Coping With Severe Test Anxiety:
Problems and Prospects for an Error-Statistical Approach to the Testing of High-Level Theories

John T. Roberts

Department of Philosophy

CB # 3125

University of North Carolina at Chapel Hill

Chapel Hill, NC 27599, USA

jtrosap@email.unc.edu

Coping With Severe Test Anxiety:

Problems and Prospects for an Error-Statistical Approach

to the Testing of High-Level Theories

Abstract:

Mayo (1996) begins the development of a general account of the epistemology of science, the "error-statistical philosophy of science" (ESPOS). The core commitments of ESPOS are that all scientific evidence takes the form of severe tests, and thatseverity requires low error probabilities. I examine the question of whether the basic commitments of ESPOS are compatible with a satisfactory account of the experimental testing of high-level theories. I argue that Mayo's arguments for the affirmative are unconvincing: Not only are severe tests of high-level theories impossible, but the strategies Mayo proposes for learning about high-level theories via severe tests are not promising. I then propose a way of extending ESPOS to make possible a satisfactory treatment of the testing of theories.

1. Introduction

Deborah Mayo's Error and the Growth of Experimental Knowledge (1996) begins the development of a general account of the epistemology of scientific testing and experiment, which Mayo calls the "error-statistical philosophy of science" (henceforth ESPOS). ESPOS focuses attention on the experimental testing of very specific hypotheses. Examples of such hypotheses include that two factors are more strongly correlated than they would be by chance, that a population

mean is within a certain interval, and that a given numerical parameter has a value within a certain range. Such hypotheses are far more specific in their content that high-level scientific theories, and the testing of such hypotheses raises different problems than the testing of high-level theories. It is an interesting question whether ESPOS can be extended into an adequate account of the testing of theories. Mayo (1996) contains some discussion of this question, but gives much less attention to it than to issues concerning the testing of low-level primary hypotheses. The present paper is a brief and preliminary exploration of the problems and the options confronting an advocate of ESPOS for dealing with the testing of high-level theories. By the end of the paper, I will have arrived at a cautiously optimistic view of one way a friend of ESPOS might treat the testing of high-level theories. I am not at all sure, however, that this is a way that Mayo herself would be happy with.

2. ESPOS: The Core Commitments

Two basic commitments at the core of ESPOS are a thesis about (scientific) evidence:

> We have good (scientific) evidence for hypothesis H just in case, and to the extent that, H has passed a severe test

(see Mayo (1996), p.177) and a criterion for severity of tests:

> When data D are gathered using test procedure T, hypothesis H passes a severe test with D just in case, and to the extent that:

(i) D fit H (that is, getting data D from test procedure T is likely or to be expected given the truth of H); and

(ii) On the assumption that H is false, the probability of T yielding data that fit H at least as well as do D is low.

(See Mayo (1996), pp.180-181.) The probability mentioned in (ii)--the *error probability*--is to be understood in frequentist terms: it represents the fraction of repeated instances of T in an actual or hypothetical long series of instances that would yield data that fit H at least as well as D, given the assumption that H is false. Probabilities are thus always measures of the reliability of test procedures; they are never assigned to hypotheses themselves. If H is a conjunction, so that there are multiple ways in which it can be false, (ii) is to be interpreted as requiring that the error probability be low given each alternative to H (Mayo (1996), p.195).

3. The Problem of High-Level Theories

The requirement that the error probability be low given each way in which the hypothesis under test could fail is a very strong one. But when the hypothesis under test is a low-level one, it is not usually prohibitive. In many cases, the hypothesis under test is a claim about (frequentist) probabilities, and so is the negation of that hypothesis. In such cases, it is often a straightforward exercise to show that the error probability is low given any alternative to the hypothesis. For example, if the hypothesis is that two factors are correlated at least to degree 0.6, then every alternative implies that the factors are correlated to a degree of less than 0.6, which in turn fixes an upper bound on the error probability for any given

experimental outcome. In other cases, the hypothesis and its negation are not directly about probabilities, but when conjoined with an (independently testable) experimental assumption to the effect that measurement errors are randomly distributed, they entail something about the probabilities. For example, if the hypothesis says that the value of a certain numerical parameter p lies within an interval (a, b), then the only alternatives are that p is less than or equal to a, and that p is greater than or equal to b. Given the experimental assumption that the errors in one's method of determining the value of p are randomly distributed, the data gathered by a test can be used to calculate the error probability given either of these alternatives, using standard mathematical techniques.

The situation is very different when the hypothesis we want to test is a high-level theory. For the set of alternatives to such a theory typically cannot be regimented into any small number of propositions each of which yields information about the error probabilities. Indeed, there are general considerations that seem to show that severe tests of high-level theories are impossible in principle. Chapter 6 of Mayo (1996) gives reasons for doubting this claim, so I will now try to make it plausible, first by considering a particular example, and then by sketching a procedure for generalizing this example.

