Can Errror-Statistical Inference Function Securely?

Kent Staley

Saint Louis University

staleykw@slu.edu

February 25, 2007

1 Introduction

Among the important tasks for a theory of ampliative inference are the classification of the kinds of "material postulates"[1] that are needed for specific inferences and the explanation of how those assumptions function in such inferences. The present paper undertakes an analysis of Deborah Mayo's error-statistical (ES) account of scientific evidence in order to clarify just these points as well as to explain and illustrate the importance of the philosophically neglected notion of the *security* of an inference [21]. After finding that, on the most straightforward reading of the ES account, it does not

[1] The term is John Norton's [18]. Norton proposes that material postulates and not formal inductive schemas underwrite inductive inference. The perspective of the present paper is that while Norton rightly emphasizes the importance of material postulates in underwriting specific inductive inferences and explaining the growing power of inductive reasoning, the material postulates do not themselves tell us how they should be used. See also [25].

succeed in its own stated aims, two remedies are considered. The first remedy introduces a *relativization* of evidence claims. The second introduces stronger *assumptions*. These two approaches may agree in the kinds of factual propositions they introduce, but they differ in how those propositions function in the inference. The choice between these approaches turns on the relative value attached to two aims of inquiry that are in substantial tension: the ability to draw strong, informative conclusions and the ability to reason securely.

In section two I situate the present discussion in the context of concepts of evidence and inference. Section three introduces the error-statistical account and the problem I will investigate. I introduce the notion of secure inference in section four, followed in section five by two proposals for solving the problem facing error-statistics. Section six introduces an example in contemporary physics, and discusses the application of these to approaches to that example. I draw some conclusions in section seven.

2 Error-statistical evidence/inference

I employ Mayo's error-statistical apparatus [14] for two related purposes: as a theory of *evidence* and as a theory of *inference*. I begin by clarifying these two notions, their relations to one another and their relations to other possible uses of the apparatus.

I follow Achinstein [2] in attributing the following significance the concept of evidence as here employed:[2] that some body of data or facts E is

[2] Achinstein [2] delineates several evidence concepts. Evidence in the sense here employed comes closes to the concept he denotes "potential evidence." For a discussion of

evidence for some hypothesis H entails that E constitutes a good reason to believe that H.

The evidence concept is, on this approach, rather strong [1]. It also turns out to be closely connected with inference: when E is evidence for a hypothesis, it also becomes the basis for an inference: 'E, therefore H'.

Certainly other aims besides true belief are at work in the scientific assessment of hypotheses. We can regard the question of what should be inferred from our data as a special case of the more general problem of "model (or hypothesis) selection." When deciding which model to select on the basis of given data, however, it matters what the purpose of the selection is: to predict future data in a given application, to develop a hypothesis regarding some distinct phenomenon, to optimize a strategy for solving some related problem, etc. The focus on evidence and inference confines us to those cases in which a model or hypothesis is selected for purposes of believing it to be true. (So as not to beg questions with regard to scientific realism, those hypotheses can be of the kind: 'proposition H is empirically adequate.')

One consequence of the present approach should be noted at the outset. If we focus on this strong concept of evidence as involving reasons for belief, then conditions ensuring merely that a particular hypothesis fares better than some specific alternatives in light of given data will in general be insufficient for evidence. In particular *inclusive* concepts of support will not

the compatibility of Achinstein's understanding of potential evidence with Mayo's error-statistics, see [22]. In that discussion, I part ways with Achinstein regarding the concept of 'a good reason to believe.'

be strong enough to capture this notion of evidence.

An *inclusive* concept of support is one such that the conditions for the truth of 'E supports H' can be met even when there exists an incompatible competing hypothesis H' such that those conditions are at least as well met by 'E supports H''.

Inclusive concepts of support can be suitable for many purposes. Inclusiveness, for example, is a feature of approaches to hypothesis assessment that focus on making comparative judgments.[3] We might wish to consider a family of hypotheses and pick one that is sufficiently predictively accurate. Suppose our procedure is such as to pick out the member of that model M (here 'model' indicates a family of hypotheses with the same structure, but differentiated by parameter values) that has the highest estimated predictive accuracy (under suitable circumstances, such a procedure can be provided by use of the Akaike Information Criterion, for example). Data E can support the selection of hypothesis H for such a purpose although another hypothesis that is not part of M might have an even higher estimated predictive accuracy [10].

If our concern is true belief, however, inclusive concepts of support will be too weak, since the conditions for such a concept can be met by one hypothesis at the same time as another incompatible alternative. If E is a good reason for me to believe H then it cannot also be a good reason for me to believe $\neg H$. The unsuitability of inclusivist accounts for present purposes will have important implications for the interpretation of the error-statistical

[3] Mayo has argued against comparativism independently [15] [16]. The present argument identifies the weakness of comparative approaches in their inclusivity.

theory, to which I now turn.

3 What do you mean, "H is false"?

The error statistical theory of evidence can be articulated in terms of the 'severe test' requirement as follows: Supposing that hypothesis H is subjected to test procedure T, resulting in data E, E constitutes evidence for H (H can reasonably be inferred from E) just in case:

SR1 E fits H,

SR2 the probability of H passing T with an outcome such as E (i.e., one that fits H as well as E does), given that H is false, is very low ([14], esp. 178–87).

