed these tests*".

[21] See Royall (2000, 760).

[22] It is more natural to speak of "reasons for believing a hypothesis", when, more properly, we should say "reasons for believing a hypothesis to some degree or other". But the more natural form of speech should not mislead in what follows.

[23] See Popper (1959, 87n. 1).

[24] Duhem (1954) famously argues that there are no crucial experiments in science. Very possibly, he would infer, and for the same reasons, that there are no severe tests. But his notion of a crucial experiment is very closely tied to a deductivist account of theory testing. Although we cannot argue for this claim here, our account of theory testing escapes Duhem's strictures.

[25] See Bandyopadhyay and Boik (1999).

[26] See Pagano et al. (2000, 137–138).

[27] A claim with which practitioners agree, as we shall see in a moment. Among several other problems with it, the false positive rate is high.

[28] Recall that this ratio does not satisfy the probability calculus; it may vary from zero to infinity.

[29] We have made these numbers up to illustrate a point, but the situation is conceivable.

[30] One might wonder why CC is satisfied in this case since the posterior probability of *H* is less than its prior probability. The fundamental point here has to do with the Bayesian requirement that the agent's beliefs be consistent. There is nothing in the decrease in the posterior probability in the HIV case that violates this requirement; the probabilities assigned represent "ideal" personal degrees of belief and are not constrained by anything other than the probability calculus (including the rule of conditional probability).

[31] See Pagano et al. (2000).

[32] Ibid., 139.)

[33] The example is made up, although again the numbers are not implausible.

[34] See especially Mayo (1996, 1997a, b). At times, Mayo (1996, 82) suggests that her account of evidence applies only to cases in which relative frequencies are not available to determine prior probabilities (i.e., applies *only* to "non-standard" cases): "*Except* for such contexts, however, the prior probabilities of the hypotheses are problematic. Given that logical probabilities will not do, the only thing left is subjective probabilities. For many, these are unwelcome in scientific inquiry".

[35] See Mayo (1997b), Forster (2000), and Giere (1997).

[36] The Bayesian, for example, thinks that actual outcomes of testing procedures are all that matters, the error statistician that we must take into account possible outcomes as well.

[37] Anyone who has conducted even elementary scientific experiments knows about mean deviances, Chi-squares, and confidence intervals.

[38] In a private communication, Mayo has made this explicit. "My notion of severity is an error probability, and unless one appeals to error probabilities, one does not get the needed guarantee of severity (as I define it)".

[39] See Carnap (1949, 1955). In light of the "statistics wars" now said to be raging, it is worth quoting his call for peace. "Today an increasing number of those who study both sides of the controversy which has been going on for thirty years are coming to the conclusion that here, as so often before in the history of scientific thinking, both sides are right in their positive theses, but wrong in their polemic remarks about the other side" (Brody, 1970, 445). We think that there are genuine issues between Mayo and us, which can be decided, but also that there is little room for "polemic remarks".

[40] Mayo suggests that her account of severity owes its origins to Popper, and might be taken to imply that, since Popper was perhaps first to emphasize the notion of severity, she has a special claim on it. Four quick points. First, as already indicated, the theme of Popper's notion of severity is that the more surprising the experimental/observational outcome, the better will be its value as a test for a particular hypothesis. But this theme is captured in both accounts, ours as well as hers. Second, Popper links severity to his idea that hypotheses can never be confirmed, only "corroborated" (i.e., to this point they have not been falsified); but neither of our accounts turns on anything like the idea of "corroboration". Third, Popper is a deductivist in this sense, that he thinks that the rules of deductive logic suffice for capturing the idea of falsification and hence the core of all scientific reasoning; in his view, the problems attendant on any so-called inductive logic are "insurmountable". But both Mayo's error-statistical and our Bayesian account are very much non-deductivist in spirit. For both of us, probability theory is crucial to an explication of "severity". Fourth, for Popper corroboration is exclusively a measure of past performance and provides no indication concerning the success of a hypothesis in the future. Whereas for both Mayo's and our own positions, a measure of severity takes account of past experiments in part to predict the course of future ones. In this sense as well, both of them are "inductivist".

[40] Mayo (96, 180). Mayo does not make our distinction between "data" and "evidence". Her (unanalyzed) notion of "evidence", *e*, is in fact akin to our "data", hence in what follows we frame her conditions in terms of "*D*".