Given Mayo's definition of severity, every test that Newtonian Gravitation Theory (henceforth NGT) ever passed failed to be even a little bit severe, just because of the existence of the General Theory of Relativity (GTR). GTR makes predictions for the results of such tests that agree with those of NGT within experimental error (given the precision available at the time the tests were carried out). (Of course, this doesn't include the tests that NGT failed, such as the one involving the orbit of Mercury. But the topic here is the tests that NGT passed.)

GTR is an alternative to NGT, and the fact that no one had ever thought of it when most of those classical tests of NGT were performed is irrelevant: If GTR is true, then the probability that we would get results that fit NGT as well as the actual results is very high. So, none of the tests passed by NGT were severe.

Furthermore, even if NGT were true and had passed all the tests it actually failed, and these tests had discredited GTR, NGT would still not have been severely tested. For these were really tests of hypotheses concerning the behavior of bodies in the solar system. One alternative to NGT is that the exponent in the denominator of the gravitation law is a variable that depends on such conditions as location in space, location in time, or the presence or absence of certain kinds of matter or certain (non-negligible values of) fields, and conditions just happen to be such that in our solar system, that exponent has the value 2 (to within limits of experimental error). Assuming that such an alternative is true, it is very likely that the results of solar-system tests of NGT will yield results that fit NGT marvelously well, even though NGT were false. Hence, the error probabilities are high for some alternatives, and the condition for severity cannot be met.

The problem can be posed in a more general form. Let T be any high-level theory, containing laws that quantify over all bodies and all of space-time. Suppose that all past experimental tests of T concerned bodies within space-time region R. (Clearly, this will always be true for some R.) There exist alternative theories to T which posit all the same basic fields and types of matter as T, but also posit other basic fields (or types of matter), and imply that when the values of the extra fields (or the density of matter of the extra types) are below some critical level, all motions will conform to the predictions of T within limits of experimental error, but will diverge radically otherwise. Take one of these alternatives to T and conjoin it with

the hypothesis that the values of the extra fields (or densities of the extra types of matter) are below the critical level in region R, and you have an alternative to T on which the probability of getting experimental results that fit T as well as could be hoped are high. This is a perfectly general recipe that should work no matter which high-level theory T is. Hence, contra Mayo (1996) (p. 191), severe tests of high-level theories are not possible. (For a different argument for this claim, see Laudan (1997), pp.313-314.)

Note that this argument does not rest on the assumption (criticized by Laudan and Leplin (1991) among others) that every theory has empirically equivalent rivals. For the general recipe just sketched is not a recipe for generating an empirically equivalent rival to any successful theory. Empirical equivalence, as traditionally understood (and as understood by Laudan and Leplin – see pp. 451-455) is the relation two theories stand in when they have all the same empirical consequences (or on the semantic variant, all the same empirical models). The theories T and T', however, have different empirical consequences, for they lead to different observational predictions for certain conditions. T' and T make the same predictions about the results of the tests of T that have been performed so far, but that is a weaker relation than empirical equivalence.

Nonetheless, my objection does appeal to underdetermination of theory by available evidence. Mayo replies to an objection based on underdetermination in her (1996), Chapter 6 (see especially p. 212). However, the objection she replies to there is distinct from the one raised here. She considers an opponent who alleges that, just because every high-level theory has logically possible alternatives that fit the existing data, ESPOS cannot endorse accepting any such theory. The objector assumes that if two hypotheses each fit a given set of data equally well, then they

will each be equally severely tested by it. But this does not follow. Even in cases where alternative, competing hypotheses both fit the data equally well, Mayo argues convincingly that it is possible for one but not the other to pass a severe test with that data (pp. 200-203). This is because, given the test procedure used to generate the data, the error probability associated with the test of one hypothesis might be quite different from that associated with the test of the other hypothesis.

The objection presented above does not work in the same way as the objection Mayo successfully dismisses. It does begin with the observation that high-level theories have logically possible alternatives that fit all the existing data equally well. But it does not then simply assume that fitting the data equally well makes for equal evidential support. Unlike the objection Mayo refutes, it attends to the second condition for severity as well as the first. If T and T' are competing high-level theories, such that if either is true then the actual results of experimental tests of T are very probable, then the error probabilities associated with the tests that yielded these results are high: If T', rather than T, were the true theory, then it would be very likely that we would get results that fit T as well as the actual results do. Since severity requires a low error probability given any alternative to T, the availability of T' renders T impossible to test severely. T, of course, returns the compliment: Its availability as an alternative to T' renders all extant tests of T' non-severe. So in the end we do get the result that 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, no severity at all. This result does not depend on the bare assumption that in general, equal goodness of fit entails equal evidential support.