Let us call an evidence claim that is warranted by the satisfaction of these two conditions a claim about "*SR* evidence." For our purposes we can assume that fit is measured probabilistically, for example in terms of the likelihood of H on E. The crucial question concerns the second criterion ("severity") and how it is to be understood.

To see the nature of the problem, consider an artificially simple example. A clinical trial is performed to determine the risk that nausea (N) will be among the outcomes of a proposed treatment (C) for asthma (A). That is to say, we wish to use data from the trial in the assessment of the hypothesis H: Individuals suffering from A subjected to the treatment C have a probability p_0 ($\pm\epsilon$) of exhibiting the outcome N.

To claim that data E support this hypothesis H (for some specific value of p), by the standards of the error-statistical theory of evidence, it is apparently required that it be very improbable that the test would yield data fitting H as well as E fits H supposing H to be false. Can this condition be met?

The answer to this depends on our interpretation of the condition 'H is false.' To clarify the issues I offer two possibilities here, both of which will turn out to be problematic. The "weak sense" of *SR2* takes 'H is false' to be a covering term for a family of alternatives to H, $\{H_p\}$, such that each H_p asserts: 'Individuals suffering from A subjected to the treatment C have a probability p of exhibiting the outcome N', where $p < p_0 - \epsilon$ or $p > p_0 + \epsilon$. The "strong sense" of *SR2* treats 'H is false' more literally: some other hypothesis is true that entails $\neg H$. (Note that on either reading 'H is false' is insufficient by itself to determine any probability distribution for the outcomes of the test.)

The advantage of taking *SR2* in the weak sense is that it makes the severity criterion tractable, so that there is no apparent problem that would prevent H from passing such a test (supposing the data to cooperate, of course). However, this interpretation jeopardizes the attempt to employ *SR1* and *SR2* in a non-inclusive theory of evidence, however. On the weak reading, the claim that E is *SR* evidence for H would seem only to license the conclusion that E confers greater support on H than on any member of $\{H_p\}$. Clearly, however, there are alternative hypotheses that are not members of $\{H_p\}$ (hypotheses that confer different probabilities on different asthma sufferers under treatment, for example), and the fact that E confers greater support on H than any member of $\{H_p\}$ does not entail that none of those alternatives will receive equally strong support from E.

On the strong reading, the *SR2* requirement would mean that under *any* incompatible alternative[4] H', if H' were true, it would be very improbable that the test would yield a result that fits H as well as E does. Thus understood, however, *SR2* cannot in general be met.

This is easiest to see for the case of maximally likely alternatives constructed post-data. Suppose that our data comprise a simple listing of individuals surveyed by our investigation, where for each individual $a_i, i = 1, \ldots, n$, we record whether the treatment was applied or not (Ca_i or $\neg Ca_i$) and whether the outcome in question was observed or not (Na_i or $\neg Na_i$). With our data in hand, we can construct a maximally likely alternative hypothesis simply by collecting together all m of the individuals for which Na_i obtains. We then relabel these individuals a_j, where $j = 2i + 1$. For all of the individuals for which $\neg Na_i$ is found, we relabel them as a_k, where $k = 2i$. We have thus assured that individuals with positive outcomes bear even indices and those with negative outcomes have odd indices. We can now formulate the following alternative hypothesis H': Individuals suffering from A that are labeled with odd indices that are subjected to the treatment C have a probability $p = 1$ of exhibiting the outcome N. Those with even numbered indices have a probability $p = 0$ of exhibiting the outcome N.

[4] Even on this reading, alternative hypotheses must be truly logically incompatible (they should not be more general explanatory theories, for example – see [14], pp. 198–99), and sufficiently definite in content to determine a probability distribution over the sample space in question.

H' is of course an absurd hypothesis produced by what Mayo calls "gellerization." Nonetheless, given the trivial possibility of constructing such an alternative for any hypothesis under consideration, it appears that a strong reading of *SR2* leaves us with the unwelcome conclusion that no hypothesis can pass a severe test given any data.

Mayo's response to this kind of example seems to take the threat to arise from the possible claim that the alternative hypothesis H' is just as well-tested as the H. She insists that in a case in which "the hypothesis is constructed on the basis of data" both the data and the content of the hypothesis must be treated as outcomes of the procedure. Thus the severe test criterion must take this into account and be formulated as

> There is a very high probability that the test procedure would *not* pass the hypothesis it tests, given that the hypothesis is false. ([14], 202)

Of course the hypothesis H' constructed as just described does not meet this criterion. The procedure employed in formulating and passing such a hypothesis will pass whatever hypothesis it generates quite independently of the truth of that hypothesis.

This response does not solve the problem, which is independent of whether H' itself is severely tested. Neither is the problem one of the goodness or reliability of a testing procedure that generates H' as an outcome. The problem for the error-statistical account is that the mere existence of H' as a possibility prevents us from saying that condition *SR2* has been met.[5]

[5] This challenge to error-statistics was first articulated, in slightly different form and context, by John Roberts [20].

