DEBORAH G. MAYO

CRITICAL RATIONALISM AND ITS FAILURE TO WITHSTAND CRITICAL SCRUTINY

PART I: THE SEVERE TESTING PRINCIPLE IN THE CRITICAL RATIONALIST PHILOSOPHY

1. INTRODUCTION

> Observations or experiments can be accepted as supporting a theory (or a hypothesis, or a scientific assertion) only if these observations or experiments are severe tests of the theory—or in other words, only if they result from serious attempts to refute the theory, and especially from trying to find faults where these might be expected in the light of all our knowledge. (Popper, 1994, p. 89)

The lack of progress in the neo-Popperian philosophy known as 'critical rationalism' may be traced to its inability to show the acceptability of the fundamental principle underlying the above quote:

Severity Principle (SP)	Data **x** count as evidence in support of a hypothesis or claim H, only if **x** constitute severe tests of H—only if data **x** (which are in accord with H) result from serious attempts to refute H.

This failure seems deeply puzzling, given the intuitive plausibility of SP, as in Popper's exhortation above. The problem is hardly limited to critical rationalists. Something like SP is endorsed far more generally in philosophy as well as in science, and yet it has been notoriously difficult to actually cash out what 'surviving serious criticism' demands, and why H's surviving the 'ordeal' is good evidence for H. My focus here is on critical rationalists, and in particular on Alan Musgrave's recent (1999) attempt.

C. Cheyne & J. Worrall (eds.), Rationality and Reality, 1—96.

What gives SP its plausible-sounding ring is the supposition that 'H's surviving serious criticism' is being used in the way it is ordinarily meant: roughly, that H has been put to a scrutiny that would have (or would almost certainly have) uncovered the falsity of (or errors in) H, and yet H emerged unscathed, i.e., that H has survived a highly reliable probe of the ways in which H might be false. However, critical rationalists, as they freely admit, do not have resources to articulate anything like 'reliable error probes', and even deny the reliability of the method they espouse. Despite exhortations as in the epigraph from Popper, critical rationalists only espouse a weaker, *comparativist principle CR*:

(CR) It is reasonable to adopt or believe a claim or theory P which *best survives* serious criticism.

But without being able to say that surviving the critical rationalist's actually affords evidence for P, a 'best surviving' claim may still have been very poorly probed, and thus P may be 'best tested' with **x**, even though **x** actually provides scant evidence for P at all.
So, while we may (and most of us do) accept the intuitive principle that CR is supposed to capture (namely the severity principle SP), we have yet to be given grounds to accept CR as instantiating the intended severity requirement. To simply declare CR is a reasonable epistemic principle without giving evidence that following it advances any epistemic goals is entirely unsatisfactory, and decidedly un-Popperian in spirit. So it does not help for the Popperian to insist 'there is no more rational procedure' than to prefer a hypothesis that is well-corroborated, i.e., that has withstood serious or severe criticism (Popper 1962, p. 51), without demonstrating the existence of testing methods that are actually severe. Yet, far from demonstrating the existence of severe error probes (or whatever one wishes to call them), the critical rationalist feels bound to deny that tests that are severe in the critical rationalist's sense are reliable tools for uncovering errors. The critical rationalist is thus guilty of a kind of 'bait and switch', getting our nod for plausible sounding exhortations, as in SP, but then serving up, not the robustly severe tests we thought we were getting, but 'tests' incapable of doing their intended job.

Granted, Popper invites this problem, due in part to his efforts to distinguish himself from the 'inductivists' of the time. The deductive resources to which Popper limited himself allows neither substantiating a claim to actually *have* a severe test or error probe, nor to say that the probability of P's passing test T is low, given P is false. Now that we know so much more about conducting severe testing in experimental practice than was evident through logical-empiricist blinders, one would have expected this weakness to be remedied by Popper's critical rationalist followers. Surprisingly, it has not been. Even so astute a thinker and upholder of common sense as Alan Musgrave (1999) has recently mounted a defence of CR that he openly concedes is circular, admitting, as he does, that such circular defences could likewise be used to argue for principles he himself regards as 'crazy'. (Why is CR a good rule? Because it is a good rule.) As if this were not bad enough, it turns out we cannot even self-referentially apply rule CR, i.e., we cannot show that CR itself is a 'best tested' rule, because it is demonstrably unreliable (and other methods are not). While Musgrave's full argument is subtle and clever, these concessions, or so I shall argue, radically undermine his goal, as they render his argument no argument at all.

I will expose the series of missteps that have landed the critical rationalist in this untenable position. In Part I, I will argue that the critical rationalist arguments, as urged by Musgrave, themselves rest on the ability to distinguish severe from insevere tests and reliable from unreliable error probes, and thus are self-contradictory when denying the possibility of doing any such thing. In Part II, I will show how to rectify the situation by (a) rejecting the erroneous conceptions of inductive or 'evidence-transcending' inference upon which their sceptical slide is based, and (b) showing how to develop an account with the resources to define and apply severe or reliable error probes. The main points for which I will be arguing are these:

1. An adequate defence of CR must characterize 'withstanding severe or serious scrutiny', and show it corresponds to classifying claims reliably, which neither Popper nor current day 'critical rationalists' have done.

