

Error-statistical Theory Assessment and Alternative Hypothesis Problems: A Role for Judgments of Plausibility?

Kent Staley

Saint Louis University

staleykw@slu.edu

May 22, 2006

Draft only – do not circulate

1 Introduction

Deborah Mayo's error-statistical theory of evidence provides considerable insight into the way in which scientists, by paying careful attention to potential sources of error in deriving conclusions from their data, can learn in a piecemeal way about specific experimental phenomena, allowing them to pursue the discovery of significant new facts while avoiding the embarrassment of publicly mistaking illusions for facts.

Nonetheless, even friendly critics have claimed that Mayo’s account faces a difficulty when it comes to talking about the context in which those experimental facts often take on the greatest significance: very general, or “high level” theories [6][10]. Whereas low-level hypotheses of the sort exemplified in many passages of Mayo’s *Error and the Growth of Experimental Knowledge* can typically be represented quite directly by means of the kind of probability distribution over possible experimental testing outcomes that form the core of the error-statistical apparatus, such can not typically be said for high-level theories such as Quantum Electrodynamics and the General Theory of Relativity (GR).

I contend that the key to understanding how the error-statistical approach can apply to high level theories lies in understanding the problem of alternative hypothesis objections generally. Thus, I begin with a discussion of Mayo’s treatment of alternative hypothesis objections. Drawing upon an argument from John Roberts, I show how Mayo’s discussion fails to vindicate the error-statistical treatment of alternative hypotheses. I then offer my own friendly amendment that will facilitate the error-statistical treatment of alternative hypotheses. I then consider the problem of high-level theories, using the example of theories of gravitation to illustrate how Mayo’s proposal for using error-statistical learning for high level theories encounters alternative hypothesis objections. Not only does my friendly amendment provide the resources to reply appropriately to those objections, it has the added virtue of conceptually unifying the treatment of low- and high-level hypotheses while also explaining why it is typically appropriate to be much

more cautious in our claims regarding even very successful high-level theories than in our commitment to well-tested low-level hypotheses.

2 Alternative Hypothesis Problems

Central to Mayo's account of the the error-statistical theory of evidence is a framework for dealing with alternative hypothesis objections. Before we consider the case of high-level theories in physics, it will be worthwhile examining how this difficulty is dealt with in the case of the kind of low-level piecemeal testing that is at the heart of Mayo's approach.

Consider a general kind of situation: We wish to learn whether there is a correlation between the application of some treatment to a population and a particular outcome pertaining to individual members of that population. This describes, for example, tests pertaining to the effectiveness of a drug at treating an illness, the capacity of a fertilizer to increase crop yields, or the potential for an educational innovation to produce greater better learning outcomes among students. An error-statistical approach to this kind of inquiry will typically consist of an attempt to determine whether there is a statistically significant different in the incidence of the outcome between a control and treatment sample.

Supposing that such a statistically significant difference is found, the procedure might then proceed from the determination of the existence of a correlation to the estimation of the strength of that correlation. Such a procedure might generate then a hypothesis of the following form H : Individuals of type

B subjected to the treatment C have a probability p ($\pm\delta p$) of exhibiting the outcome O .

For such a hypothesis to be supported by the evidence E , under the error-statistical theory, H must be subjected to, and pass, a severe test with the result E . To meet this criterion the following conditions must be met:

- E fits H ,
- 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 ([13], esp. 178–87).

How these criteria should be understood is put into clearer light if we consider how Mayo proposes to deal with alternative hypothesis objections. The critic of error-statistics alleges that, since for any given data there is always an alternative hypothesis H' that fits the data at least as well as any hypothesis for which support is being claimed, evidence alone cannot single out any hypothesis as the best supported.

Mayo denies this claim. Her first point is that, because the severe test requirement demands more than just fitting a hypothesis, the fact that another hypothesis fits E as well as H does fails to establish that another hypothesis is equally well supported. Of course, this is insufficient by itself, for the critic could grant this point, but maintain that the price paid for this success is that *no* hypothesis then can be said to pass any severe test. Even though the alternative hypothesis H' may not *itself* pass any severe test with E , its mere existence suffices to establish that we can never assert that, supposing

H to be false, a result such as E would be improbable. For the data-fitting alternative H' renders E probable, and entails that H is false. Thus there is at least one scenario in which H is false, yet an outcome such as E is not improbable.

