Error-Statistical Elimination of Alternative Hypotheses

Kent Staley (staleykw@gmail.com)
Saint Louis University

Abstract. I consider the error-statistical account as both a theory of evidence and as a theory of inference. I seek to show how inferences regarding the truth of hypotheses can be upheld by avoiding a certain kind of alternative hypothesis problem. In addition to the testing of assumptions behind the experimental model, I discuss the role of judgments of implausibility. A benefit of my analysis is that it reveals a continuity in the application of error-statistical assessment to low-level empirical hypotheses and highly general theoretical principles. This last point is illustrated with a brief sketch of the issues involved in the parametric framework analysis of tests of physical theories such as General Relativity and of fundamental physical principles such as the Einstein Equivalence Principle.

Keywords: error-statistics, evidence, alternative hypothesis, comparativism, gravity, equivalence principle, parametric framework

Keywords: error-statistics, evidence, alternative hypothesis, comparativism, gravity, equivalence principle, parametric framework

1. Introduction

Drawing upon Deborah Mayo's error-statistical account as both a theory of evidence and a theory of inference, I explore how error-statistical reasoning lends support to conclusions regarding the truth of hypotheses. In particular, I argue that such truth inferences require avoiding a certain kind of alternative hypothesis problem that I call the problem of the unconsidered alternative. I seek to show how error-statistical analysis can avoid this problem by using two forms of reasoning to justify exclusion of some alternatives from consideration: the *validation by testing of the experimental model* and *judgments regarding the implausibility of some alternatives*. A benefit of my analysis is that it reveals a continuity in the application of error-statistical assessment to low-level empirical hypotheses and highly general theoretical principles.

2. Error-statistical evidence/inference

I employ Mayo's error-statistical apparatus (Mayo, 1996) for two related purposes: as a theory of *evidence* and as a theory of *inference*. I begin by clarifying how these purposes relate to one another.

I follow Achinstein (2001) in attributing the following significance to the concept of evidence as here employed:[1] that some body of data or facts E is evidence for some hypothesis H entails that E constitutes a good reason to believe that H.

Let us call concepts of evidence for which the entailment holds *strong* (Achinstein, 2000). Strong evidence is closely connected with inference: when E is evidence for a hypothesis, it also becomes the basis for a reasonable inference: 'E, therefore H'. A person who makes an inference of this form treats E, which is believed to be true, as the reason for holding (either acquiring or maintaining) the belief that H is true. But

[1] Achinstein (2001) delineates several evidence concepts. Evidence in the sense here employed comes closest to the concept he denotes "potential evidence." For a discussion of the compatibility of Achinstein's understanding of potential evidence with Mayo's error-statistics, see Staley (2005).

the inference itself is only reasonable if E is indeed a good reason for believing H, and the error-statistical account as I treat it here sets out conditions which, if satisfied, ensure that this is so.

The focus here, then, is on the problem of how to expand our knowledge, i.e., how to acquire or maintain beliefs reliably. I do not deny that other aims besides true belief are at work in the scientific assessment of hypotheses. But I do assume that central to scientific concerns are such truth-related purposes as truth, empirical adequacy, or statistical adequacy.

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.

To clarify, consider the problem in the abstract. Suppose we begin with a question Q and a class of incompatible answers to that question $\{H_1, \ldots, H_n\}$. Furthermore, suppose that our theory of evidence selects some H_i as being the best supported of these based on data E. If $\{H_1, \ldots, H_n\}$ does not include *all* of the possible answers to Q that have not been ruled out already, then the fact that H_i is the *best* supported of the set will be an insufficient basis for *inferring* H_i. There may, after all, be an answer that is better supported by E that is not in the set. I will call this "the problem of the unconsidered alternative."

In the next section, I seek to show how the error-statistical theory of evidence avoids the problem of the unconsidered alternative.

3. Escaping the problem of the unconsidered alternative

The error statistical theory of evidence can be articulated in terms of Mayo's '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, and

SR2 the probability of H passing T with an outcome such as E (i.e., one that fits H at least as well as E does), given that H is false, is very low (Mayo, 1996, 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.