[42] Mayo (1996, 179).

[43] Mayo apparently rejects our attempt to put her criteria in terms of likelihoods. Thus in a private communication she writes: "Severity is an error probability – NOT a likelihood. EGEK (Mayo, 1996) spends most of its time demonstrating that mere likelihoods and likelihood ratios fail to control error probabilities". In another private communication she writes: "... one cannot appeal to error probabilities if one accepts the Likelihood Principle

LP. Thus, the choice is to renounce error probs or renounce the LP"; if you renounce error probs, "then you must also renounce severity in the sense I define it".

Although we shall have more to say about the issues involved in the text, it is necessary here to make three points.

First, it is necessary to distinguish sharply between the use of likelihoods and the Likelihood Principle. Our use of likelihoods is intended to make the formal differences between our positions sharper and does not, so far as we can see, beg any important questions in favor of the Bayesian position. A variety of non-Bayesians, Akaike statisticians for example, make similar use of likelihoods (although the Akaike Information Criterion, on which Akaikean statistical theory rests, violates the Likelihood Principle). Indeed, Mayo makes one crux of the dispute between us the claim that "if we find the probability of getting such a value [i.e., how much evidence there would be], under the assumption that H is false, then H has NOT passed a severe test".

Second, although there is some disagreement among Bayesians concerning the status of the Likelihood Principle, we too reject it as an account of *severity*. That is, we accept LP as a condition on *evidence*; indeed, its acceptance justifies ESC. As we have already suggested, and will later argue, any acceptable account of "evidence" must subscribe to LP. But our first condition, CC, violates LP, and hence so too does our account of "severity" (see our earlier discussion of the HIV case and footnote 28). On this point, as on several others, we agree with Mayo. Our differences with her account lie elsewhere.

Third, we distinguish the concepts of evidence and severity, for reasons to be made clear. Mayo interdefines them. Once again, the concept of evidence must make use of LP, the concept of severity cannot do so. Both of our conditions, of course, make use of likelihoods.

[44] Mayo (1996, 180).

[45] This is the sort of "Bayesian" assumption that Mayo very much rejects. But again, such idealization is characteristic of scientific practice (Giere (1996, S183), speaks of "over-idealization" without, however, providing a criterion) and, while the two accounts may be difficult to compare in their terms, aims, abstractions, and presuppositions, we need to generate some quantitative assessments of specific tests if we are to compare them at all.

[46] "... the statement of high probability need not be obtained by reference to a statistical calculation; some of the strongest arguments from error are based on entirely qualitative assessments of severity" (Mayo 1997b, S203).

[47] Note that if a test satisfies our conditions of severity, it satisfies Mayo's as well. On the other hand, as this Table makes clear, a test that satisfies Mayo's conditions of severity does not necessarily satisfy ours. Our account is in this sense "more severe" than hers, possibly a result of our distinguishing more sharply than she does between the concepts of evidence and severity.

[48] Achinstein (2001, 134–136) sets out some very simple and schematic cases to make the same point, that neglect of prior probabilities, whether they are interpreted in an objective or a subjective way, leads to counter-intuitive results. In a way very different, and perhaps less sympathetic, than ours, he goes on to argue that Mayo cannot reconstrue her account so as to save it from these results.

[49] The mistake often made by overlooking the notion of prior probability in analyzing diagnostic test results has been discussed in the literature (Schulzer 1994). Among other considerations, the prevalence of the disease in question is crucial. Galen and Gambino (1975, pp. 4–5) provide this real-world example: "... if a test were so unbelievably good that there were never any false-negative results in patients with the disease and there was

only 1 false-positive result per 10,000 subjects who were free of the disease – but the disease had a very low prevalence of between 5 and 10 per 100,000 (e.g., phenylketonuria) – then the predictive value of an abnormal result obtained with this unbelievably superb test would be only 50%! This means that half of the abnormal results would occur in subjects *who did not have true phenylketonuria* if the test were performed on *all* newborns – which is exactly what happened when legislators mandated routine PKU screening".

[50] We will have more to say on the connection between probabilities and utilities in the final section of this paper.

[51] Geoffrey Hellman has urged us to make this point explicit.