Returning to our example, the kind of alternative hypothesis that is most threatening is that which fits the data just as well as or better than our H , but takes a different mathematical form than H . An example of this kind discussed by Mayo is the “maximally likely hypothesis.” Suppose that our data comprise a simple listing of individuals surveyed by our investigation, where for each individual $a_i, i = 1, \dots, n$, 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 (Oa_i or $\neg Oa_i$). With our data in hand, we can construct a maximally likely alternative hypothesis simply by collecting together all m of the individuals for which Oa_i obtains. We then relabel these individuals a_j , where $j = 2i + 1$. For all of the individuals for which $\neg Oa_i$ 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 of type B that are labeled with odd indices that are subjected to the treatment C have a probability $p = 1$ of exhibiting the outcome O . Those with even numbered indices have a probability $p = 0$ of exhibiting the outcome O .

H' is of course an absurd hypothesis produced by what Mayo calls “gel-erization.” Nonetheless it would seem that, given the trivial existence of

hypotheses such as H' , no H -type 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

SC with hypothesis construction: There is a very high probability that the test procedure would *not* pass the hypothesis it tests, given that the hypothesis is false. ([13], 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 however, which as mentioned above, does not concern the possibility that the alternative hypothesis is also severely tested. Neither is the problem one of the goodness or reliability of a testing procedure that generates H' as an outcome. The source of the problem for the severe-testing account is the mere existence of H' as a possibility, which, it is claimed, prevents us from saying that the severe test requirement has been met.

Our objector is pursuing a line of argument that has been presented most

cogently by John Roberts, although Roberts makes the argument specifically for the case of high-level theories. He considers the difficulty to arise from the fact that “high -level theories have logically possible alternatives that fit all the existing data equally well.” Consequently, whenever we might wish to support theory T with the results of a severe test, we find that we can always come up with a theory T' that fits the same data equally well. Hence, if T' were true and T false, it *would* be probable that we would get results that fit T as well as our actual results do. As a result, “in the case of competing high-level theories, if each fits the extant data equally well, then the two are tested with equal severity — specifically, without any severity at all” [17].¹

This would be a devastating setback for the error-statistical approach were there no way to refute the accusation or fix the problem without abandoning the central commitments of the error-statistical way. Roberts proposes to take the latter approach by advocating 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 assump-

¹Alan Chalmers goes further and argues that theories in general cannot pass severe tests as Mayo conceives them [7]. I believe the approach I here defend avoids the problems posed by both Roberts’ and Chalmers’ argument, although I will not argue for the latter claim.

tions that have been made *explicit*, that make possible the *measurement* of parameters that potentially describe nature, and that can be (non-severely) *tested* (and hence possibly refuted) by means of multiple independent measurements of those same parameters.

Although Roberts intends this as a friendly amendment to the error-statistical account of theory testing, an approach that avoids relativization might fit better with the aims of the error-statistical account, at least as Mayo has expressed them. Consider two reasons that support this judgment.

First, although Roberts’s proposal puts constraints on what relativizations we should find “impressive,” the fact remains that such relativized evidential assessments are themselves insulated in a way from error. All else being equal, if I judge that theory T has been severely probed (in some sense) relative to assumption A, but then learn that A is false, my initial relativized assessment has not itself been shown to be in error, but only my judgment of how impressive this fact is. Central, however, to the spirit of the error-statistical approach is *corrigibility*.²

Second, as Alan Chalmers notes with regard to Mayo’s rejection of a comparativist approach [7], assessments based on severe tests need to be relatively stable, so that error-statistics can achieve its aim of accounting for the *accumulation* of experimental knowledge. Relativizing such judgments without providing for sufficient stability of the basis on which such relativizations rest threatens to leave scientific knowledge without a firm ground for cumu-

²A similar point is made in Mayo’s approving quotation of Kyburg’s remarks about the incorrigibility of Bayesian prior distributions ([13], 83).

lative growth. These two considerations provide some reason to consider the prospects for a non-relativized approach.