These considerations appear to show that, given the basic commitments of ESPOS, no high-level theory can be severely tested, and hence we cannot have any

good evidence for a high-level theory. One might reply that this is not a fatal objection to ESPOS, on the grounds that strictly speaking experiments do not test high-level theories, but only particular, local consequences of them. Still, it would be implausible to suggest that experimental testing has nothing to do with the empirical support and criticism of high-level theories. If ESPOS can be developed into a satisfactory general theory of the epistemology of science, it ought to be able to illuminate the epistemological relation between experimental tests and high-level theories. Mayo has argued that there are two in which experimental tests can be relevant to the epistemological assessment of theories, even in the absence of severe tests of the theories themselves. The first way is by probing a theory for specific errors; the second is by "squeezing the space of theories," by locating the true theory within some relatively narrow range of possible theories. In the following two sections, I'll examine these ideas, using tests of gravitational theories as examples. The upshot will be pessimistic. It will turn out that in some interesting cases, and perhaps in typical cases, severe tests are not even available for claims that rule out particular errors in a high-level theory, or claims that narrow down the range of possible theories (cf. Earman (1992), p. 177). However, this negative conclusion will point the way toward one more hopeful for the friend of ESPOS, which will be the subject of section 7.

4. Probing a Theory for Errors: The Case of the Eclipse Expeditions

Mayo suggests that even when it is not possible to test a high-level theory severely, scientists can test particular low-level hypotheses that say that the theory is free of this or that particular error. By performing severe tests of these low-level

hypotheses, scientists probe the high-level theory for specific errors. Laudan (1997) labels this procedure "balkanization," (p. 315) and claims that it amounts to an evasive change of subject (p. 313). The question we originally wanted to answer was that of how empirical evidence can provide epistemic support for a high-level theory. But the question answered by this piecemeal approach is the distinct question of how empirical evidence can be used to pose severe tests for lots of low-level hypotheses derived from a high-level theory (together with other assumptions). Given the first core commitment of ESPOS, this entails an answer to the question of how we can have good evidence for all those low-level hypotheses, but it does not seem to provide an answer to our original question. Mayo can justly point out that even if one grants for the sake of argument that "balkanization" fails to provide severe tests of high-level theories, nonetheless ESPOS explains how empirical evidence can be used to probe such a theory for specific errors, and knowledge concerning which particular errors a high-level theory is and is not guilty of is a valuable thing. (Cf. Mayo 1997, 332.) So even if severe tests of high-level theories aren't available, it seems that the piecemeal approach nonetheless allows experiments to answer important questions about high-level theories. In this section, though, I'll point out some problems for even this modest claim.

GTR implies that when light passes through a non-negligible gravitational field, its spatial path is not straight; it gets deflected by an amount that depends on the strength of the field. The famous eclipse expeditions of 1919 used observations of the stars (apparently) near the sun during a solar eclipse in order to test this consequence of GTR. By comparing the apparent positions of stars during the eclipse with their apparent positions a few months later, the amount of deflection of light passing by the limb of the sun -- conventionally called λ -- was estimated.

Several photographic plates were taken both during the eclipse (the eclipse plates) and at a later time (the control plates). No plate can be assumed to be perfect, but if the errors in the apparent positions recorded by the plates are randomly distributed (an experimental assumption that can be checked independently), then one can use straightforward classical statistical techniques to estimate the value of the deflection parameter λ. GTR predicts a certain value for λ; an experiment that successfully estimates λ to lie within a small interval containing the predicted value severely tests the hypothesis that λ lies in this range. In this way, a consequence of GTR is subjected to a severe test. What is tested is the hypothesis that GTR is not guilty of a particular error, namely an error (of at least a certain size) in its prediction concerning the deflection of light near the sun.

Things aren't quite as simple as the preceding sketch makes them seem, however. It is a familiar point that in order to derive testable consequences form a high-level theory, auxiliary hypotheses are required. ESPOS provides a helpful way of locating the role of such auxiliaries: they show up in the experimental models. In Mayo's scheme, an experimental model is a model of an experimental situation that reveals a connection between the data gathered by a test procedure and the hypothesis under test (see Mayo (1996), p.133). This is exactly what auxiliary hypotheses are traditionally thought to do; they allow us to derive testable predictions from the hypothesis under test.

Which experimental assumptions are needed in the eclipse experiment? It depends on which hypothesis is getting tested, and unless we are careful about attending to this dependence, we are liable to be misled. Mayo argues that many underdetermination objections to ESPOS arise from carelessness about this very point (see Mayo (1996), p.189, pp.199-200). In the case of the eclipse experiment

of 1919, we can distinguish several candidates, including (but not limited to) the following:

1. GTR;
2. GTR's implications concerning the deflection of light by gravitational fields in general;
3. The hypothesis that, at any time, the magnitude of the deflection of light passing by the limb of the sun, caused by the sun's gravitational field, is the predicted value 1.75" (or within some narrow interval including 1.75");
4. The hypothesis that, at any time, the magnitude of the deflection of light passing by the limb of the sun is the predicted value 1.75" (or within some narrow interval including 1.75");
5. The hypothesis that the magnitude of the deflection of light passing by the limb of the sun at the time of the eclipse of 1919 is the predicted value 1.75" (or within some narrow interval including 1.75");
6. The hypothesis that the difference in the apparent positions of the particular stars involved, between the times when the eclipse plates were exposed and that when the control plates were exposed, is within some narrow interval containing 1.75".