2. Musgrave's argument that all epistemic principles can only be defended circularly, if they are defended at all, is unsound, and confuses 'self-subsuming' methods with 'self-sealing' (circular) methods.

3. In distinguishing 'crazy' and 'non-crazy' methods, Musgrave must assume a reliable classification scheme, which, if drawn out, already goes several steps further than what is alleged by the critical rationalist.

4. Critical rationalists assume falsely that justifying claim P requires either showing it to be true or probable.

5. A satisfactory articulation of withstanding a severe test can achieve the intended goals, without illicit 'justificationist' or metaphysical inductive appeals.

2. BETWEEN SKEPTICISM AND IRRATIONALISM: THE WEDGE IS NOT ENOUGH

The reason that there are only a dozen odd self-styled 'Popperians' in philosophy of science these days, Musgrave ventures, is that 'Popper's chief contribution to philosophy has still not been understood' (p. 314), in particular, philosophers of science have failed to appreciate 'Popper's critical rationalism and the solution to the problem of induction which it contains.' He sets out to rectify this. The secret, in Musgrave's view, is to appreciate fully the way critical rationalism drives a 'wedge' between skepticism and irrationalism (e.g. 1999, p. 322): we can be skeptics about inductively inferring or warranting hypothesis or claim P, but still regard it as reasonable to believe P, or to put it in Popper's locution, to accept or 'prefer' P.

2.1 The Probabilist's View of Justifying Claims is Rejected

What enables this 'wedge' to be 'driven between skepticism and irrationalism' (p. 322), Musgrave thinks, is the critical rationalist's rejection of the traditional justificationist's principle (J) (p. 321):

> (J) A's believing that P is reasonable if and only if A can justify P, that is, give a conclusive or inconclusive reason for P, that is, establish that P is true or probable.

Blithely accepting that an 'inconclusive reason' for P must be understood as assigning P a probability, Musgrave touts a philosophy wherein 'P is reasonable' never means having to say there's even an

inconclusive reason for P.[1] Later on, I will come back to question and reject this conception of warranting an 'evidence-transcending' claim, but for now we want to trace out Musgrave's reasoning, and his reasoning is this:
By rejecting principle (J), the critical rationalist is free to hold:

(1) it is or may be reasonable to prefer hypothesis P despite the lack of any warrant for the truth of P.

Quoting Popper: 'although we cannot justify a theory…we can sometimes justify our *preference* for one theory over another; for example if its degree of corroboration is greater' (Popper 1976, p. 104), Musgrave proposes to replace Popper's 'justify our preference for' P with 'show the reasonableness of believing P'.[2] So, Musgrave replaces (1) with:

(1´) it may be reasonable to believe hypothesis P despite the lack of any warrant for the truth of P.

Although my own preference is to avoid talk of beliefs altogether (and (1´) is more contentious than (1)), since Musgrave seems prepared to identify belief in P with adopt P as true (p. 327), I will follow his terminology in this throughout Part I.

2.2 Some 'Unsavoury' Wedges

Musgrave cites, as precedents, examples that help to construe the manner in which one can uphold (1´): 'Pascal's wager gives a reason for believing God exists which is not a reason for God's existence. The pragmatic vindication of induction is a reason for believing that nature is uniform which is not a reason for the uniformity of nature' (p.322). This suggests a clearer and less contentious formulation of the thesis in (1'):

---

1 In speaking of P's truth, there is no realist assumption. That P is true or correct may be cashed out in terms of a specific error P asserts to be absent. It may mean that a given claim is adequate in a number of senses: that an assertion about a genuine effect, a causal factor, or a parameter estimate is correct, possibly with margins of error attached (in quantifiable cases).

2 The use of 'justify' here presumably does not mean show it has a high probability—even critical rationalists have trouble keeping up the linguistic summersaults their defences require.

(1´) we may have information **x** that shows belief in P to be reasonable, even though **x** does not show P to be true or probable.

By upholding thesis (1´), Musgrave argues, the critical rationalist can concede there are no good reasons for inferring an evidence-transcending hypothesis P, while nevertheless maintaining that it is reasonable to believe that P.
In particular, we may have a method or procedure (M) for classifying claims as reasonable or not. That is:

M: P $\rightarrow$ {reasonable, unreasonable}

Associated with each such M is an epistemic principle (EM):

(EM) It is reasonable to believe P if and only if P is classified as reasonable by method M, i.e., iff P satisfies a criterion set out by method M.

We can readily agree with Musgrave that P may be classified as reasonable to believe by method M, even without there being reasons for regarding P as true or probable, while still demanding that the chosen classification method M have some warrant or justification.[3]
He himself claims his two illustrative examples 'are unsavoury ones', though one wishes he had explained why. Does he regard them as unsavoury because in each case the reasons for belief are merely pragmatic and do not supply support for the truth of the claims in question (God's existence, nature is uniform, respectively)? That would seem strange, since the whole point of this exercise is to uphold the idea that data **x** may give perfectly good reasons for believing P although **x** fails to supply reasons that P is the case.
In fact, there are features of these 'unsavoury wedges', at least in their intent, that would seem to offer the kind of strategy that would appeal to a critical rationalist. Their linchpin, after all, is their claim to demonstrate that *whether or not the belief in question is true*, there is a payoff attached to adopting a given attitude with respect to P. At times, Popper himself drops hints along these lines in arguing for CR (e.g., 'if we have made it our task…' (1962. p. 51)). Perhaps their

[3] Although I will strive mightily to abide by the language the critical rationalist wants us to adopt, I see no reason to share his fear that using the word 'justification' will force me to adopt enumerative induction. For me it is just a synonym for 'warrant'.

unsavouriness, then, is that they fail to ensure the promised payoff (pragmatic or epistemological)?