[52] Although according to Good (1983, 132), likelihoodists conflate the evidence and belief questions.

[53] A claim spelled out in detail, although in somewhat different directions from ours, by Glymour (1980).

[54] Following Grandy's (1967) account.

[55] Hempel (1965), chapter 2).

[56] "Following a practice common to testing approaches, I *identify* 'having good evidence (or just having evidence) for H' and 'having a good test of H', That is, to ask whether e counts as good evidence for H, in the present account is to ask whether H has passed a good test with e" (our italics) Mayo (1996, 179). Staley (2001) has re-cast Mayo's error-statistical position on evidence.

[57] See Bernard (1983, 57), Gelman et al. (2000, 190), and Berger (1985, 33).

[58] Briefly, we characterize evidence in terms of the ESC, that is, in terms of likelihood ratios which satisfy LP. We then characterize severity in terms of the ESC and posterior probability (whose calculation involves certain simplicity considerations). Finally, we characterize acceptability in terms of severity.

[59] In addition to Sober (see also Sober (1988)), Hacking (1965, 103–109) sets out arguments for actualism, one of them in connection with the problem of optional stopping rules.

[60] As Eells (1993) puts it, "we have to work with what we have to work with".

[61] He first raised it in connection with Good's concept of "weight of evidence". Good's paper is in (1985), as is Seidenfeld's concise response (1985, 264–266). He raised it again (in private correspondence) as a problem for our account of severity. We have adapted the objection slightly so as to fit it more precisely to our present account.

[62] Seidenfeld (1985) goes on to say, conflating degrees of belief with weight of evidence in ways that we do not, that we should have the same degree of belief in the two hypotheses. Since on our account this does not follow, then once again, according to him, our account should be rejected.

[63] See Spanos (1999, 12–13).

[64] In our case, contra Seidenfeld, when we say that D, the datum, provides evidence for H as against *not-H*, our discussion presupposes the usual framework of a statistical model. His attempted counter-example does not have force against our account unless it is cast in this same framework.

[65] Horwich (1982, 105).

[66] See Hellman (1997).

[67] See Earman (1992) on a related theme.

[68] We are indebted to one of the anonymous referees for pointing this out.

[69] Perhaps our four model cases could be ranked in the following rough way with respect to the utilities involved "in making very sure" that one had or didn't have the disease: HIV, PAP, MAL, TB.

[70] See Levi (1960).

[71] We have followed the discussion of Jeffrey (1956) in making this point.

[72] *Aspects of Scientific Explanation*, 5. For an example of how Carnap runs evidence and belief together, see Brody (1970, 441).

REFERENCES

Achinstein, P.: 1983, *The Nature of Explanation*, Clarendon Press, Oxford.

Achinstein, P. (ed.): 1984, *The Concept of Evidence*, Clarendon Press, Oxford.

Achinstein, P.: 2001, *The Book of Evidence*, Clarendon Press, Oxford.

Bandyopadhyay, P. and Boik, R.: 1999, 'The Curve-Fitting Problem: A Bayesian Rejoinder', *Philosophy of Science* (Proceedings), S390–S402.

Berger, J.: 1985, *Statistical Decision Theory and Bayesian Analysis*, second edition, Springer, New York.

Berger, J. et al.: 1987, 'Testing a Precise Hypothesis with Discussion', *Statistical Science* **3**, 317–352.

Berger, J. et al.: 1996, 'The Intrinsic Bayes Factor for Model Selection and Prediction', *Journal of American Statistical Association* **91**, 109–122.

Berger, J. et al.: 1997, 'Unified Frequentist and Bayesian Testing of a Precise Hypothesis (with Discussion)', *Statistical Science* **12**, 133–160.

Bernard, G.: 1987, 'Comments', in Bernard, De Groot, Lindley, and Smith (eds), *Bayesian Statistics*, Vol. II, North-Holland Publishing Company, Amsterdam, pp. 57–60.

Berry, D.: 1996, *Statistics: A Bayesian Perspective*, Duxbury Press, Belmont.

Berry, D. et al.: 1996, *Bayesian Biostatistics*, Marcel Dekker, Inc., New York.

Birnbaum, A.: 1969, 'Concepts of Statistical Evidence', in Morgenbesser, Suppes, and White (eds), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, St. Martin's Press, New York, pp. 112–143.