No matter which hypothesis is under test, some assumptions about the experimental situation will be needed in order to determine to what degree the actual results fit this hypothesis. The needed assumptions will vary depending on which hypothesis is being tested. If 6 is the hypothesis under test, then assumptions concerning the randomness of measurement errors and the reliability of the method

of exposing the plates will be needed, but perhaps little else. If 5 is the hypothesis under test, then we will need a further assumption, in order to insure that the measured value of λ is a reliable indication of the amount of deflection suffered by the light while it passed by the limb of the sun (as opposed to deflection suffered at other points on its trajectory). Suppose that there exists an object, such as a dust cloud, lying between the sun and the earth at the time of the 1919 eclipse expedition. Suppose that this object has anomalous optical properties that enable it to act as a lens. Light from the measured stars would pass through this object on the way to the eclipse plates, but not on the way to the control plates. In this case, the measured value of λ would not be a good indication of the deflection of light as it passed by the limb of the sun. The same would be true if there were some lens-like object or region that was traversed by starlight on its way to the control plates but not by starlight on its way to the eclipse plates. The use of the eclipse experiment to test hypothesis 5, then, requires the assumption that apart from that due to deflection of light as it passes by the sun on the way to the eclipse plates, there is no non-negligible difference in the direction of the light reaching the two sets of plates.

Unlike hypothesis 5, hypothesis 4 generalizes over time. Hence, a test of 4 requires the further assumption that there was nothing special about the time of the eclipse experiment that affected the results. If, for example, some unusual event inside the sun resulted in temporary anomalous optical properties of the sun's corona, which would have added an extra deflection to light passing through it, then the actual results of the eclipse experiment would not fit 4, though they would fit 5.

Hypothesis 3 refers to cause of the deflection, and so a test of it requires the assumption that during the eclipse experiment, non-gravitational causes of light-deflection near the sun (e.g., optically inhomogeneous gases surrounding the sun) are not present or else negligible. (Hypotheses 1 and 2 quantify universally over all of space-time, and hence, by the argument of section 3, are not severely testable at all.)

Not only are such assumptions required in order to determine how well the actual results fit the hypothesis under test; they are also typically needed in order to determine the error probabilities of the test. For example, if the assumption of no non-gravitational causes of light-deflection near the sun is false, then depending on the nature of whatever such causes are present, it may well be very likely that results fitting hypothesis 3 at least as well as the actual results would be obtained even if 3 were false. Hence, auxiliary assumptions are needed to assess how well *each* of the two criteria of severity are satisfied. An assessment of how severe a test a given hypothesis has passed presupposes the truth of some such assumptions. How strong and varied those assumptions are depends on how strong the hypothesis under test is. Hypotheses 1-6 are in decreasing order of strength, and so are the experimental assumptions needed to design and assess tests of them.

Suppose that if certain experimental assumptions are true, then hypothesis H passes a test T, and the test is a severe one. In order for this to count as good evidence for H, is it (i) sufficient that these assumptions be true? Or is it (ii) also necessary that we have good evidence for them (i.e., severe tests that they have passed)? Answer (i) seems out of step with the spirit of ESPOS: The project is to explain the rationality of science, not just its reliability. It shouldn't be enough to

have performed a test in which the two requirements for severity are in fact met, if we have no good evidence that they are met. (See Mayo (1996), p. 161.)

Answer (ii), on the other hand, imposes a very heavy burden. Consider what would be required of a severe test of the experimental assumption which, I argued above, is needed for testing hypothesis 5: that the only difference in the deflection suffered by the light reaching the eclipse plates and the light reaching the control plates is the deflection suffered by the former upon passing the limb of the sun. The experimental assumption here generalizes over all non-solar causes of differential light deflection, known and unknown. Any possible non-solar cause is a way that the assumption could fail. Hence, a severe test of this assumption must be such that, given any such possible cause, the probability of getting results that fit the assumption as well as the actual results is low. There are many conceivable mechanisms that might cause such deflection, and we ought not ever assume that we have thought of them all. Moreover, there are a tremendous number of possible spatiotemporal locations for such a cause. The experimental assumption will fail if there is some cause of light deflection, working by any possible mechanism, lying anywhere along the space-time path traversed by the starlight that reached one of the sets of plates, so long as it does not lie on the space-time path traversed by the starlight that reached the other set of plates. Answer (ii) requires a severe test of the hypothesis that there is no such mechanism anywhere--a very tall order.