It is of interest to note that contemporary statistical hypothesis testing, e.g., Neyman-Pearson (NP) tests, exemplify Musgrave's wedge: tests use data **x** to classify statistical hypotheses 'acceptable' or not, without assigning them degrees of probability; however, they will be regarded as good tests only insofar as it can be shown they very infrequently classify false hypotheses as true (or true hypotheses false), i.e., they must be shown to be reliable in this sense, namely, they have low error probabilities. Tests with high error probabilities are 'unsavoury'. (I return to this in Part II.)

Thus, merely giving us a 'wedge' (between evidence for the reasonableness of believing in P versus evidence for the truth of P) does not take Musgrave very far. We are led to the question of what if any grounds Musgrave provides for the 'belief-adoption' method M championed by the critical rationalist, What grounds are there that CR is not also unsavoury?

3. M_{CR} IS UNRELIABLE

The critical-rationalist position that Musgrave endorses espouses the following method:

(M_{CR}) P satisfies classification method M_{CR} (at time t) if and only if P has best withstood serious criticism (at time t).

The corresponding epistemic principle is that it is reasonable or rational to prefer or believe the comparatively best tested P. That is, the epistemic principle corresponding to method M_{CR}, which we may write as EM_{CR}, is CR, only now he writes it with a time index:

(EM_{CR}) It is reasonable to believe P (at time t) if and only if P has best withstood serious criticism (at t).

The particular 'wedge' offered by CR, then, is this: "if a hypothesis has withstood our best efforts to show that it is false, then this is a good reason to believe it *but not a good reason for the hypothesis itself"* (Musgrave 1999, p. 322*)*.

Even granting the 'wedge', surely the critical rationalist (of the Popperian or Musgrave stripe) wishes to incorporate certain requirements or demands which must be satisfied before it can be said to be reasonable or rationale to believe P, and surely, then, they must regard CR as embodying a method capable of promoting those aims. The question then is: what aims does M_{CR} achieve, such that it makes sense to adopt this principle?
To sum-up this part, we can grant Musgrave's instantiation of (1′):

(1′) we may have information **x** that P is best tested[4] (by method M_{CR}), even though **x** does not show P to be true or probable.

But we deny **x** shows belief in P to be reasonable if it turns out that M_{CR} is unreliable. That is, we insist on:

(2) if with high probability method M_{CR} deems P 'best tested' even if P is false, then the passing result **x** fails to show belief in P to be reasonable.

As we will see, method M_{CR} may deem P 'best tested' even if little or nothing has been done to uncover the ways P can be in error, i.e., even if the 'test' would be regarded as having little or no 'severity' at all. If this is so, then M_{CR} may fail its intended task of capturing the severity principle (SP) with which we began.

3.1 CR Fails to Give a Necessary Condition for Reasonable Belief

Remembering that CR is equivalent to our EM_{CR}, we see that the 'only if' in EM_{CR} is false:

(CR⇒) It is reasonable to believe P (at time t) only if P has best withstood serious criticism (at t).

To satisfy the antecedent, it is required only that there be reasons, **x**, to believe P (at t), and by Musgrave's 'wedge', we need not expect that **x** supplies reasons in support of P's truth. So, **x** might be reasons of prudence, pragmatics, or any number of things. Consider for example information **x**:

---

[4] We substitute in (1′) as Musgrave directs us to: replacing '**x** shows belief in P to be reasonable' with '**x** shows that P is best tested' (i.e., has best survived the critical rationalist's notion of a severe test).

(**x**) evidence from medical trials shows a high correlation between tolerating the treatment given for disease D and adopting an optimistic belief that one can make disease D vanish by will.

Proposition P is that one can make disease D vanish by will. Evidence **x** may make it reasonable for a patient with D to believe P, even where **x** does not constitute any evidence that P has withstood serious criticism or any kind of criticism. Indeed, P may have failed tests (suppose no patient has ever been able to will disease D away). But it may be prudent to believe P in order to better tolerate the treatment. Or it may be reasonable to believe P if there is evidence **x** that one will otherwise be killed, even where **x** is not evidence that P has survived any kind of criticism or probe of P's falsity. In fact, P may be known to be false.

Thus, CR as an 'if and only if' claim is plainly false: having withstood serious criticism is not a necessary condition for reasonable belief, as Musgrave understands the 'wedge'. But since it is the 'if' claim that seems mostly to be doing the work for Musgrave, we can put this qualm aside for now. However, the 'if' clause, on which Musgrave's argument depends, is also highly problematic.