Consider a simplistic example: We seek to estimate the probability p of a coin coming up heads using a confidence interval. We conduct an experiment consisting of $N = 100$ tosses of the coin. The test statistic is defined as $\bar{X} \equiv n(heads)/N$. We find that $\bar{X} = 0.6$ and infer that $H : p = 0.6 \pm 0.1$. Assuming that the underlying data-generating distribution is approximately Normal, that the probabilities of possible outcomes on a given trial are independent of outcomes from previous trials, and that those probabilities are identically distributed for each trial (NIID), this constitutes a 95% confidence interval estimate, and is equivalent to passing H with high severity against the (compound) alternative J: $p \geq 0.7$ or $p \leq 0.5$.

Under the error-statistical account, the data gathered via this testing procedure constitutes evidence for H on the grounds that the conditions *SR1* and *SR2* are satisfied. If this concept of evidence is to be a strong concept, then the satisfaction of these conditions must also provide a basis for a reasonable inference from the data to H. Note, however, that merely asserting that these conditions have been met does not make a cogent argument for H. In particular, a cogent argument should also provide some justification for the assumptions that underwrite the severity assessment. In particular, the *statistical model* of the experiment, with its assumption of IID trials, must be justified. A particular strength of the error-statistical model is precisely its emphasis on this point, and on the development of testing strategies suitable for this purpose (Mayo and Spanos, 2004) (Spanos, 1999).

As indicated above, it is crucial for the suitability of error-statistics for this purpose that it not fall prey to the problem of the unconsidered alternative. And it would seem that it does not, since the severe test requirement *SR2* demands that we consider the probability of an outcome that fits H as well as or better than E under the apparently exhaustive alternative *H is false*.

My main purpose in this section is to consider some complications that might be thought to point to the opposite conclusion: that error-statistics does face a problem of the unconsidered alternative, leading to merely comparative conclusions regarding competing hypotheses. I show that these considerations do not lead to such a conclusion. In the process, I will underscore and clarify the sources of fallibility in the error-statistical approach, and indicate where I think substantive, though typically weak, assumptions enter into error-statistical inferences.

To begin, recall a fairly obvious point: typically the denial of a statistical hypothesis does not define a unique probability distribution. "H is false" is a stand-in phrase, with definite probabilities supplied by *specific* alternatives to H that entail its negation. The content of

these alternatives will be provided by the experimental model. Mayo emphasizes this point:

> 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 .... (Mayo, 1996, 190)

One might think, then, that we should restate *SR2* as follows:

SR2' for any H' within the experimental testing model such that $H' \vdash \neg H$, the probability of H passing T with an outcome such as E (i.e., one that fits H at least as well as E does), given H', is very low.

But there is a reason why *SR2* and not *SR2'* is the correct formulation: *SR2'* is too easily satisfied, simply by choosing a very restrictive experimental model that eliminates a lot of rival hypotheses. It is precisely because the error-statistical account seeks to underwrite a strong concept of evidence that it requires hypotheses to be tested in ways that avoid the problem of the unconsidered alternative. If we begin with a question Q, to which H is intended as a possible answer, then the model must include as well alternative hypotheses that offer other possible answers to Q. (Excluded are "alternatives" that really address different questions, such as hypotheses concerned with "higher level" explanations than H is concerned with, etc. (Mayo, 1996, 199–200).) If we already have good reason to believe that such an alternative is false, then it can safely be excluded from consideration, but otherwise it must be included.

So far so good, but one might wonder about the prospects for providing such an exhaustive experimental model. For example, in the coin tossing example under consideration, all of the hypotheses under consideration represent possible answers to the question: What is the probability of obtaining heads as an outcome in a toss of this coin? As answers, they all share the following assumption: that the same distribution is an equally good probability model for each toss of the coin, .i.e., that the outcomes of all trials are identically distributed.

If this assumption is justified, then excluding alternatives that violate it is also justified. Here is an example of such an alternative:

> H': For the first 100 tosses of this coin, $p = 0.6$, and for all tosses thereafter $p = 0.2$.

This alternative is excluded from the experimental model because it contradicts the assumption that the data are IID, since H' postulates

a change in the distribution after the 100th toss. This IID assumption is used furthermore in calculating the error-probabilities for the confidence interval, i.e., it is used in evaluating severity.