This observation makes the prospects for a severe test of hypothesis 5 (for which the needed experimental assumptions are rather weak, compared to what is required for testing hypotheses 1-4) seem quite dim, if we accept answer (ii). Of course, this does not amount to a proof that there cannot be a severe test of the experimental assumption required for testing 5; perhaps there really is some way of

ensuring low error probabilities for every one of the possibilities discussed in the last paragraph. But this seems to bring little comfort to the friend of ESPOS who wants to go with answer (ii). For the common opinion of physicists is that the data gathered by the eclipse expedition of 1919 already provides pretty good evidence for the general theory of relativity (or at least for hypothesis 5). ESPOS together with answer (ii), however, implies that this data is not even good evidence for 5 yet, and that it will not be, until we somehow manage to severely rule out the presence of lens-like systems all the way from here to the distant stars. Since a severe test of hypotheses 1-4 would require even stronger experimental assumptions than would a test of 5, these considerations all carry over to them.

The problems may be more tractable the weaker the hypothesis under test: If all we want to test is hypothesis 6, then assumptions concerning the photographic apparatus and the randomness of the measurement errors might be all we need. Thus, the case for the possibility of severe tests is stronger the weaker the hypothesis under test is. But we have seen reason to doubt that severe tests are in principle available for hypotheses 1-5. The problem this raises for ESPOS is that the weaker hypotheses are the ones whose connection with the high-level theory of interest is most tenuous. The idea behind the piecemeal approach is that severe tests can be used to learn about high-level theories, by probing those theories for errors. In the list of hypotheses above, the only one where severe tests seem to be available is 6, which does not declare GTR to be free from error, but rather asserts a particular empirical fact. This fact is not a consequence of GTR–it is only a consequence of GTR plus a set of auxiliary assumptions, severe tests of some of which are not feasible. Further, GTR is not a consequence of the severely tested fact, and we cannot say anything informative about how probable this fact would be

if GTR were false (unless we want to resort to interpreting probabilities as degrees of belief, which would be to abandon the spirit of ESPOS). So, even if 6passes a severe test, that won't tell us anything about which errors GTR is and is not guilty of. In short, the severe tests that are available do not probe the high-level theory of interest for errors. What it probes for errors is something else, namely a specific claim about the apparent deflection of light from certain stars on a certain occasion.

This is the difficulty faced by the friend of ESPOS who insists that for a hypothesis to be counted as passing a severe test, the two conditions for severity (fit and low error probabilities) must not depend on any experimental assumptions that have not themselves been severely tested. This difficulty seems to be intolerable. But is there any alternative for a friend of ESPOS? (Assuming that one does not want to revert to the answer (i), effectively embracing a form of reliabilism, and giving up the project of explicating the rationality of experimental method.) A remark of Mayo's hints at one; immediately after a discussion of the importance of the testability of experimental assumptions, she writes:

> In any event, experimental assumptions are part of the statistical report and can be challenged by others; they are not inextricably bound up with subjective degrees of belief in the final sum-up. (Mayo (1996), 161.)

The idea seems to be that even in cases where some of the assumptions of the experimental model have not been severely tested, at least they have been made explicit, so that other researchers can challenge them. This is enough to counter the real threat Mayo sees in the area–the idea that in the end, all testing has to rely on untested auxiliary assumptions that we countenance simply because we have a high

degree of belief in them, and that the buck must stop at those untested auxiliaries.

These remarks inspire the following proposal: The severity of a test should be relativized to a certain set or auxiliary assumptions that are used to assess the two requirements of severity (fit and low error probability). This proposal would allow us to say that hypotheses higher up the list than hypothesis 6 have been severely tested, and that we have good evidence for them, relative to a certain set of assumptions. Since some of these hypotheses are more closely linked to GTR itself than 6 is, such results look more like real probes of GTR for specific errors. (Laudan (1997) similarly proposes (pp. 314-315) that the friend of ESPOS address the problems posed by high-level theories by relativizing the claim that a hypothesis has been severely tested. But his proposal is quite different from the one made here, for he proposes relativizing this claim to the set of currently available alternatives to the theory under test, whereas the present proposal relativizes the claim to the set of background assumptions used in determining whether the two requirements for passing a severe test have been met.)