3.2 Having 'Best Withstood Criticism' is not Sufficient for Reasonable Belief

I will argue that it is also false to claim that:

(CR⇐) If P has best withstood serious criticism (at time t), then it is reasonable to believe P (at t).

It is very important, in evaluating CR to consider, not what we would ordinarily mean by surviving serious criticism, because, as already said, this assumes tests with capabilities that the critical rationalist has no intention of supplying. To begin with, the comparative nature of the rule entitles P to receive the 'best-tested' medal, even if poorly tested, it may be the first ever tested, or slightly less poorly tested than an existing rival! But such a comparative principle of testing is highly unreliable (Mayo 1996). I return to this in Section 10.

In deeming M_{CR} unreliable, I mean that P may be the best-tested so far without P having been probed in the least, and thus it would seem that this does not suffice for it to be reasonable to believe P. (Even if there

are other, non-evidential reasons to believe P, e.g., pragmatic considerations, this is no thanks to the antecedent being satisfied.) Why then do critical rationalists settle for comparativist method M_{CR} when principle SP is non-comparative? Presumably, it is felt that the comparatively-best-tested principle is all that can be demanded if the principle is to be applicable. But this just underscores the fact that 'best-tested' in the critical rationalist's sense, need not mean well-tested at all (else the non-comparativist principle SP would be retained).

3.3 Popper on Severe Testing and Corroboration

This comparativism was clearly embraced by Popper. According to Popper, hypothesis P best survives test T with data **x** so long as:

(i) P entails (or otherwise 'fits') **x**

and

(ii) **x** is not predicted, or is counterpredicted, by P's existing rival(s).[5]

As Popper's critics observed from the start (e.g., Gruenbaum 1978) satisfying (ii) does not warrant stronger claims such as:

(ii´) **x** would not be expected were P false

or

(ii´´) there is a low probability of **x**, given that P is false

although at times Popper suggested it did. The reason is that (as Popper was aware) 'P is false' includes the disjunction of all possible hypotheses or claims other than P that would also 'fit' or accord with **x**—the so-called 'catchall hypothesis'—including those not even thought of. Existing data **x** would be just as probable were one of the catchalls true, and P false.

---

[5] As Musgrave has elsewhere noted, Popper's condition (ii) may be construed even more weakly, allowing it to be satisfied even if existing alternatives to P say nothing about the phenomenon in **x**. P may pass all the tests that rival(s) P´ do, even if all are silent about certain results.

Therefore, P may be the 'best-tested' hypothesis so far, even without P's having been probed especially well at all. So long as P is not falsified, even if no alternative to P exists, P would, on this requirement, count as 'best-tested', or so it seems. But why should it be reasonable to believe in the first hypothesis put forward, say, to account for a phenomenon? Or believe in a full blown theory when only a small portion has been tested? (Mayo 2002a, Laudan 1997). The intuition behind the severity demand is that mere accordance between **x** and P—mere survival of P—is insufficient for taking **x** as genuine evidence for P. Such survival must be something *that is very difficult to achieve* if in fact P deviates from the truth (about the phenomena in question). The intuition is sound, but Popperian logical computations between statements of hypotheses and data never gave us a way to characterize severity adequately. (In part II, I shall describe an account that enables the needed characterization.) Popper himself seemed to concede that the various formal definitions C(P,**x**) he proffered were only *potential* measures of the degree to which **x** corroborates P: in order for it to genuinely measure corroboration, Popper claimed, **x** would have to *actually* be the result of a severe test, a notion which was perhaps beyond formalization.

> In opposition to [the] inductivist attitude, I assert that C(P,**x**) must not be interpreted as the degree of corroboration of P by **x**, unless **x** reports the results of our sincere efforts to overthrow P. The requirement of sincerity cannot be formalized—no more than the inductivist requirement that **x** must represent our total observational knowledge. (Popper 1959, p. 418. I substitute his h with P and e with **x** for consistency with Musgrave's notation.)

The important kernal of rightness here is that these inductive logics make it too easy to find evidence in support for hypotheses *without satisfying the requirement of severity.* Unfortunately, Popper's computations suffered from just this weakness.

3.4 How Might Musgrave Respond?

Now Musgrave might respond in two ways:

(a) He might maintain that he (and other critical rationalists) do or would go beyond Popper by fleshing out the demand that '**x** report the results of our sincere efforts to overthrow' or find fault with P.

But how? I doubt he would be satisfied with some sort of subjective or psychologistic 'sincerity' requirement that could not be intersubjectively checked (Musgrave 1974b). Musgrave has, after all, been a long-time proponent of one or another 'objective' novelty requirements, and he might maintain that this allows him to exclude problematic cases. For example, he might classify under 'not sincerely trying to find fault with P' cases where P has been deliberately constructed to account for given data **x**, and no independent evidence for P exists (Musgrave 1974). But the novelty requirement Musgrave endorses, 'theoretical novelty',[6] boils down to Popper's comparatively best-tested requirement (3.3); and, as noted, this fails to provide tests that are actually severe and is, moreover, neither necessary nor sufficient for good evidence (see Mayo 1996).