The error-statistician thus faces a dilemma. A weak reading of *SR2* will permit the conditions for an *SR* evidence claim to be sometimes met, but the resulting evidence concept will be too weak to escape inclusivism. That a test outcome constitutes *SR* evidence for H on such a reading would not rule an alternative not in $\{H_p\}$ receiving even stronger support from that test, were it to be included in the class of hypotheses under consideration. A strong reading of *SR2* will permit the reading of 'E is *SR* evidence for H' in a non-inclusive sense, but at the unreasonable cost of never being in a position to make an evidence claim at all.

I would like to consider two options for the error-statistican to escape this dilemma. The choice between these two options will rest largely on consideration of the *security* of inference, and so I turn first to an explanation of that concept.

4 Secure inference

I propose that we reason securely to the degree that our inferences or evidential judgments are stable under changes in our background information that are possible, given what we know (i.e., changes that are *epistemically possible*). The idea here concerns what happens as we learn new things. The aim of secure inference is not that our beliefs about how things are remain the same, but that we be able to continue to regard our past judgments about what should be concluded from the then-available data as having been correct, so that our acquisition of new information allows us to build on our past judgments rather than requiring us continually to start over.

Security of inference derives its importance largely from the fact that evidential judgments rest on fallible assumptions, such that the failure of those assumptions can defeat the judgments in question. For example, Galileo's opponents in his arguments for Copernicanism incorrectly held that perpendicular free fall is evidence that the earth is stationary. Newton held that the fortuitously-placed and nearly circular orbits of the planets is evidence that they were the result of choice on the part of a providential deity. Many of Darwin's critics incorrectly held that domestic breeding provided evidence against evolution by natural selection [19]. In each of these cases an evidence claim was predicated on a false empirical assumption, leading to an error that could only be corrected by discarding or replacing that assumption.

One of the most important applications of security is in understanding the value of robust evidence, i.e., evidence that rests on multiple independent tests. As I have explained in detail elsewhere [21], the advantage enjoyed by robust evidence is that the failure of an assumption that is not shared by all of the tests on which that result depends does not necessarily defeat (though it might weaken) the evidence claim itself. This protection against failure of background assumptions enhances the security of inferences from robust evidence.

Security of inference is not to be confused with reliability of inference, where reliability is to be thought of in terms of the rate at which a type of inference yields true beliefs. If I draw my inferences according to the standards *SR*, for example, then in virtue of the resulting low error rate of my procedure, I am following a reliable inference procedure (at least if we construe *SR2* in the strong sense). I might not be reasoning securely, however. My assessment of the error probabilities of my procedure might rest on very shaky grounds, such that many plausible scenarios regarding future relevant information would require that I give up my belief (I never actually calibrated my instruments, arbitrarily trusting that the factory settings would be adequate). That could be the case even though as a matter of fact the error probabilities of my procedure are really as low as I take them to be ("as chance would have it," the instruments are calibrated quite well right out of the box!). Thus, while reliability considerations are externalist in character, security incorporates a more internalist perspective, requiring the consideration of what future developments might take place given what the inquirer knows.

As important as security is, the more secure inference is not always the best kind of inference to draw. This will be important to keep in mind as we return to the error statistician's dilemma.

5 Relativization or implausibility judgment?

John Roberts, who first argued for the alternative hypothesis difficulty discussed above in the context of assessing "high-level" theories by severe testing, has proposed what he regards as a friendly amendment to the error statistical account. Roberts proposes that judgments about high-level theories in light of piecemeal severe testing be *relativized* to those assumptions that are not themselves severely tested, and are sufficiently strong to rule out competing hypotheses that threaten to nullify severity assessments for such tests. Such assumptions must meet certain constraints, however. Roberts proposes that we endorse the results of severe testing as it relates to high-level theories only when those test outcomes are relativized to assumptions that have been made *explicit*, that make possible the *measurement* of parameters that potentially describe nature, and that can be (albeit non-severely) *tested* (and hence possibly refuted) by means of multiple independent measurements of those same parameters.

Such an approach certainly could be adapted to the context of such low-level empirical claims as that in our toy example above. If one assumes that there is some definite and uniform probability that a randomly selected asthma sufferer will experience nausea subsequent to receiving treatment C, then the rate at which members of an appropriately gathered sample of C-treated asthma sufferers experiences nausea can be taken as a measurement of that probability. Furthermore, although that assumption may not itself be susceptible to *severe* testing, it can be tested in a weaker sense. If, for example, upon repeated sampling, we find that our estimate of the probability does not appear to be converging on any number, this might indicate that our working assumption is incorrect. Most importantly, of course, this assumption effectively cuts off our need to consider such exotic alternatives as H'.

Here I propose an alternative (also discussed in [24]) that does not rely on relativization. I will compare the the two approaches in section six.

What is needed is a way to reduce the class of hypotheses to be considered as alternatives without relativizing our evidential judgments. We certainly *could* begin to incorporate hypotheses such as H' into our testing models. Our testing procedures could remain useful for learning, so long as we did not use the data to construct these alternatives. But of course we know that if we brought in a hypothesis similar to H' as an alternative without using such a construction procedure, it would almost certainly fail any test to which we subjected it. There is no need to be worried about the mere existence of alternative hypotheses which are such that, were they to be subjected to any genuinely informative test, would be almost certain to fail.