To complete the treatment of the alternative hypothesis problem, we have to go further and scrutinize the way in which the objector has posed the problem. Mayo's discussion holds a helpful clue. She notes that

Within an experimental testing model, the falsity of a primary hypothesis H takes on a specific meaning. If H states that a parameter is greater than some value c , not- H states that it is less than c ; if H states that factor x is responsible for at least p percent of an effect, not- H states that it is responsible for less than p percent ([13], 190)

Encouraged in part by Mayo's own formulation of the severity criterion, we have been supposing that the bare fact of H' being a logical possibility that entails the denial of H suffices to render H' a legitimate alternative hypothesis.³ Apparently Mayo rejects (or at least should reject) this supposition. What is needed is a basis for such a rejection. Based on the passage just quoted, I'd like to consider two possibilities.

One possibility is that error-statistics restricts legitimate alternative hypotheses to those that have been explicitly formulated within a testing model.

³Roberts clearly – and understandably – understands the severity criterion in just this way : “the *logical possibility* of T' , and its incompatibility with T , guarantees that any exhaustive list of alternatives to T will include one that is consistent with T' , and on that alternative, the error probability will not be low” (ibid., emphasis added).

Such an approach certainly allows us to avoid the problem of gellerized alternative hypotheses. We simply do not include these kinds of hypotheses in our testing models. However, this solution succeeds all too well. If we have a favored hypothesis that we would like to see succeed, we can greatly increase its prospects by limiting the alternative hypotheses that we test it against. Such a solution leads to the unwelcome result that we can avoid being worried about alternative hypotheses by avoiding thinking about them.

The better way to extend Mayo's idea is to incorporate plausibility judgments into our account of learning through experimental testing. 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.

A few points regarding this 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.⁴ The role played here by plausibility judgments is not like the role

⁴However, it remains possible that an investigator's beliefs might be *causally* relevant

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. 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.⁵ Thus their status differs from that of prior probability judgments in logical probability approaches.

Finally, plausibility judgments, as I conceive them, do not result in an assignment of probabilities of any kind. They are judgments about the extent to which our empirical information in the broad sense gives us reason to consider a hypothesis as being of a kind that is sometimes accurately descriptive of some aspect of the world.

Clearly, much more work is needed to develop the present proposal and to consider in particular the kinds of reasoning employed in arriving at such plausibility judgments. For the moment, however, I want to explore a point insofar as they influence the error probabilities of a testing procedure — a situation experimenters are at pains to avoid (see [18], ch. 7).

⁵Such reasoning may be largely analogical in character — a more exact characterization remains for me to provide. I will not do so in this paper.

tential benefit of developing the error-statistical framework in this way by showing how it might help us to see in what ways high-level theories in physics can, and cannot, be supported by the outcomes of severe tests.

3 Theories of Gravity

Mayo herself has taken the experimental investigation of theories of gravity as an example of how she sees error statistical considerations applying to high-level theories [14] [15], and Roberts has developed his critique in the context of this same example. In effect, her approach, following up on suggestive comments in *EGET* ([13], 191), relies on combining the results from individual “piecemeal” tests of parametric hypotheses, so that, from a large class of gravitational theories one can eliminate all but those whose parameters take values lying within certain intervals.

The background to this way of approaching the problem lies in the Parametrized Post-Newtonian (PPN) framework.

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 \mathbf{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. ([20], 22)

The ability to compare such theories is facilitated by using a common framework for writing out the metric \mathbf{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” ([21], 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” ([20], 22)⁶. According to LLI, the outcome of any “local nongravitational test experiment” is independent of the velocity of the experimental apparatus, and LPI states that the outcome of any such experiment is independent of its space-

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

time 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 ([20], 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.” [15]

The ability to draw positive conclusions about gravitational phenomena from the results of such tests is crucial for Mayo’s account, insofar as she wishes to distinguish her approach from a strict falsificationism. She is committed to being able to say more, on the basis of the outcome of a severe

test, than that certain possible theories have been refuted.