To understand how error-statistics avoids the problem of the unconsidered alternative, we need to understand how the decision to exclude such alternatives from the experimental model is justified. In particular, we want to know whether this decision is itself justified by severe testing.

As has been emphasized (Mayo and Spanos, 2004), good error-statistical practice demands that in evaluating the severity with which a hypothesis passes a test given the data, the adequacy of the statistical model used in that evaluation must itself be tested, and Mayo and Spanos (e.g., Mayo and Spanos, 2004; Spanos, 1999) have provided powerful insights into how such testing can proceed. A full quantitative treatment of severity requires that we work within a specified experimental testing model, one part of which will be the statistical model of the data-generating mechanism just mentioned. Consideration of the $SR2$ requirement is made in reference to this model.

This goes far toward reducing the burden of alternative hypothesis problems. It underscores, for example, that the mere fact that an alternative hypothesis can be stated that confers a high probability on the outcome E is not enough to undermine a severity assessment – many such alternatives are ruled out in the process of validating the statistical model.

These considerations, however, leave one problem to be addressed. In the scenario described here, the alternative hypothesis has been formulated precisely to make the failure of the statistical model *invisible*, as it were, to the very testing procedures that would serve to validate the statistical model. Departures from the IID assumptions take place only beyond the size of the sample in hand. Within the sample, the IID assumption would appear to be satisfied.

We would normally take ourselves to be justified in excluding H' from our testing model if the IID assumption itself passed appropriate tests of validity (Spanos, 1999, 739–53). However, if H' were true, then we would expect the data in our $N = 100$ sample to appear to satisfy IID, although IID in fact is false, since under H' outcomes of the first 100 trials *are* identically distributed; it is only for later trials that the assumption is violated. What's more, if H' were true, then H would be false, *and* the probability of getting a result that fits H as well as E does, or better, would not be low at all.[2]

[2] Of course, if we were setting out to test H', we would not use the same measure of fit as used in our example. However, any difference in these measures would only

The existence of such alternatives as H' raises two distinct worries, calling for two different error-satistical responses. I discuss the first worry and review the error-statistical reply already worked out by Mayo. The second worry requires a different kind of reply.

3.1. The skeptical gellerizer

The first worry is that the availability of unconsidered alternatives provide a basis for resisting all evidence claims based on severe testing. We can easily dispense with this worry.

Let us suppose Experimenter Esther has performed the coin-tossing example described above, with 100 coin tosses. Upon careful examination of the data, including the testing of underlying assumptions about the statistical model, Experimenter Esther concludes hypothesis H above, as a 95% confidence interval. Suppose that Perverse Pete wishes to deny any evidence claim that Esther might make, come what may (or, equivalently, that he seeks to avoid learning anything from these data).

Perverse Pete might choose to exploit such hypotheses as H' to pursue his skeptical aims. When Esther presents her data and infers from them the hypothesis H, Pete would reply, "You haven't included H' in your model, but H' *could* be true! So you don't really have evidence for H after all!"

A first comment is to emphasize the basic fallibilist insight that the bare possibility of error is not of itself a reason for doubt. But let's go farther than that: As Mayo has noted, this "gellerizing" procedure is subject to the following flaw: Pete's method will always lead him to reject hypotheses even when they are true. In that sense, it is 100% unreliable. It is precisely its unreliability that makes the method appropriate to Pete's aim of not learning anything. By using the data in hand to formulate an unconsidered alternative, and then taking that unconsidered alternative as a reason to doubt any evidence claim put forth, Pete will reject all hypotheses, regardless of the data, even when the hypothesis is true. This is the cost of pursuing the strategy of invoking unconsidered alternatives as skeptical foils.