The notion of an "informative" test here is meant to refer to something analogous to, but weaker than, a severe test. To say that a hypothesis H passes an informative test with a given result E means that, supposing *some* alternative hypothesis H' to be true, it would be highly improbable that H would pass the test with a result that fits as well as E. In effect, by eliminating from consideration hypotheses that we expect to fail any informative test we are eliminating only those that we antecedently judge would only pass tests that were "rigged" in advance to favor those hypotheses (like our gellerized example H').

In short, my proposal (*IJ* for "implausibility judgment") is that we exclude from our class of relevant hypotheses those which we antecedently judge to be nearly certain to fail any informative test to which they are subjected. We can then take the 'H is false' locution in *SR2* as a covering term for all of the incompatible alternatives to H exclusive of those antecedently judged to be (in this sense) too implausible to consider. There does not seem to be any general basis for skepticism regarding the possibility of satisfying *SR2* conceived in this way.

Two points regarding the *IJ* approach to alternative hypothesis objections are worth emphasizing.

First, evidential relations remain objective on the present account in the sense that they obtain or fail to obtain independently of our beliefs about them. The role played here by plausibility judgments is not like the role of prior probabilities in Bayesian accounts. An investigator's judgment that H' is not a plausible alternative hypothesis in the sense that it would almost certainly not pass any informative testing procedure is a fallible background assumption in the judgment that H has passed a severe test. This assumption, and hence the severity assessment made based on it, may be mistaken. This feature distinguishes such judgments from prior probability assessments in personalist Bayesian approaches.

Second, these judgments are empirical, making their status different from that of prior probability judgments in logical probability approaches. Although just how such empirical judgments are made needs to be worked out in more detail than I will here undertake, let me just sketch how such an account might go.

Our determination that a hypothesis regarding a particular phenomenon is not a legitimate alternative is based largely on our knowledge of the kinds of patterns of behavior found in other natural phenomena. Such reasoning may be largely analogical in character. In our example, we exclude H' from consideration because it in effect postulates a kind of “conspiracy of nature” in which the outcome of treatment in a given case correlates with the attachment of a certain index to the data in an arbitrary labeling scheme. Our previous experience of data collection across a broad range of phenomena provides us with ample reason to believe that nature is indifferent to our schemes for data-labelling. Although this very general proposition does not seem susceptible to severe testing (what are the probability distributions determined by its denial?), to the extent the present case resembles other cases in which data-labelling has failed to correlate with the phenomenon under study, we can expect an analogous indifference to obtain.

What both the relativization proposal and *IJ* share is that they require, prior to the drawing of inferences from the results of severe testing, an additional step that cannot in general be made on the basis of severe testing. The conclusions that the two approaches permit one to draw, however, are quite different in character, as can be seen most clearly by considering the consequences of the relevant assumptions being overturned by subsequent findings. To explain and to emphasize the potential import of these considerations, let me introduce a less artificial application: the use of parametric frameworks for testing some fundamental principles of physics.

6 Theories of Gravity

6.1 The PPN formalism

Our first example of such an enterprise comes from the experimental investigation of theories of gravity through the Parametrized Post-Newtonian (PPN) framework, an example discussed already by both Mayo and Roberts [15] [16] [20].

The PPN formalism was developed to enable the comparison of metric theories of gravity with each other and with the outcomes of experiment, at least insofar as those theories are considered in the slow-motion, weak-field limit. Metric theories of gravity can be characterized by three postulates:

1. spacetime is endowed with a metric **g**,
2. the world lines of test bodies are geodesics of that metric, and
3. in local freely falling frames (Lorentz frames) the nongravitational laws of physics are those of special relativity. ([26], 22)

The ability to compare such theories is facilitated by using a common framework for writing out the metric **g** as an expansion, such that different theories are manifested by their differing values for the constants used in the expansion. As Clifford Will writes, "The only way that one metric theory differs from another is in the numerical values of the coefficients that appear in front of the metric potentials. The [PPN] formalism inserts parameters in place of these coefficients, parameters whose values depend on the theory under study" ([27], 29).

Crucial to the issues at hand is the fact that the PPN framework only encompasses metric theories of gravity. Such theories, which treat gravity as a manifestation of curved spacetime, satisfy the Einstein Equivalence Principle (*EEP*). *EEP* is equivalent to the conjunction of three apparently distinct principles — Local Position Invariance (*LPI*), Local Lorentz Invariance (*LLI*) and the Weak Equivalence Principle (*WEP*).

WEP holds that "if an uncharged test body is placed at an initial event in spacetime and given an initial velocity there, then its subsequent trajectory will be independent of its internal structure and composition" ([26], 22)[6]. According to *LLI*, the outcome of any "local nongravitational test

[6] The test body in question must have negligible self-gravitational energy, according to Newtonian gravitational theory, and negligible coupling to inhomogeneities in any external

experiment" is independent of the velocity of the experimental apparatus, and *LPI* states that the outcome of any such experiment is independent of its spacetime location. Here a "local nongravitational test experiment" is understood to be an experiment in a freely falling laboratory shielded and small enough to render inhomogeneties in external fields negligible throughout its volume and in which self-gravitational effects are negligible ([26], 22).