Brody, B.: 1970, *Readings in the Philosophy of Science*, Prentice-Hall, Englewood Cliffs, N.J.

Carnap, R.: 1949, 'The Two Concepts of Probability', in Feigl and Sellars (eds), *Readings in Philosophical Analysis*, Appleton-Century-Crofts.

Carnap, R.: 1955, 'Statistical and Inductive Probability', pamphlet published by the Galois Institute of Mathematics, Brooklyn, N.Y., reprinted in Brody (1970), pp. 440–450.

De Finetti, B.: 1951, 'Recent Suggestions for the Reconciliation of Theories of Probability', *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley & Los Angeles, pp. 217–225.

Demissie, et al.: 1998, 'Bayesian Estimation of the Asthma Prevalence, and Comparison of Exercise and Questionnaire Diagnostics in the Absence of a Gold Standard', *Annals of Epidemiology* April **8**(3), 201–208.

Duhem, P.: 1962, *The Aim and Structure of Physical Theory*, Atheneum, New York.

Earman, J.: 1992, *Bayes or Bust?* MIT, Cambridge, MA.

Edwards, Lindman, and Savage: 1963, 'Bayesian Statistical Inference for Psychological Research', *Psychological Review* **70**(3), 193–242.

Eells, E.: 1993, 'Probability, Inference, and Decision', in J. Fetzer (ed.), *Foundations of Philosophy of Science*, Paragon Press, New York, pp. 192–208.

Epstein et al.: 1996, 'Bayesian Imputation of Predictive Values when Covariate Information is Available and Gold Standard analysis is Unavailable', *Statistics in Medicine*, March **15**(5), 463–476.

Forster, M.: 2000, 'The New Science of Simplicity', in H. A. Kreuzenkamp et al. (eds), *Simplicity, Inference and Economic Modeling*, Cambridge University Press, Cambridge.

Galen, R. and Gambino, R.: 1975, *Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses*, John Wiley & Sons, New York.

Gelman, A. et al.: 2000, *Bayesian Data Analysis*, Chapman & Hall, London.

Giere, R.: 1997, 'Scientific Inference: Two Points of View', *Philosophy of Science* (supplement), **64**, S180–S184.

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

Good, J.: 1983, 'Some Logic and History of Hypothesis Testing', in *Good Thinking*, University of Minnesota Press, Minneapolis, pp. 129–148.

Good, J.: 1992, 'The Bayesian/Non-Bayesian Compromise', *Journal of the American Statistical Association* **87**, 597–606.

Grandy, R.: XXXX, 'Some Comments on Confirmation and Selective Confirmation', *Philosophical Studies* **18**, 19–24.

Greenhalgh, T.: 2001, *How to Read a Paper: The Basics of Evidence Based Medicine*, BMJ Books, London.

Groopman, J.: 1999, 'Contagion', *The New Yorker*, September 13, pp. 34–49.

Guth, A.: 1997, *The Inflationary Universe*, Boston, Addison-Wesley.

Hacking, I.: 1965, *Logic of Statistical Inference*, Cambridge University Press, Cambridge.

Hellman, G.: 1997, 'Bayes and Beyond', *Philosophy of Science* **64**, 191–221.

Hempel, C.: 1965, *Aspects of Scientific Explanation*, Free Press, New York.

Horwich, P.: 1982, *Probability and Evidence*, Cambridge University Press, Cambridge.

Howson, C.: 1997a, 'Error Probabilities in Error', *Philosophy of Science* (Supplement), **64**, S185–S194.

Howson, C.: 1997b, 'The Logic of Induction', *Philosophy of Science* **64**, 268–290.

Howson, C.: 2000, *Hume's Problem: Induction and the Justification of Belief*, Clarendon Press, Oxford.

Jeffrey, R.: 1956, 'Valuation and Acceptance of Scientific Hypotheses', *Philosophy of Science* **33**, 236–246.

Jeffrey, R. et al.: 1988, 'Probabilizing Pathology', *Proceedings of the Aristotelian Society*, vol. 89, part 3.

Jeffrey, R.: 1992, 'Radical Probabilism', *Probability and the Art of Judgment*, Cambridge University Press, Cambridge.