Many philosophers of science have held that evidential support is relative to a set of background assumptions (e.g., Feyerabend (1962), Kuhn (1962), Lakatos (1978), Glymour (1980), Longino (1990), Laudan and Leplin (1991)); the present proposal seeks to wed the core commitments of ESPOS with this idea. An advocate of ESPOS might well feel butterflies upon contemplating such a marriage. ESPOS aims to find a criterion of evidence with real teeth, eschewing reference to what a scientist happens to believe, or what paradigm a scientist happens to subscribe to, and relativizing severe tests to background assumptions seems to risk giving up this ambition. The most serious worry is that the severity of tests, and the quality of evidence supplied by severe tests, will be trivialized by this proposal,

since it is not hard to imagine that one could come up with severe tests of the most outlandish and arbitrary hypotheses just by choosing the right set of background assumptions. I'll propose a way of alleviating such worries in the final section. But first, I'll turn to the second way in which Mayo suggests that low-level experimental tests may be relevant to high-level theorizing.

5. Squeezing the Space of Theories: The Case of Metric Theories of Gravity

GTR belongs to a large and interesting class of theories called "metric theories of gravitation" because what they have in common is that they treat gravity not as a force but rather as a manifestation of metrical phenomena, including space-time curvature. More precisely, all and only metric theories imply the Einstein Equivalence Principle (to which we will return shortly).

On any metric theory of gravity, in the Newtonian limit (in which gravitational fields are relatively weak), certain "super-equations of motion" are satisfied. These are equations of great generality, containing ten numerical parameters, such that for certain values of the parameters these equations yield the equations of motion. By setting the ten parameters to different numerical values, one obtains (the Newtonian limits of) different metric theories, and every metric theory can be obtained by setting those parameters to one set of values or another. Hence, we can think of each metric theory as occupying a point in a ten-dimensional space, isomorphic to the ten-dimensional parameter space defined by the ten parameters of the equations of motion. This use of the super-equations of motion, and the ten-dimensional space whose points represent metric theories, is known as the Parametrized Post-Newtonian (PPN) formalism. (See Will (1993), pp. 103-4).)

Much recent experimental work on gravity can be understood as a project in measuring the values of the ten PPN parameters. Of course, a hypothesis that assigns a precise value to a parameter whose value is a real number cannot be tested severely. But what can often be tested severely is the hypothesis that such a parameter lies within a certain interval. By estimating the values of the ten parameters to lie in narrower and narrower intervals, experimenters can effectively narrow down the region of the ten-dimensional space in which the true theory lies. That is, they can zoom in on the true theory, by locating it within smaller and smaller ranges of possible theories. Ideally, then, we might be able to test severely the hypothesis that the true theory of gravity lies within some very small region of the 10-dimensional space. This would not be the same thing as severely testing one such theory. But it would still be pretty impressive. This illustrates a way in which impressive experimental knowledge concerning high-level theories could be attainable, consistently with ESPOS.

Alas, there is a catch. Measurements of the ten PPN parameters are meaningful only if the true theory really does lie somewhere within that ten-dimensional space. That is, these measurements have the physical content they are supposed to have only if some metric theory of gravitation is true–in other words, only if the Einstein Equivalence Principle (henceforth EEP) is true. The hypothesis that the true theory lies somewhere in a certain small region of the ten-dimensional PPN space is then equivalent to a conjunction of eleven separate hypotheses: EEP, and a hypothesis restricting the range of each parameter. We cannot severely test this hypothesis without severely testing its first conjunct, EEP.

6. Testing EEP

EEP is itself a conjunction of three propositions: The Weak Equivalence Principle, Local Lorentz Invariance, and Local Position Invariance (Will (1993), p. 22). I will focus on the Weak Equivalence Principle (though the same point could be made by focusing on either of the other two conjunct):

> *The Weak Equivalence Principle (WEP):* If an uncharged test body is placed at an initial event in space-time and given an initial velocity there, then its subsequent trajectory will be independent of its internal structure and composition.

EEP can be severely tested only if WEP can. (If Schiff's conjecture is true, then testing WEP is also a sufficient condition for testing EEP. See Will (1993), p. 38).

WEP is itself a high-level theoretical hypothesis, quantifying over all of space-time and all bodies. So the argument of section 3 shows that it cannot pass a test that is severe in Mayo's sense. It will be worthwhile to look at this case in a little more detail, to see exactly why severe tests are unavailable for WEP.

A broad class of experiments known as E $\bar{t}\bar{v}$ $\bar{s}$ experiments have been used to test WEP. Such experiments effectively place limits on the possible values of the E $\bar{t}\bar{v}$ $\bar{s}$ ratio, η, for various forms of matter-energy, where η is a measure of the size of violations of WEP ($\eta = 0$ means no violation). Results have been extremely

impressive: For some materials, $|\eta|$ has been estimated at less than 10^{-12} (Will (1993), p.27).

These experiments test *something* severely – namely, that in the space-time regions where the tests were performed, the values of η for the forms of matter and energy tested differ from 0 by at most very tiny amounts. This is not sufficient for a severe test of WEP itself. For a severe test of WEP requires a low error probability given *any* alternative to WEP. There are alternatives to WEP that take the following forms:

(A) η is 0 for the following forms of matter and energy: ... but diverges from 0 for other forms

(where the ellipsis is filled in with a list that includes all forms of matter or energy that have been used in E $\bar{t}\bar{v}$ $\bar{s}$ experiments to date).