3.2. A non-skeptical question about evidence

Now suppose we reframe our scenario. As before, Experimenter Esther performs her experiment and presents her data as evidence for H. But Curious Carl replies, "Esther, I accept that these data constitute

be manifested for $N > 100$. Even using an appropriately modified measure of fit, any result that fits H' will also fit H as long as $N \leq 100$.

evidence for H. I notice, however, that a premise in your argument for that claim is that the probability of getting data that fit H as well as these do or better, assuming H is false, is very low. Is it true, then, that the probability of getting data that fit H this well is low if H' is true?" Esther replies, "Well, no, but I have not included H' in my experimental model." Carl asks, "So are you just saying that H is the best supported of the hypotheses that are included in your experimental model?" "No Carl, I'm using a strong concept of evidence. The data don't just indicate that H is better than the alternatives I've considered, but that H is true." Carl asks, "Then I take it you do not think H' needs to be included in your experimental model for you to draw that conclusion?" Esther replies, "That's right, Carl." "Good, Esther, I agree that it does not. But why is it safe to exclude it?" "Honestly, Carl, I'm really not worried about such a crazy alternative hypothesis. I'm already pretty confident that it isn't true, so I see no need to conduct a test to rule it out."

The worry raised by this scenario is not that alternative hypotheses provide a basis for resisting evidence claims, but that we do not yet understand just what *justifies* strong evidence claims for hypotheses as against alternatives that have not explicitly been ruled out by testing, either within the experimental model or in the validation of the experimental model. Whether Carl asks his question or not, Esther needs some justification for excluding H' (and other hypotheses like it) from her experimental model. She cannot do so arbitrarily without undermining her inference. In the next section I discuss how to address this worry.

To summarize: I have described a situation that should constitute a clear, paradigmatic example of a hypothesis H severely passing a test with outcome E. I have also described an alternative hypothesis H' against which H has not been tested in this example. Yet this hypothesis is such that, if it were true, then (a) H would be false, (b) it would not be improbable that H would pass this test with a result that fits it as well or better than E does, and (c) the statistical model used in the original severity assessment would be invalid, but would still appear to be valid given the data in hand (because the departures from that model would all occur outside of the data sample itself). Although we might, of course, be able to eliminate H' from our worry by gathering a larger data sample of size $N' = N + m$, it is trivial to show that if N' is finite another hypothesis H'' will remain to which these comments still apply.

4. Excluding implausible alternatives

Newton's fourth rule of reasoning asserts that when a law is "gathered from phenomena by induction," then it "should be considered either exactly or very nearly true notwithstanding any contrary hypotheses, until yet other phenomena make such propositions either more exact or liable to exceptions" (Newton, 1999, 796). Newton recognized that the mere fact that an alternative hypothesis could be formulated that could be made to fit the data does not by itself constitute a reason to doubt an inductively supported generalization. Similarly, Peirce asserted as the "first rule of logic" that one must not "block the way of inquiry" (Peirce, [1898] 1998). Certainly to use the mere possibility of error, in the absence of any real doubt, as an obstacle to accepting the result of a sound probable inference, would be to violate Peirce's rule.

Applying both Newton's and Peirce's insights to the present issue, we can go farther. Already built into the error-statistical standards of evidence is a discrimination between those hypotheses that merely fit the data and one that has withstood being severely probed for error. Yet in order to deploy these standards we need to apply Peirce's rule at some point, lest alternative hypotheses that lie just beyond the reach of our data present an obstacle to the business of severe testing.

Surely, this is a sensible attitude to take, but can it be given a basis that goes beyond a pragmatic concern with being able to "move forward"?

In the above example, H' could of course be easily subjected to a severe test, simply by collecting a larger data sample. But it hardly seems necessary to conduct this test to justify the fact that in the example as originally described, H' is not included as an alternative to be tested against, while nonetheless the conclusion drawn entails that H' is false. Indeed, unless we had reason to think we were dealing with a very odd sort of coin, H' would almost certainly fail any "real" test to which we subjected it (i.e., any test that is informative in a sense to be explained below). There is no need to be worried about the mere existence of alternative hypotheses which are such that, were they to be subjected to some genuinely informative test, would be almost certain to fail.

That is to say, it is justified to exclude from an experimental model M (taken to include a set of hypotheses to be tested) an hypothesis H if H is such that, were it to be subjected to an *arbitrary informative test*, it would almost certainly fail.

The notion of an "informative" test here is meant to refer to something analogous to, but weaker than, a severe test. Specifically, H passes an informative test T with result E if:

IT1 E fits H, and

IT2 the probability of H passing T with an outcome such as E (i.e., one that fits H at least as well as E does), given *some* alternative H', is very low.