Mayo's account emphasizes the positive role played by the PPN framework in facilitating, not only the comparison of existing theories, but also the construction of new alternatives ("straw men" in Will's phrase) as a means of probing the various ways in which General Relativity (GR) could be in error. In addition, she argues that the resulting proliferation of alternatives to GR was not a manifestation of a theory in "crisis," but rather of an exciting new ability to probe gravitational phenomena and prevent the premature acceptance of GR. She claims various advantages for her account over the approaches of Bayesians and "comparativists." A key to these advantages, it seems, is the way in which the PPN formalism allows for the combination of the results of piecemeal hypothesis tests, not only to show that some possibilities have been eliminated, but to indicate in a positive sense the extent to which gravitation is a phenomenon that GR (or theories similar to GR) gets, in some respects, right: "By getting increasingly accurate estimates, more severe constraints are placed on how far theories can differ from [GR], in the respects probed" [16]. Note here how the results of this investigation are taken to give us more than merely comparative conclusions. Mayo is

fields

committed to being able to say more, on the basis of the outcomes of such tests, than that certain possible theories have been refuted.[7]

As John Roberts argues [20], Mayo's approach does not quite work in the way that she would like. While the "squeezing" of "theory-space" can be brought about by combined piecemeal tests as Mayo claims, the space that is squeezed is not the space of all possible theories of gravity, or even of all theories of gravity that have been actually formulated. It is only the space of all *metric* theories of gravity, i.e., those satisfying *EEP*. Nonmetric theories are certainly possible, and some have been proposed (though none so far that are compatible with empirical results).[8] Non-metric theories as a class could be ruled out on error-statistical grounds, according to Roberts, only if we could carry out a severe test of the Einstein Equivalence Principle (*EEP*). Such a test would require at a minimum a severe test of *WEP*.[9]

However, this is not possible. *WEP* quantifies over all spacetime and all bodies of a certain kind. The principle could be violated either in regions of spacetime remote from ours or by kinds of matter that have not yet been tested. For precisely this reason, the very high precision with which some of the PPN parameters have been measured is not, without further

[7] Another noteworthy discussion of the PPN framework is due to William Harper, who assimilates it to an extension of the methodology developed by Newon in the *Principia*, in which the standard of empirical success incorporates the accurate, theory-mediated measurement of parameters of the theory by predicted data, and in which theoretical propositions are accepted as a guide to future research [8] [9].

[8] One such theory, discussed by Lightman and Lee [12] as well as Will [26] is the Belinfante-Swihart theory [4] [5] [6].

[9] If Schiff's conjecture that *EEP* is equivalent to *WEP* is true, then it would also require no more than severely testing *WEP*.

assumptions, equivalent to the high severity (in the strong sense of *SR2*) with which corresponding hypotheses about the values of those parameters have been tested.

6.2 Applying the two approaches

Both relativization and *IJ* seek to close the gap between theory and experiment so that the results of severe tests can be made relevant to the theory under consideration. But the two approaches give quite different answers to the question of what is learned from such testing.

The relativization approach endorses the *EEP* assumption on the grounds that doing so allows for the measurement by multiple, independent means of parameters that only have a meaning *within* metric theories of gravity, and that it is susceptible to being, though it has not been, shown to be in error. What we then conclude from the results of testing is that *relative to* this assumption, numerous sources of experimental data provide evidence that the values of those parameters that characterize gravitational phenomena differ very little, if at all, from the values assigned by *GR* (call the latter claim *GRP*).

If on the other hand the *IJ* approach is followed, then these same results become evidence for *GRP* not merely relative to *EEP*, but *in fact*, unless our assumption that *EEP* holds to the needed degree of accuracy turns out to be incorrect. That is not a trivial worry, either. In effect, on the *IJ* proposal, one would take the all of the previous tests of *EEP* on various physical systems in various locations as a basis for holding it to be implausible that types of physical systems not yet tested and physical systems in locations not yet tested would behave differently with regard to *EEP*. This is, of course, a substantive empirical assumption. It cannot be severely tested, but insofar as it is reasonable to regard untested systems and locations as analogous to those that have been tested, it will be a reasonable assumption. (Physicists are presently engaged in programs of testing *EEP* that employ parametric frameworks similar to PPN. See the appendix for a brief discussion.)

The advantage of the relativization approach is clear: our judgments of relativized evidence enjoy a degree of security not shared by the kind of unrelativized claim emerging from *IJ*. Even if *EEP* turns out to be false, it can (and will, assuming we analyzed the data correctly and understand correctly the relevant characteristics of our testing procedures) remain nonetheless true that *relative to EEP*, the data constitute evidence for GRP. In the *IJ* approach, learning that *EEP* is false amounts to having to give up our claim that the data constitute evidence for GRP.

Security, however, is not the only value to consider here. The motivation for developing a non-inclusive account of scientific evidence was to allow us to develop a strong concept of evidence that would be suitable for serving as the basis of scientific inference. That is, we wanted to be able to assert that when E is evidence for H, then one can reasonably infer from E to the truth of H. We can therefore ask whether relativized *SR* evidence claims meet this criterion.

If we apply our definition of inclusivism above to relativized *SR* evidence claims, we discover an ambiguity in our definition. On the one hand it is true that if conditions *SR* are met by result E of test T, relative to assumptions B, then it will not be the case that E will support an incompatible hypothesis

H' relative to B. On the other hand, E might well support an incompatible alternative H' relative another set of assumptions B'.