Finally, even if one granted a given test was a severe and reliable probe of errors, we would still be in need of an account of how to obtain evidence **x** that P has actually *withstood* this test. This a non-trivial task that (as Popper admits) demands evidence of a 'reproducible' or reliable effect, not merely 'non-reproducible single occurrences' (Popper 1959, p. 86). (The mere perceptual claims that Musgrave (1999) is prepared to accept so long as they are not known to fail scrutiny will hardly do.) Musgrave's remark that 'existing critical rationalist literature goes a good way to provide [a theory of criticism]' (ibid. p. 323) must remain a mystery: one finds nothing approaching such a thing in that literature.

Indeed, the whole 'secret' to critical rationalism, as he sees it, is that it escapes the demanding task of developing an account of severe testing that can be shown to be reliable.

(b) Musgrave might, in this vein, insist he cares nothing for the reliability of method M_{CR}.

'Critical rationalists' Musgrave tells us 'deny that the process they commend is reliable' (p. 346). But this will not do. It is one thing to deny one is 'commending' a method as reliable*; it is quite another to be confronted with the blatant unreliability of a method and yet deny*

[6] Musgrave also propounded a notion of 'deductive novelty' which demanded being able to identify if data **x** were required as premises in constructing P (Musgrave 1989). For a discussion of the relationship between novelty and severity see Mayo 1991, 1996.

that this matters. Further, since accounts of severe testing exist which are not unreliable, it follows that the critical rationalist's testing method M_{CR} itself fails to survive even moderately severe scrutiny! (More on this later.)

4. THE FALSITY OF THE ALLEGED NECESSITY OF CIRCULAR DEFENCES OF EPISTEMIC PRINCIPLES

Musgrave does spend considerable effort addressing the question of how to defend the critical rationalists' epistemic principle, but remarkably, he does not take up concerns such as those I have just raised. Notably, he does not seem to think he has to. In handling the question, "Why is P's being best-tested (as the critical rationalist understands this) a reason for believing P?" he allows he can do no better than simply repeat the epistemic principle under question!

> There is nothing more rational than a thorough and searching critical discussion. Such a discussion may provide us with the best reason there is for believing (tentatively) that a hypothesis is true—though not, of course, with a conclusive or inconclusive reason for that hypothesis. (p. 324)

Buying the traditional inductivists' (probabilistic) notion of 'a reason' for a hypothesis ((J) above), Musgrave is forced to embrace circularity. Of course this is just what Popper said, the difference is that Musgrave wishes to bite the bullet of circularity. But are we to accept that the promised cornerstone for avoiding irrationality is no more than a declaration that there is a method M such that, by definition, M is rational? Amazingly, it seems that, according to Musgrave, we are:

> Even if it is accepted that CR withstands criticism better than rival epistemic principles (a big 'if'), another objection immediately presents itself. All this is circular! The critical rationalist is saying that it is reasonable to adopt CR by CR's own standard of when it is reasonable to adopt something! (p. 330)

Not that he is happy about it. Indeed, Musgrave concedes that an analogous circular move would countenance arguing for the reasonableness of so crazy a method as:

M_{Mus}: It is reasonable to believe anything said in a paper by Alan Musgrave

since the assertion is made in a paper by Musgrave (p. 330). (I return to this 'crazy' method in Section 5.) Nevertheless, Musgrave declares

that as all epistemic principles can only be defended circularly, it is no special reason to find fault with critical rationalism! According to this, we know in advance that no epistemic principle could be faulted, thanks to the availability of its surviving a circular defence. It would follow that claims about methods are non-testable! Could this really be the long-sought for defence of Popper?

4.1 Musgrave's Remarkable Argument

It is easier to express horror at the final destination of Musgrave's reasoning than it is to show just where one is warranted in getting off his train of argument. His argument, while unsound, or so I shall argue, is subtle and interesting. His argument is this:

> Any general epistemic principle is either acceptable by its own lights (circularity), acceptable by other lights (hence irrational by its own lights and inviting an infinite regress), or not rationally acceptable at all (irrational again). So even though the rational adoption of CR involves circularity, this cannot be used to discriminate against it and in favour of some rival theory of rationality. (p. 331)

Although our interest is in CR, let us analyse this striking general argument. Any general epistemic principle is either

(A) acceptable by its own lights (circularity), or

(B) acceptable by other lights (hence irrational by its own lights and inviting an infinite regress), or

(C) not rationally acceptable at all.

So if a general epistemological principle is rationally acceptable (i.e., (C) is false), he concludes, either (A) or (B) is the case (i.e., its acceptability will be circular, or irrational and inviting a regress). By a general epistemic principle, Musgrave has in mind an 'if-and-only-if' claim (doubtless the reason he expressed EM_{CR} as such), so as to ensure the claim about the acceptability of the principle comes under the principle itself, i.e., that it is *self-subsuming*. To make his intent clear, the form of a 'general epistemological principle' EM is this:

(EM) a claim P is acceptable iff it is classified as acceptable or beliefworthy by belief-classification method M.

EM, itself being a claim, would be subsumed under method M. By contrast, an epistemic principle concerning, say, purely mathematical claims, would not itself be a mathematical claim and so would not be self-subsuming.