In effect, by eliminating from consideration hypotheses that we expect to fail any informative test we are eliminating just those that we antecedently judge would pass only tests that were "rigged" in advance to pass those hypotheses, no matter what. (For example, the skeptical gellerizer Perverse Pete might be thought to take H' as passing a test of some kind, but this is a test that is rigged to pass hypotheses independently of whether they are true.)

In short, my proposal is that we exclude from our class of relevant hypotheses those which we antecedently judge to be such that, were they subjected to an arbitrary informative test, they would almost certainly fail. This allows for the construction of an experimental model that does not include all logically possible alternative hypotheses, and yet allows us to establish strong evidence claims.

Three 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. 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, and may be corrected. 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, but especially in phenomena that are most similar to the case in hand. Such reasoning may be largely analogical in character. In our example, we exclude H' from consideration because it in effect postulates an abrupt change in the behavior of a physical system that

is of a kind well known to us. To the extent that we are justified in regarding the coin in question as resembling other coins that we have experienced, we are justified in expecting it to display broadly similar patterns of response to forces that are imposed upon it.

Third, it is important to distinguish between the kind of unconsidered alternative that would, on this proposal, be justifiably excluded from consideration and unconsidered alternatives that a given test really is incapable of discriminating against. In the coin tossing example above, acceptance of the experimental model and of H as a reasonable conclusion should be understood as not committing us to think that the probability of heads on tosses of the coin will remain within the 95% confidence interval around 0.6 *forever*, as we would expect that eventually wear on the coin might change its behavior. But this distinction might not be easy to make sharp.

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 explorer a potential 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.

5. Theories of gravity

Mayo herself has taken the experimental investigation of theories of gravity in the Parametrized Post-Newtonian framework as an example of how she sees error-statistical considerations applying to high-level theories (Mayo, 2002) (Mayo, forthcoming, 2008). Her approach 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 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. (Will, 1993, 22)

The PPN approach facilitates comparison of such theories 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" (Will, 2006, 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).

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 as a means for 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. A key to the strength of this approach 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." (Mayo, forthcoming, 2008)

John Roberts alleges that Mayo's approach does not quite work in the way that she would like (Roberts, 2006). 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 of all theories of gravity that have been 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). 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.

However, this is not possible, he claims, because "high level" theoretical claims generally 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" (Mayo, forthcoming, 2008).

The important point about this in Roberts's argument is that WEP quantifies over all spacetime and all bodies of a certain kind. He concludes that it therefore cannot in principle be severely tested. I do not assert, as Roberts does, that principles with such a universal scope cannot be severely tested. Rather, I want to point out a similarity between the *way* that severity applies to such very general principles and the way that the concept applies to such "low-level" claims as the example discussed in section three. More specifically, an understanding of the role of implausibility judgments in error-statistical reasoning puts these considerations into their proper light.

Clifford Will claims that we can fruitfully focus our attention on the metric theories that can be characterized within the PPN framework. What could justify Will's position?[3] In other words, what forms the basis for the assumption that EEP holds? After all, EEP could very well *fail* to hold for some applications that have not yet been examined. As Catalina Alvarez and Robert Mann note, although many tests of EEP 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 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) (Alvarez and Mann, 1997).

Although I can hardly begin to address the issue of the experimental status of EEP in a brief essay, I will point to where I think the answer lies. This is in the use of other parametric frameworks that facilitate the testing of WEP, LLI, and LPI. I suggest that the systematic and progressive elimination of possibilities for the violation of EEP 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, which is all that the PPN results require.

[3] Will does acknowledge that "The structure of the PPN formalism is an assumption about the nature of gravity that, while seemingly compelling, could be incorrect" (Will, 1993, 207), but elsewhere writes that tests of EEP "accurately verify that gravitation . . . must be described by a 'metric theory' of gravity" (ibid., 10).

Another formalism developed to systematize the search for violations of EEP that functions analogously to the PPN framework for tests of GR is the $TH\epsilon\mu$ formalism developed by Lightman and Lee (1973).[4] 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.[5] 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.

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" Will, 1993, 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). Eötvös experiments have tested WEP and yielded limits on "non-metric parameters" of $|\Gamma_0| < 2 \times 10^{-10}$ and $|\Lambda_0| < 3 \times 10^{-6}$.