As a consequence, given that E is *SR* evidence for H relative to B, it will be reasonable to infer from E to H only insofar as it is reasonable to believe B to be true. But notice that on the relativization account at hand, the basis for relativizing to B is not directly connected to it being reasonable to think that B is true, but rather is connected to the usefulness of B in permitting the measurement of parameters, to the testability of B, and to B's not having been falsified. Thus it can be reasonable on the relativization proposal to employ assumptions that underwrite an evidence claim, without that evidence claim supporting a reasonable inference.

The *IJ* proposal imposes on the assumptions behind our evidential judgments a restriction directly related to truth. The price to be paid is that our evidential judgments are subject to a kind of defeat that is avoided by the relativization approach. Because future developments that are perfectly compatible with what we presently know could defeat, for example, our assumption that *EEP* holds to the necessary degree of accuracy, our inferences based on that assumption are rendered to a degree insecure.

Such an outcome should not be shocking, however. Compare the kind of insecurity that holds for the evidential status of *GR*, for example, with that involved in the low level hypothesis assessment discussed in our original example. The working assumption in the low level case amounted to little more than the assumption that nature does not conspire against us. Although it is *possible* that nature could somehow conspire against us, it is reasonable to assume that this is not the case. But claims about the nature of gravity are bound to be attended by greater epistemic risk, and on the current proposal that risk enters through the stronger assumptions that must be made to enable evidential judgments to be drawn.

7 Conclusion

Any given error-statistical inference will rely on a material postulate concerning the error-rates, or severity, of the testing procedure behind that inference. If, however, the error-statistical approach is to provide a concept of support that, by escaping inclusivism, provides a good basis for a theory of scientific inference, then an additional material postulate will be needed. This second postulate will need to provide the basis for taking the severity postulate in a strong sense.

Here I have considered two different *ways* in which this second postulate might enter our reasoning: either as a basis for relativized evidence claims or as a claim to the effect that certain possibilities do not deserve consideration. Both approaches bring their own problems. The choice ultimately must be referred to the aims of inquiry. If we suppose that the ability to make measurements and be explicit in the assumptions underlying such measurements constitute sufficient aims for scientific inquiry, then relativization constitutes a more secure way to pursue those aims. On the other hand, if we suppose that scientific inquiry is aimed at discovering what it is reasonable for us to believe, then relativized evidence claims seem insufficiently strong, and we will have to accept the epistemically risky nature of the scientific enterprise.

8 Appendix: From PPN to $TH\epsilon\mu$

There is another formalism that has been developed to systematize the search for violations of *EEP* that functions analogously to the PPN framework for tests of GR. This formalism, dubbed $TH\epsilon\mu$, was first developed by Lightman and Lee [12] for purposes of proving Schiff's conjecture for a restricted class of theories. The class of theories that can be described within the $TH\epsilon\mu$ formalism includes all metric theories. It also includes many, but not all, non-metric theories.[10] The ability to put non-metric theories into a common framework such that limitations can be put on *EEP* violations in a systematic way provides a powerful extension of the program of testing within PPN. However, just as PPN is limited by its exclusion of non-metric theories, $TH\epsilon\mu$ is limited by including only some non-metric theories. It is precisely for this reason that Schiff's conjecture is still called a conjecture.[11]

[10] The restriction, more specifically, is to theories that describe the center-of-mass acceleration of an electromagnetic test body (effects from weak and strong forces are neglected) in a static, spherically symmetric (SSS) gravitational field, such that the dynamics for particle motion is derivable from a Lagrangian. The parameters T and H appear in the Lagrangian; ϵ and μ appear in the "gravitationally modified Maxwell equations" (GMM). Lightman and Lee argue (in 1973) that "all theories we know of" have GMM equations of the type needed, and that all but one theory (which they treat separately) can be represented in terms of the appropriate Lagrangian, although this may require (as in the case of Belinfante-Swihart theory) a "reformulation" of the theory [12].

[11] It is noteworthy that Lightman and Lee, in introducing the formalism, express skepticism about the possibility of an unrestricted proof of Schiff's conjecture precisely because doing so would require a "moderately deep understanding" of all theories of gravity satisfying *WEP*, "including theories not yet invented" (ibid., 364). Such epistemic modesty with regard to "all possible" claims is central to the error-statistical emphasis on "learning

$TH\epsilon\mu$ focuses on the behavior of charged particles in an external static spherically symmetric gravitational field with potential U. The motion of charged particles in this external field is described by two arbitrary functions $T(U)$ and $H(U)$, while $\epsilon(U)$ and $\mu(U)$ describe the response of the electromagnetic fields to U. The following identity is satisifed by every metric theory:

$$\epsilon = \mu = (H/T)^{1/2} \tag{1}$$

for any U.

This formalism has proven to be adaptable to the pursuit of tests of null hypotheses for each of the components of *EEP*. By taking various combinations of the four $TH\epsilon\mu$ parameters, one can define three "non-metric parameters," Γ_0, Λ_0, and Υ_0, such that if *EEP* is satisfied then $\Gamma_0 = \Lambda_0 = \Upsilon_0 = 0$ everywhere.