Johnston, D.: 1990a, 'Interpreting Statistical Insignificance: A Bayesian Perspective', *Psychological Reports* **6**, 115–121.

Johnston, D.: 1990b, 'Sample Size and Strength of Evidence', *Abacus* **26**, 17–35.

Kyburg, H.: 1970, 'Conjunctivitis', in *Epistemology and Inference*, University of Minnesota Press, Minneapolis.

Levi, I.: 1960, 'Does Science Make Value Judgments?' *Journal of Philosophy* **57**, 345–357.

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

Mayo, D.: 1997a, 'Duhem's Problem, The Bayesian Way, and Error Statistics, or "What's Belief Got to Do With It?" and "Response to Howson and Lauden" ', *Philosophy of Science* (June), 222–244, 323–333.

Mayo, D.: 1997b, 'Error Statistics and Learning from Error: Making a Virtue of Necessity', *Philosophy of Science* (Supplement), **64**, S195–S212.

Mayo, D. and Kruse, M.: 2001, 'Principles of Inferences and Their Consequences', in Cornfield and Williamson (eds), *Foundations of Bayesianism*, Dordrecht, Kluwer Academic Publishers, pp. 381–404.

Pagano et al.: 2000, *Principles of Biostatistics*, second edition, Duxbury Thomson Learning, Pacific Grove, CA.

Popper, K.: 1959, *The Logic of Scientific Discovery*, English Edition, Hutchinson, London.

Raftery, A.: 1995, 'Bayesian Model Selection in Social Research', in Marsden (ed.), *Sociological Methodology*, Blackwell, Cambridge, MA.

Raftery, A. et al.: 1997, 'Computing Bayes Factor by Combining Simulation and Asymptotic Approximations', *Journal of American Statistical Association* **92**(439), 903–915.

Rawls, J.: 1971, *A Theory of Justice*, Harvard University Press, Cambridge, MA.

Rothman, K. et al. (1998) *Modern Epidemiology*, second edition, Lippincott-Raven, Philadelphia.

Royall, R.: 1997, *Statistical Evidence: A Likelihood Paradigm*, Chapman & Hall, London.

Royall, R.: 2000, 'On the Probability of Observing Misleading Statistical Evidence (with Discussion)', *Journal of American Statistical Association* **95**(482).

Royall, R.: 2001, 'The Likelihood Paradigm for Statistical Evidence', in M. Taper et al. (eds), *The Nature of Scientific Evidence*, University of Chicago Press, Chicago, pp. 760–780.

Savage, L. (ed.): 1962, *The Foundations of Statistical Inference: A Discussion*, Methuen, London.

Schulzer, M.: 1994, 'Diagnostic Tests: A Statistical Review', *Muscle and Nerve*, July **17**(7), 815–819.

Seidenfeld, T.: 1985, 'Comment on Good's Philosophy', in Bernard, De Groot, Lindley, and Smith (eds), *Bayesian Statistics*, Vol. II, North-Holland Publishing Company, Amsterdam, pp. 264–266.

Shafer, G. and Vovk, V.: 2001, *Probability and Finance*, New York, John Wiley.

Sober, E.: 1993, 'Epistemology for Empiricists', *Midwest Studies in Philosophy* **XVIII**, 39–61.

Spanos, A.: 1999, *Probability Theory and Statistical Inference*, Cambridge University Press, Cambridge.

Staley, K.: 2001, 'What Experiment Did We Just Do? Counterfactual Error Statistics and Uncertainties About the Reference Class', *Philosophy of Science* **69**, 279–299.

Steel, D.: 2003, 'A Bayesian Way to Make Stopping Rules Matter', *Erkenntnis* **58**, 213–222.

van Fraassen, B.: 1980, *The Scientific Image*,Oxford University Press, London.

Wheeler, G.: 2000, 'Error Statistics and Duhem's Problem', *Philosophy of Science* **67**, 410–420.

Prasanta S. Bandyopadhyay
Department of History and Philosophy
Montana State University
2-155 Wilson Hall
Bozeman
MT 59717-2320
U.S.A.
E-mail: psb@montana.edu

Gordon G. Brittan, Jr.
Department of History and Philosophy
Montana State University
2-155 Wilson Hall
Bozeman
MT 59717-2320
U.S.A.
E-mail: uhigb@montana.edu