(B) η diverges from 0 in many space-time regions, but is extremely close to 0 throughout the region occupied by the solar system.

(A) and (B) are not merely philosophers' skeptical nightmare scenarios (Cartesian demons, brains in vats etc.). They could each be consequences of general, fundamental physical theories that do not differ radically in form from actual fundamental physical theories. For example, (B) could be a consequence of a theory according to which the motions of bodies in free fall depend on the coupling of matter with a certain field, which happens to have value zero (or else a negligible

value) throughout the region of space-time we inhabit (perhaps this field began with a high value everywhere, which gradually shrank as the universe expanded). (A) could be a result of a fundamental theory of matter according to which violations of WEP show up only for particles of a certain type. This type of particle could be rare and exotic (tachyons?) which would explain why we have not yet encountered them. But it needn't be. Perhaps it is a fairly typical kind of particle that just happens not to have very many instances in our galaxy. Or perhaps such particles were very numerous in the past, but due to frequent collisions with the corresponding anti-particle, there are not many left at this stage of cosmic history. On any such scenario, WEP would be false, and so would all metric theories of gravity, yet it would be very probable that the WEP would pass the tests posed for it by E $\bar{t}\bar{v}$ $\bar{s}$ experiments. Hence, by Mayo's criterion of severity, such experiments do not constitute severe tests of WEP.

Alternatives of forms (A) and (B) are speculative, of course, and I don't mean to suggest for a moment that the probative value of E $\bar{t}\bar{v}$ $\bar{s}$ experiments are compromised by them. Imagine someone objecting to the claim that we have good scientific evidence for the WEP on the grounds that such alternatives have not been ruled out. A typical physicist, I imagine, would be rather impatient with the objector. "Of course there are other possibilities that we haven't ruled out. So what? That's always the situation in scientific research of this kind. We think it's reasonable to ignore such speculative possibilities unless and until some good empirical reason for taking them seriously arises." A Bayesian could endorse this response to the objector in good conscience. Physicists assign such speculative hypotheses a low prior probability; thus, they don't destroy the value of evidence from E $\bar{t}\bar{v}$ $\bar{s}$ experiments as confirmations of WEP. But an advocate of ESPOS cannot be so carefree about this matter. What difference does it make if the

speculative alternative has no independent empirical support at this time? What difference does it make if contemporary scientists have a low degree of belief in the alternative? What matters for an advocate of ESPOS is the severity of tests, which depends on the smallness of error probabilities. In the case of hypotheses that can fail in more than one way, the error probability has got to be small for each alternative. But in the case at hand, there are alternatives for which the error probabilities are large. Hence, E $\bar{t}\bar{v}$ $\bar{s}$ experiments are not severe tests of WEP.

If we require the "squeezing of the space of theories" to involve severe tests of hypotheses estimating the values of parameters, then the squeezing of the space of metric theories is possible only on the assumption that EEP is true. In other words, EEP is a crucial experimental assumption of severe tests of hypotheses that estimate the values of the ten PPN parameters. But a severe test of EEP itself is not possible. If one of the requirements on a severe test is that all of its experimental assumptions themselves be susceptible of severe testing, then this means that the squeezing of the space of metric theories is not possible.

If, however, we adopt the proposal of section 5, then this problem needn't arise. If EEP itself is relied on as an auxiliary hypothesis, then there appears to be no insuperable obstacle to severe tests of hypotheses that narrow down the possible ranges of the ten PPN parameters. But the severity of such tests, and the quality of the evidence they provide, will be relativized to a set of background assumptions including EEP itself. EEP has clearly been "tested" in an important sense – the variety and precision of E $\bar{t}\bar{v}$ $\bar{s}$ experiments that have been performed are extremely impressive. It's just that such experiments don't live up to Mayo's demanding requirement of severity.

7. Adding Teeth to the Proposal: Measurement and Vindication

The proposal to relativize severity of tests to sets of background assumptions, which must be made explicit but need not themselves be severely tested (or even severely testable) simultaneously with the hypothesis under test, threatens to trivialize the ESPOS program, by making it possible to severely test all sorts of crazy things just by selecting the right background assumptions. In order for this trivialization to be avoided, we need some way of distinguishing between "interesting" and "uninteresting" background assumptions, where a severe test, and the evidence it affords, should impress us only to the extent that the assumptions it relies on are interesting. Letting the interesting background assumptions just be the true ones would be an option of considerable appeal, if only we were in a position to know which background assumptions are true. Considering the interesting background assumptions to be the ones that have been severely tested leads to difficulties encountered above. If we suppose that the interesting assumptions are the ones we have some good evidence for, then we are going to need some criterion of good evidence other than the one espoused by ESPOS (good evidence = severe tests). Perhaps we ought to say that the interesting assumptions are the ones that, for whatever reason, strike us as worth taking seriously. But this move would compromise the spirit of ESPOS, letting our subjective beliefs or degrees of belief play a role in evidential reasoning. Is there nowhere to turn?