Tests of the components of *EEP* can then be investigated in terms of null tests for these parameters. A non-zero value for Υ_0 is a sign, for example, of a failure of *LLI*. Will describes how the results of the Hughes-Drever experiment ("the most precise null experiment ever performed" [26], 31) can be analyzed so as to yield an upper bound of $\Upsilon_0 < 10^{-13}$ and concludes that "to within at least a part in 10^{13}, Local Lorentz Invariance is valid" (ibid., 62).

The point made previously about the PPN formalism applies here as well. To regard such tests as showing (by means of severe testing) that *LLI* must be valid to within the cited accuracy, we must rely on some plausibility assumptions.

about" rather than conclusively establishing general and fundamental theories in physics.

We should first note that, just as for the case of low-level hypotheses, one can trivially construct "conspiracies of nature" that will predict such experimental outcomes in a way that is compatible with the failure of the principle being tested. In particular, one could explicitly introduce terms into the Lagrangian that yield two arbitrarily large violations of LLI that are equal (or very nearly so) but opposite in sign. There is even precedent in physics for such theories, in the sense that theory has sometimes *required* such "fine-tuned" balancing of two oppositely signed contributions to yield a very small quantity [17] [23]. In any case, at least the same kind of "no-conspiracy" assumptions that are called for in low-level hypothesis testing will be needed here.

More substantively, recall that the $TH\epsilon\mu$ formalism, like the PPN formalism, can only be applied to a restricted class of theories (although this class is less restricted than that of PPN). Thus the analysis that allows for the limit in question to be generated does require that we assume that the correct theory of phenomena in the weak-field, slow-motion limit is among those theories. This assumption is weaker than the assumption of EEP (or of LLI), which is needed for the application of the PPN formalism. Nonetheless, just as with *EEP*, It is very unclear just how such an assumption could itself be subjected to a severe test. On the present account, this assumption need not pass a severe test in order to be reasonable.

Returning to the example of Lorentz invariance (LLI), the Hughes-Drever experiment cited above is an example of a "clock comparison" experiment. In the version performed by Drever at Glasgow University [7], the "clock" was constituted by the transition frequencies of the $J = 3/2$ ground state of the ^{7}Li nucleus in an external magnetic field. The magnetic field introduces a splitting of the ground state into four levels. Any perturbation introduced by a preferred direction in space would result in a further splitting, resulting in an inequality of the spacing between the lines. Similar experiments have been performed on numerous systems subsequently; none have uncovered any signs of Lorentz violation. Clock comparison tests are just one of a growing variety of tests of *LLI*, each of which is specific to a particular type of matter-energy. Mattingly gives a helpful review of a vast number of such tests in [13].[12]

In concluding his review, Mattingly notes that "over the last decade or two a tremendous amount of progress has been made in tests of Lorentz invariance. Currently, we have no experimental evidence that Lorentz symmetry is not an exact symmetry in nature," and asks, "When have we tested enough?" Without quite answering that question, he notes the difficulty of fitting any Lorentz-violating terms into existing field theories consistently with experiment and concludes that "It therefore seems hard to believe that Lorentz invariance could be violated in a simple way." [13]

Where does this leave us with respect to the status of PPN tests of gravity? Considering only *LLI* and neglecting the other components of *EEP* (and acknowledging that an actual argument for my claim would require

[12] Mattingly discusses these results and others in the context of yet another (!) formalism, the Standard Model Extension (SME) [11] , that is even broader than $TH\epsilon\mu$ and that is useful for, among other things, systematizing the testing of *LLI*. Bailey and Kostelecký [3] discuss how the SME can be applied to the gravitational sector, noting that the phenomena that can be described in the PPN and the SME are overlapping, but distinct. Each has its blind spots.

a much more detailed discussion of the existing experimental situation), it seems that, although there are *possible* ways that *LLI* (and hence *EEP*) could fail, these plausibly fall into the following three categories: (1) conspiracies of nature, (2) violations involving forms of matter not yet tested, and (3) phenomena outside the scope for which the PPN approach claims validity.

It is the second category that is the most troubling for the error-statistical approach, and which distinguishes the alternative-hypothesis worries for low-level from those for high-level hypotheses on that account. In both contexts, I have argued, error-statistical assessment gets under way only *after* we assume that nature does not conspire against us. But the kind of universality involved in a principle such as *LLI* demands a stronger assumption before we can hope to invoke severe tests on behalf of the principle. Nonetheless, I believe that such assumptions can not only be made, but can be justified, even if doing so does involve some risk. So, for example, clock comparison tests have only been made using first-generation matter (up and down quarks and electrons). No such test (to my knowledge) has been carried out using second- or third-generation matter (such as muon or tau leptons, charm, strange, bottom, or top quarks). However, there are good plausibility arguments for expecting such tests, were they to be carried out, to fall in line with the results on first-generation matter, since the known physics for all three generations is essentially the same. Still (and here is where the risk lies), no one knows *why* more than one generation of matter exists, and if we did understand the answer to that question, it is at least possible that we would have a reason to think that there would be a difference in their adherence to Lorentz invariance.

Finally, the third category is by far the more interesting as far as physics is concerned, but is no embarrassment from the standpoint of the error-statistical account of theory assessment defended here. Indeed, much of the testing of *LLI*, and of *EEP* more generally is directed not so much at establishing greater support for those principles, but in the active search for the manner in which they might fail, as such failures, should they be found, are among our current best hopes for developing the fundamental physics that we do not yet possess.