To engage his argument, and see how it goes wrong, let us resist drawing any distinctions of levels and grant Musgrave's claim that a general epistemic principle itself comes under the domain of claims that M classifies as acceptable or not, let us grant that it is self-subsuming. (He equates 'self-subsuming' with 'circular' but, as we will see, the latter term has importantly different connotations.) Let us suppose, with respect to a given method M, that premise (C) is false: M *is* rationally acceptable , and allow that M is itself classified as beliefworthy by M, i.e., EM is acceptable 'by its own lights', premise (A).

But this does not yet entail that the *only* warrant for EM is EM, i.e., the only warrant for EM is that it has been classified as 'acceptable' by M! Musgrave confuses a 'self-subsuming' method with what may be called a 'self-sealing' method.

4.2 Self-Subsuming is Not Self-Sealing

Consider an example which I shall try to design so as to concede as much as possible to Musgrave. There is a principle, let us imagine, for deciding to accept or believe claims about books in print in 2004:

(EM_{BIP}) accept claims concerning which books are in print (in year 2004) if and only if they are found in the comprehensive *Handbook of Books in Print* for 2004 (BIP),

where we stipulate, for purposes of the illustration, that it really is exhaustive of the finitely many books in print in 2004.

Again, let us resist any attempt to suggest EM_{BIP} is itself a 'meta-claim', and allow that it is subsumed under itself. Suppose in fact that the assertion EM_{BIP} is found on the first page of the BIP. So EM_{BIP} is acceptable 'by its own lights', but does this entail any circularity? No. Being 'self-consistent' is not the same as being 'self-warranting', i.e., warranted only by dint of the self-subsumption: *self-subsuming is not self-sealing*. In arguing for the acceptability of EM_{BIP}, one might allude to such things as the scrupulousness with which each publisher

is checked to keep the listing of books in print up to date. In other words, one would allude to the reasons that assertions find their way into the BIP handbook to begin with, the criteria which must be met before inclusion, thereby ensuring that all claims therein have certain qualities (examples such as this can easily be multiplied). According to Musgrave, appealing to these various facts about the criteria used to include assertions in the BIP handbook instantiates premise (B) rendering the warrant for BIP irrational and/or leading to an infinite regress! But this is clearly false.

Compare this with recommending *Sloppy Joe's Books in Print* which we may imagine is very sloppy, imcomplete and outdated.

(EM $_{\text{SJ'S BIP}}$) accept claims concerning which books are in print (in year 2004) if and only if they are found in *Sloppy Joe's Books in Print.*

And again suppose this assertion is itself on page one of Sloppy Joe's volume. Following Musgrave's reasoning, even *Sloppy Joe's Books in Print* would be as acceptable as the authoritative BIP! But in fact, we would adduce many reasons for regarding its listing as unreliable, out of date, and so on.

Musgrave's argument presents us with a false dilemma: it appears to go through only by assuming that if epistemic principle EM is acceptable then the *only* warrant that may be given for the claims M classifies as beliefworthy is the fact that M classifies them as beliefworthy!

But the if and only if claim in EM does not entail this. He is confusing self-subsumption with self-sealing. In other words, the left-to-right conditional in EM is:

(EM⇒) P is acceptable only if P is classified as beliefworthy by method M,

Musgrave conflates this with an entirely different claim, one asserting that any test of M is self-sealing:

Self-Sealing Test of M: P is acceptable *only because* P is classified as beliefworthy by method M.

The latter claim asserts that the only grounds for the acceptability of P is that P is classified as acceptable by method M. Were the 'self-

sealing' test the only one available for a method (for classifying claims as acceptable), then Musgrave would be right to allege that his defence of CR is no worse off than for any other. But this is false.

4.3 Sum-up of the Confusion Between Self-Subsuming and Self-Warranting

We see that Musgrave's remarkable argument assumes and does not show that *only* a self-sealing defence is possible for method M (and thus for an epistemic principle espousing M). In so doing, Musgrave assumes the very thing he is claiming to argue for, i.e., he is guilty of question-begging. Moreover, since we have seen there are grounds to reject his claim, his question-begging adherence to it has no weight. That a principle is not self-refuting hardly entails that there is no test or means of scrutiny (independent of the classification scheme itself) of whether the method in question is, or is not, capable of satisfying the desired aims in applying method M. Otherwise we would not have been able to mount the criticism in Section 4 of the comparativist account of severe tests, nor criticize *Sloppy Joe's Books in Print*. Nothing in Musgrave's arguments show otherwise.

It is not that Musgrave is not pained by having to assume his favored principle in order to (deductively) defend it. He is. If we permit epistemic principles whose sole support is circular, Musgrave freely admits, we can easily argue in favour of all manner of crazy procedures such as procedure M_{Mus}:

M_{Mus}: It is reasonable to believe anything said in a paper by Musgrave

since M_{Mus} occurs in Musgrave's paper. I feel his pain, and have been setting the stage for its extirpation. Before administering the anesthesia, however, let us twist the knife a bit further—to learn more about the critical-rationalist infirmity with which Musgrave saddles himself, and just how devastating the malady really is. Were self-sealing defences the most one could give for epistemic principles, then the critical rationalist should close up shop: he would have to concede there are no better grounds for CR than for any other principle, even one that ignores all evidence and counsels accepting whatever Musgrave endorses!