As John Roberts argues persuasively, 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).⁷ 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 (if Schiff’s conjecture that EEP is equivalent to WEP is true, then it would also require no more than severely testing WEP).

However, this is not possible, he claims, because to do so would require a severe test of exactly the kind of “high level” theoretical claim that he has argued cannot be severely tested. This is directly at odds with Mayo’s account in which, drawing upon comments from Will, “This principle [WEP] is inferred with severity by passing a series of null hypotheses (e.g., Eötvös experiments) that assert a zero difference in the accelerations of two differently composed bodies.” This severity assessment is warranted in turn by the “high precision with which these null hypotheses passed” ([15], 27).

⁷One such theory, discussed by Lightman and Lee [11] as well as Will [20] is the Belinfante-Swihart theory [3] [4] [5].

The important point about this in Roberts’s argument is that 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. These possibilities, Roberts notes, “are not mere philosophers’ hoked-up skeptical scenarios. They could each be consequences of general, fundamental physical theories that do not differ radically in form from actual fundamental physical theories” ([17],13). It is worth adding to this that “high precision” in a test of a hypothesis, although desirable, is not equivalent to severity of the test with respect to that hypothesis.

More specifically, while granting that some Eötvös experiments have indeed yielded extremely small limits on the difference in the accelerations of two differently composed bodies,⁸ an additional step is needed to infer from the fact that the limits drawn from tests on particular test bodies are so small to the claim that it is very improbable that one would get such small limits if in fact the WEP, understood as a claim about the relative accelerations of all test bodies whatsoever, were false. On this point, I agree with the argument given by Roberts. Indeed, as Catalina Alvarez and Robert Mann note with regard to EEP, although many tests have been conducted on systems dominated by nuclear electrostatic energy, “there are many physical systems dominated by other forms of mass energy for which the validity

⁸In terms of parameters definable within the $TH\epsilon\mu$ formalism (see below) Will notes that Eötvös experiments yield limits on “non-metric parameters” of $|\Gamma_0| < 2 \times 10^{-10}$ and $|\Lambda_0| < 3 \times 10^{-6}$.

of the equivalence principle has yet to be empirically checked” (including for example second and third generation matter such as charmed or top quarks and quantum vacuum fluctuations), and analogous comments almost certainly apply to WEP [1].

I propose, however, that an understanding of the role of plausibility judgments in error-statistical reasoning puts these considerations into their proper light. It is interesting to note that, although EEP cannot be severely tested, Clifford Will claims that we can fruitfully focus our attention on the metric theories that can be characterized within the PPN framework.

If, as I have claimed, Roberts is right about the obstacles to severely testing EEP, what could justify Will’s position? (I assume for the sake of argument that it is justified.) According to Roberts, such justification would appeal to the facts that EEP 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. I hold that this is an insufficient basis for the affirmations that Will expresses.⁹ Furthermore, this does not sufficiently take into account the extensive and ongoing experimental testing of the components of EEP themselves. It does not seem that Roberts’ account provides us with the resources to see how the mere possibility that such tests could show EEP

⁹Will does acknowledge that “The structure of the PPN formalism is an assumption about the nature of gravity that, while seemingly compelling, could be incorrect” ([20], 207), but elsewhere writes that tests of EEP “accurately verify that gravitation . . . must be described by a ‘metric theory’ of gravity” (ibid., 10).

to be in error, in conjunction with the fact that they have not done so, could warrant relying on an assumption of EEP.

Physicists are presently engaged in various programs of testing directed at WEP, LLI, and LPI. There is much more to be said about these programs than I can even suggest here. My present aim is merely to indicate how the systematic and progressive elimination of possibilities for the violation of EEP, though falling short of a severe test of the principle, can become the basis for the judgment that it is plausible to expect that any violations of EEP will be relegated to domains beyond the expected range of viability for GR.

3.1 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 [11] 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

¹⁰The 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).

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

$TH\epsilon\mu$ 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 satisfied 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 T and H , 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 [11].

¹¹It 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 about” rather than conclusively establishing general and fundamental theories in physics.

tions 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” [20], 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.