References

[1] Achinstein, Peter. “Why Philosophical Theories of Evidence Are (and Ought to Be) Ignored by Scientists,” *Philosophy of Science* 67(Supplement) (2000): S18092.

[2] Achinstein, Peter. *The Book of Evidence* (New York: Oxford University Press, 2001).

[3] Bailey, Quentin and V. Alan Kostelecký. “Signals for Lorentz Violation in Post-Newtonian Gravity,” (2006). URL (cited on May 20, 2006): `http://lanl.arxiv.org/abs/gr-qc/0603030`.

[4] Belinfante, F. and J. Swihart. “Phenomenological Linear Theory of Gravitation. I. Classical Mechanics,” *Annals of Physics* 1 (1957): 168–95.

[5] Belinfante, F. and J. Swihart. "Phenomenological Linear Theory of Gravitation. II. Interaction with Maxwell Field," *Annals of Physics* 1 (1957): 196–212.

[6] Belinfante, F. and J. Swihart. "Phenomenological Linear Theory of Gravitation. III. Interaction with Spinning Electron," *Annals of Physics* 2 (1957): 81–99.

[7] Drever, R.W.P. "A Search for Anisotropy of Inertial Mass Using a Free Precession Technique," *Philosophical Magazine* 6 (1961): 683–87.

[8] Harper, William. "Newton's Argument for Universal Gravitation," in I. Bernard Cohen and George E. Smith (eds.) *The Cambridge Companion to Newton*, (Cambridge: Cambridge University Press, 2002), pp. 174–201.

[9] Harper, William. "Newtons Methodology and Mercurys Perihelion before and after Einstein." Paper presented at PSA 2006, Vancouver, British Columbia, November 2006.

[10] Hitchcock, Christopher and Elliott Sober. "Prediction Versus Accommodation and the Risk of Overfitting," *British Journal for the Philosophy of Science* 55 (2004): 1–34.

[11] Kostelecký, V. Alan. "Gravity, Lorentz Violation, and the Standard Model," *Physical Review D* 69 (2004): 105009. Related online version URL (cited May 20, 2006): `http://arxiv.org/abs/hep-th/0312310`.

[12] Lightman, Alan and David Lee. "Restricted Proof that the Weak Equivalence Principle Implies the Einstein Equivalence Principle," *Physical Review D* 8 (1973): 364–76.

[13] Mattingly, David. "Modern Tests of Lorentz Invariance," *Living Reviews of Relativity* 8 (2005): 5. URL (cited on May 20, 2006): `http://www.livingreviews.org/lrr-2005-5`.

[14] Mayo, Deborah. *Error and the Growth of Experimental Knowledge* (Chicago: University of Chicago Press, 1996).

[15] Mayo, Deborah. "Theory Testing, Statistical Methodology, and the Growth of Experimental Knowledge," in P. Gärdenfors, J. Wolinski, and K. Kijania-Placek (eds.) *In the Scope of Logic, Methodology, and Philosophy of Science*, (Dordrecht: Kluwer, 2002), pp. 171–90.

[16] Mayo, Deborah. "Learning from Error: Severe Testing and the Growth of Theoretical Knowledge." Unpublished ms (2006).

[17] Murayama, Hitoshi. "Supersymmetry," in Shoichi Yamada, ed., *Physics with High Energy Colliders: Proceedings of 22nd INS International Symposium* (Singapore, World Scientific, 1995).

[18] Norton, John. "A Material Theory of Induction," *Philosophy of Science* 70 (2003): 647–70.

[19] Richards, Richard. "Darwin and the Inefficacy of Artificial Selection," *Studies in the History and Philosophy of Science* 28 (1997):75–97.

[20] Roberts, John. “Coping With Severe Test Anxiety: Problems and Prospects for an Error-Statistical Approach to the Testing of High-Level Theories.” Unpublished ms (2006).

[21] Staley, Kent. “Robust Evidence and Secure Evidence Claims.” *Philosophy of Science*, 71 (2004): 467–88.

[22] Staley, Kent. “Agency and Objectivity in the Search for the Top Quark.” In Peter Achinstein (ed.), *Scientific Evidence: Philosophical Theories and Applications.* Baltimore: Johns Hopkins University Press, 2005, pp. 165–84.

[23] Staley, Kent. “Anti-matter, Supersymmetry, and God: On Fine-tuning and Scientific Inquiry,” unpublished ms (2006).

[24] Staley, Kent. “Error-statistical Theory Assessment and Alternative Hypothesis Problems: A Role for Judgments of Plausibility?” Unpublished ms. `http://www.error06.econ.vt.edu/Staley.pdf`.

[25] Steel, Daniel. “The Facts of the Matter: A Discussion of Norton’s Material Theory of Induction,” *Philosophy of Science* 72 (2005): 188–97.

[26] Will, Clifford. *Theory and Experiment in Gravitational Physics* (New York: Cambridge University Press, 1993).

[27] Will, Clifford. “The Confrontation between General Relativity and Experiment,” *Living Reviews in Relativity* 9 (2006): 3. URL (cited on May 20, 2005):
`http://www.livingreviews.org/lrr-2006-3.`