Thus, Musgrave's defence of critical rationalism defeats itself. For, what is wrong with a self-sealing test of a general epistemic principle EM (about method M)? What is wrong is that the epistemic principle is guaranteed to pass even if it is false. Even if method M does not satisfy its intended aim, *whatever it is*, it will nevertheless still be permissible to classify M as an acceptable method. That EM passes a self-sealing test is tantamount to its passing a test it had no risk of failing, a test that utterly lacks severity or probative power in the ordinary sense upon which CR is parasitic. What is more, his arguments are self-contradictory. My goal in showing this, I should emphasize, is a positive one: to show how to get beyond where critical rationalism thinks it can go.

In at least two places Musgrave's arguments assume the existence of reliable methods: (i) he assumes there are non-crazy methods, ones that are at least not utterly unreliable for the intended job of uncovering flaws and errors, and (ii) he also assumes there is a reliable method for distinguishing crazy methods from non-crazy methods (for classifying claims as acceptable). But since at the same time he denies these assumptions, his arguments are self-contradictory. Moreover, suppose we perform this substitution in M_{CR}: replace 'P is best tested' with the phrase 'P is endorsed in a paper by Musgrave', yielding method M_{CR*}. Now M_{CR*} is identical to M_{Mus}. Cashed out this way, Musgrave would presumably deny the corresponding principle EM_{CR*} and he would adduce non-circular reasons for this.

5. WHY BELIEVE THAT 'BELIEVING WHATEVER MUSGRAVE WRITES' IS A CRAZY RULE?

Musgrave declares that '"It is reasonable to believe anything said in a paper by Alan Musgrave"…is a crazy epistemic principle' (1999, p. 330), and I want to know why. He evidently regards its craziness as fairly obvious, and so I believe he has reasons for this judgment. I take it that he does not regard all epistemic principles for adopting beliefs as similarly crazy, else he would not be mounting efforts to argue in favor of the critical-rationalist epistemic principle (CR). For sure, it is bizarre to hold, as he seems to in his circularity concession, that CR

has no better grounds than does M_{Mus}, while at the same time denying, as we may presume he does, that the two are similarly crazy.[7] Moreover, it is implicit in Musgrave's discussion that there are some criteria for *distinguishing* (crazy from non-crazy) epistemic principles such that CR withstands this scrutiny and rules like M_{Mus} do not. The scrutiny cannot be a matter of whether they may be defended non-circularly, since we have already seen he denies that (even though we have rejected his arguments). What I want to consider is what Musgrave *could mean* in making a distinction between crazy and non-crazy belief-classification rules. Articulating the grounds behind his self-deprecating critique, ironically, takes us several steps further than he declares is possible toward an account of 'evidence transcending' or 'inductive' inference. But, in so doing, we expose a contradiction in critical rationalism, at least as he describes it, and the self-defeating nature of his defence of it.

5.1 Musgrave's Method for Classifying (belief-classification) Methods as Crazy or Not

If Musgrave does not regard all epistemic principles as crazy, if, for example, he does not regard following M_{CR} as just as crazy as following M_{Mus}, then, he must have a procedure, or criteria to apply, that discriminates crazy from non-crazy methods, or is at least capable of identifying a clearly crazy one. But then it would seem that there is at least one method that may be defended with good reason: Musgrave's method for condemning following M_{Mus} as 'crazy'. In other words, Musgrave has a method, or discriminating capacity, that pigeonholes under the rubric 'crazy' method M_{Mus} and under 'not crazy' (but rather, rational) method M_{CR}.

What makes the rule M_{Mus} crazy? Why is it *correct* to classify it under the rubric 'crazy'? Why is it *incorrect* to classify it as 'rational' (or non-crazy)?

Does Musgrave classify it as crazy because he thinks it an unreliable procedure to follow (i.e., that Musgrave often publishes flawed or incorrect claims)? Or because the mere fact that a paper by Musgrave

---

[7] He cites two other crazy or unwarranted epistemic principles : 'Granny told me I ought to believe everything she tells me', and 'The Pope declared *ex cathedra* that everything declared *ex cathedra* by the Pope is a matter of faith' (p. 330).

claims P does not, in and of itself, provide evidence that P is correct or well-supported? But these all seem to be at odds with his insistence upon the 'wedge'. So perhaps Musgrave regards it as a crazy method because the mere fact that a paper by Musgrave claims P does not, in and of itself, make it reasonable for others to accept or believe P (where this may be construed pragmatically or otherwise).
But we need not pretend to know what classification method Musgrave is using here for the argument that I am now interested in making. Musgrave must allow there is an adequate 'metamethod' whereby he classifies M_{Mus} as crazy, in contrast to an inadequate metamethod, say, a procedure that willy-nilly classified as crazy any and all rules for adopting beliefs, or made the determination by flipping a coin. Compare two metamethods:

($MetaM_{Mus}$) accept claims about whether or not a method (for belief-classification) is crazy in accordance with Musgrave's pronouncements about (crazy/non-crazy)

where this alludes to whatever classification method Musgrave is using in this paper (Musgrave 1999); and one based on, say, coin-flipping:

($MetaM_{Coin}$) accept claims about whether or not a method (for belief-classification) is crazy in accordance with the outcome of a fair coin toss.