However, 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 implausibility judgments. Because the $TH\epsilon\mu$ formalism can only be applied to a restricted class of theories, these limits require that we judge any theory outside of that class *not* to be a candidate for the correct theory of phenomena in weak-field, slow-motion limit. This claim, however, is *weaker* than the assumption of EEP (or LLI), which is needed for the application of the PPN formalism. Thus, by developing new frameworks relying on ever weaker assumptions, physicists are able to set limits on the violation of such fundamental physical principles

[4] See (Mattingly, 2005) for a recent review of this and other parametric frameworks as applied to tests of Lorentz invariance, such as the Standard Model Extension (Kostelecký, 2004).

[5] The restriction, more specifically, is to theories that describe the center-of-mass acceleration of an electromagnetic test body in a static, spherically symmetric 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 argued 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 (Lightman and Lee, 1973).

that are increasingly secure, and those principles in turn can underwrite the application of severe testing to particular physical theories.

6. Conclusion

To underwrite strong evidence claims, error-statistical inferences must avoid the problem of unconsidered alternatives. Testing of the experimental model's assumptions goes far toward achieving this aim. Here I have argued that such a strategy needs to be supplemented by weak, though substantive, assumptions regarding the implausibility of some alternative hypotheses.

By noting the implausibility judgments involved in keeping the "way of inquiry" open for error-statistical inference, I also suggest a path toward a unified error-statistical treatment of both low-level and high-level empirical claims.

References

Achinstein, P.: 2000, 'Why Philosophical Theories of Evidence Are (and Ought to Be) Ignored by Scientists', *Philosophy of Science* **67**, S180–92.

Achinstein, P.: 2001, *The Book of Evidence*, Oxford University Press, New York.

Alvarez, C. and Mann, R.: 1997, 'Testing the Equivalence Principle in the Quantum Regime', *General Relativity and Gravitation* **29**, 245–50. Related online version (cited May 20, 2006): `http://lanl.arxiv.org/abs/gr-qc/9605039`.

Kostelecký, V. A.: 2004, 'Gravity, Lorentz Violation, and the Standard Model', *Physical Review D* **69**, 105009. Related online version URL (cited July 23, 2007): `http://arxiv.org/abs/hep-th/0312310`.

Lightman, A. and Lee, D.: 1973, 'Restricted Proof that the Weak Equivalence Principle Implies the Einstein Equivalence Principle', *Physical Review D* **8**, 364–76.

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

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

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

Mayo, D.: 2008, 'Severe Testing, Error Statistics, and the Growth of Theoretical Knowledge', in Mayo, D. and Spanos, A. (eds.), *Error and Inference: Recent Exchanges on the Philosophy of Science, Inductive-Statistical Inference, and Reliable Evidence*, Cambridge, Cambridge University Press. unpublished ms.

Mayo, D. and Spanos, A.: 2004, 'Methodology in Practice: Statistical Misspecification Testing' *Philosophy of Science* **71**, 1007–25.

Newton, I.: 1999, *The Principia: Mathematical Principles of Natural Philosophy*, tr. Cohen, I. B. and Whitman, A., University of California Press, Berkeley.

Peirce, C. S.: [1898] 1998, 'The First Rule of Logic,' in Peirce Edition Project (eds.), *The Essential Peirce: Selected Philosophical Writings*, vol. 2, Indiana University Press, Bloomington.

Roberts, J.: 2006, 'Coping With Severe Test Anxiety: Problems and Prospects for an Error-Statistical Approach to the Testing of High-Level Theories', unpublished ms.

Spanos, A.:1999, *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press, Cambridge.

Staley, K.: 2005, 'Agency and Objectivity in the Search for the Top Quark', in Achinstein, P. (ed.), *Scientific Evidence: Philosophical Theories and Applications*, Johns Hopkins University Press, Baltimore, pp. 165–84.

Will, C.: 1993, *Theory and Experiment in Gravitational Physics*, Cambridge University Press, New York.

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

7. Acknowledgments

I am grateful for very helpful discussions with Deborah Mayo and John Roberts. Thanks also to Aris Spanos, Bill Rehg, members of the SLU HPS study group, and participants in the ERROR 2006 conference and the LSE/Pitt Confirmation, Induction, and Science conference, where earlier versions of some of these thoughts were presented.