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 [16] [19]. However, such measures appear to function as a last resort, and do not reflect a judgment that such arrangements are independently plausible. 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 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 [8], the “clock” was constituted by the transition frequencies of the $J = 3/2$ ground state of the ${}^7\text{Li}$ 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

[12].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.” [12]

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 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

¹²Mattingly discusses these results and others in the context of yet another (!) formalism, the Standard Model Extension (SME) [9], that is even broader than $TH\epsilon\mu$ and that is useful for, among other things, systematizing the testing of LLI. Bailey and Kostelecky [2] 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.

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. [cite an example]

3.2 Conclusion

In Mayo's own account, severe testing is presented as the sole basis for evidential assessment, and high-level theories are assessed by putting them into a framework in which the accumulation of piecemeal severe tests is used to determine how "far" the correct description of relevant phenomena could be from a particular theory. I have argued that, in effect, severe testing cannot begin until some kinds of plausibility judgments have been made, and that this holds both for low- and high-level hypotheses.

I then applied this idea to the testing of fundamental physical principles in the context of theories of gravity. The use of parametric frameworks such as PPN and $TH\epsilon\mu$ shows how tests of such principles that fall short of being severe nevertheless enhance the plausibility in some respects of the principles, but also show how even such judgments must be qualified by an awareness that, while "conspiracies" might be plausibly ruled out, more systematic failure will remain not only plausible, but at some level to be expected.

References

- [1] Alvarez, Catalina and Robert Mann. "Testing the Equivalence Principle in the Quantum Regime," *General Relativity and Gravitation* 29 (1997):

245–50. Related online version (cited May 20, 2006):

<http://lanl.arxiv.org/abs/gr-qc/9605039>.

- [2] 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>.
- [3] Belinfante, F. and J. Swihart. “Phenomenological Linear Theory of Gravitation. I. Classical Mechanics,” *Annals of Physics* 1 (1957): 168–95.
- [4] Belinfante, F. and J. Swihart. “Phenomenological Linear Theory of Gravitation. II. Interaction with Maxwell Field,” *Annals of Physics* 1 (1957): 196–212.
- [5] Belinfante, F. and J. Swihart. “Phenomenological Linear Theory of Gravitation. III. Interaction with Spinning Electron,” *Annals of Physics* 2 (1957): 81–99.
- [6] Chalmers, Alan. “Experiment 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. 157–70.
- [7] Chalmers, Alan. “Can Scientific Theories Be Warranted?.” Unpublished ms (2006).

- [8] Drever, R.W.P. “A Search for Anisotropy of Inertial Mass Using a Free Precession Technique,” *Philosophical Magazine* 6 (1961): 683–87.
- [9] 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>.
- [10] Laudan, Larry. “How About Bust? Factoring Explanatory Power Back into Theory Evaluation,” *Philosophy of Science* 64 (1997): 303–16.
- [11] Lightman, Alan and David Lee. “Restricted Proof that the Weak Equivalence Principle Implies the Einstein Equivalence Principle,” *Physical Review D* 8 (1973): 364–76.
- [12] 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>.
- [13] Mayo, Deborah. *Error and the Growth of Experimental Knowledge* (Chicago: University of Chicago Press, 1996).
- [14] 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.
- [15] Mayo, Deborah. “Learning from Error: Severe Testing and the Growth of Theoretical Knowledge.” Unpublished ms (2006).

- [16] Murayama, Hitoshi. “Supersymmetry,” in Shoichi Yamada, ed., *Physics with High Energy Colliders: Proceedings of 22nd INS International Symposium* (Singapore, World Scientific, 1995).
- [17] 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).
- [18] Staley, Kent. *The Evidence for the Top Quark: Objectivity and Bias in Collaborative Experimentation* (New York: Cambridge University Press, 2004).
- [19] Staley, Kent. “Anti-matter, Supersymmetry, and God: On Fine-tuning and Scientific Inquiry,” unpublished ms (2006).
- [20] Will, Clifford. *Theory and Experiment in Gravitational Physics* (New York: Cambridge University Press, 1993).
- [21] 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>.