Consider a striking feature of EEP that it exhibits when used as an auxiliary hypothesis for tests of estimations of the PPN parameters: It makes possible the measurement of certain numerical parameters, which otherwise would not be measurable. This is because if EEP were false, then no metric theory of gravity would be true at all; hence, the super-equations of motions of the PPN formalism

might not hold in the Newtonian limit, and the PPN parameters only even make sense in the context of those equations. (They are, by definition, parameters occurring in those equations.) Without the assumption of EEP, we can't even be sure that there are such parameters; with EEP, and other pieces of background knowledge, we can actually estimate their values. Hence, adopting EEP plays a crucial role in making certain empirical descriptions of reality possible.

Furthermore, EEP (together with other items of background knowledge implicated in tests of metric theories of gravity) makes possible multiple, independent measurements of one and the same numerical parameter. (For example, there are now a variety of ways of estimating the light-deflection parameter λ ; see Will (1993), pp. 166-173). It is logically possible that repeated measurements of the same parameter, or measurements of the same parameter by means of different experimental procedures, will fail to give results in agreement with one another. The fact that, thus far, this hasn't happened, is striking, and reinforces one's faith that some metric theory of gravity is true. Repeated measurements, and independent measurements, of the same parameter thus constitute a kind of test of the background assumptions they rely on: Nature is given an opportunity to show us that we aren't really measuring what we thought we were measuring, that something must be wrong with our background assumptions. This does not constitute a severe test of background assumptions such as EEP, since in general we have no way of estimating the needed error probabilities. But it does constitute a "risky test" (in an intuitive sense that I won't try to make precise here), since we run the risk of having our background assumptions shown to be in error.

Here then, we have two striking features of the background assumptions

used in tests of theories of gravity, that aren't shared by all logically possible background assumptions: They make possible new kinds of empirical description of nature, and they run the risk of being shown to be in error by means of multiple and independent measurements of the same quantities. They do both of these things by virtue of making it possible to measure quantities that otherwise wouldn't be measurable. (I have argued elsewhere (Roberts 2005) that this features if exactly what is distinctive of *nomological* hypotheses in physical theories.) My final proposal is that severe tests and the evidence they afford be considered impressive only when the background assumptions they are relativized to have these same features.

The proposal represents a view of scientific methodology that could be considered a modest version of ESPOS. It analyzes quality of evidence in terms of severity of tests; it uses Mayo's two criteria of degree of fit and lowness of error probabilities in order to assess severity; it dispenses with the idea that subjective probabilities have any role to play in scientific methodology. Its modesty consists in its relativizing severity of tests and quality of evidence to background assumptions, a move Mayo has not made, and would perhaps be averse to making. But this move allows one to get around serious difficulties that ESPOS faces with respect to explaining how experimental testing relates to high-level theorizing, and as I have just tried to show, does not trivialize experimental methodology. None of this comes close to establishing that this modest version of ESPOS is correct, but it does seem to show that it is a view worth taking seriously.

References:

EARMAN, J. (1992): *Bayes or bust? A critical examination of bayesian confirmation theory* (Cambridge, MA, MIT Press).

FEYERABEND, P. K. (1962) Explanation, Reduction and Empiricism, in: H. FEIGL and G. MAXWELL (Eds.), *Minnesota studies in the philosophy of science*, volume 1 (Minneapolis, University of Minnesota Press).

GLYMOUR C. (1980) *Theory and evidence* (Princeton, Princeton University Press).

KUHN, T. S. (1996) *The structure of scientific revolutions*, third edition (Chicago, University of Chicago Press).

LAKATOS, I. (1978) *The methodology of scientific research programmes* (Cambridge, Cambridge University Press).

LAUDAN, L. (1997) How about bust? Factoring explanatory power back into theory evaluation, *Philosophy of Science*, 64, pp. 306-316.

LAUDAN, L. and LEPLIN, J. (1991) Empirical equivalence and underdetermination, *Journal of philosophy*, 88, pp. 449-472.

LONGINO, H. (1990) *Science as social knowledge: values and objectivity in*

scientific inquiry (Princeton, Princeton University Press.)

MAYO, D. G. (1996) *Error and the growth of experimental knowledge* (Chicago, University of Chicago Press).

MAYO, D. G. (1997) Response to Howson and Laudan, *Philosophy of Science*, 64, pp. 323-333.

ROBERTS, J. T. (2005) Measurability and Physical Laws. *Synthese* 144(3): 433-447.

WILL, C. (1993) *Theory and experiment in gravitational physics* (Cambridge, Cambridge University Press).