Methods like $MetaM_{Coin}$, presumably, would not be adequate for the task of discriminating crazy from non-crazy methods. That is because the latter procedures are poor tools for accomplishing the intended job (of correctly classifying rules as crazy). (Indeed, they are themselves crazy tools for the job!) Since Musgrave clearly thinks there are good reasons for regarding M_{Mus} as crazy, I should think he would regard as a poor metamethod one that declares M_{Mus} a non-crazy rule; or one that bases its pronouncements on irrelevancies (e.g., coin flips, or whether its adoption as a good rule would make money for Musgrave). A possible criterion, then, for evaluating metamethods might be:

(5.1) *Criterion for Evaluating Metamethods*: MetaM is a poor classification method if it erroneously classifies methods as crazy as often as not.

For example it would be poor if the test it employs to decide whether to classify a method as crazy uses a criterion with no correlation to the method's actually being crazy (however this is defined).

Now I do not know what test rule Musgrave is applying so as to declare M_{Mus} crazy, my point is only that I believe he has one and that were he to spell out the criterion behind it, we may evaluate its properties for performing the job at hand. It would follow that there are perfectly good, *non-circular*, reasons for endorsing some methods for classifying methods as crazy or not, while rejecting others as not up to the job.

5.2 Learning From the Failure of Musgrave's Defence

Our critique of Musgrave's attempted defence of critical rationalism bears positive fruits, as severe critiques should. A method for classifying methods (what I called a metamethod) is precisely on par with any method for classifying claims as acceptable or not, so from the above discussion, and the criterion in (5.1), we extract the following:

(5.2) Method M is a poor classification method (a crazy method) if it often classifies claims as acceptable when they are not, i.e., if its classification scheme is an unreliable indicator that P is acceptable (however one defines acceptability).

Even so weak an assertion as (5.2), which is itself just a start, already breaks through the critical rationalist's obstacles to progress. To begin with it gives a basis for distinguishing 'crazy' and 'non-crazy' methods, as well as grounds for criticizing arguments claiming to show why a given method is acceptable. Unless an argument gives assurance that a method avoids threats of unreliability, it fails utterly as a defence of the method. Musgrave's defence of the critical rationalist's classification method fails on these grounds: being classified as 'best-tested' by his critical rationalist makes it too easy to classify claims 'acceptable' without warrant.

Musgrave mistakenly assumes that demonstrating reliability would be tantamount to justifying enumerative induction, but enumerative induction, Musgrave declares, is 'unreliable' (p. 346), in contrast to perceptual beliefs which he claims are 'reliable'. Assuming the only

kind of justification an evidence transcending claim can receive is to find it true or highly probable i.e., accepting (J), he rejects justifying ampliative inferences altogether.

> Critical rationalists deny that induction is a reliable process. Critical rationalists also deny that the process they commend is reliable—or at least, they must deny this if they [are] to avoid the widespread accusation that they smuggle into their theory either inductive reasoning or some metaphysical inductive principle. (pp. 246-247)

I shall now turn to showing how to characterise the severe testing requirement, avoiding all the shortcomings of the critical rationalist. Viewing induction in terms of severe testing, as I define it, lets us warrant induction or evidence transcending methods as reliable without smuggling in 'probabilism' or a metaphysical inductive principle!

PART II. THE SEVERE TESTING PRINCIPLE IN THE ERROR STATISTICAL PHILOSOPHY

6. HIGHLY PROBABLE VERSUS HIGHLY PROBED

The modern-day critical rationalist (Musgrave being the best among them) has not cleared away the stumbling blocks that stymied Popper; like the 'inductive logician' of old he retains the assumption that a justification or warrant for an evidence transcending inference is either to show it conclusively true (whatever that might mean) or assign it a probability. Denying the former, the inductivist looks for a probabilistic computation, most often by appealing to the statistical definition of conditional probability or *Bayes's Theorem*: P(H|e) = P(e|H)P(H)/P(e)[8]. Computing P(H|e), the *posterior probability*, requires starting out with a probability assignment to all of the members of 'not-H,' the *prior probabilities*. Insofar as the computed degrees of confirmation are viewed as analytic and a priori—as in the 'logical probability' notions often favoured by Popperians—their relevance for predicting and learning about empirical phenomena is questionable; insofar as they measure subjective degrees of belief, they are of questionable relevance for giving objective guarantees of

[8] Where P(e) = P(e|H)P(H) + P(e|not-H) P(not-H).

reliable inference. The search for an inductive logic as purely formal rules for relating statements of evidence to hypotheses has largely been abandoned, but, oddly enough the underlying conception of the nature of inductive or statistical inference appears to remain firmly entrenched in the critical rationalist's program.

The most flagrant mistake of the critical rationalists, like the inductivists and probabilists, is to suppose that an 'inconclusive' reason or warrant for an evidence-transcending claim should come in the form of a probability assignment. In fact, inductive uncertainty or inconclusiveness is not well-captured by a posterior probability assignment, in any of the senses that probability has been defined. Even if one is a frequentist about probability, as I am, the inductive job is not accomplished by attempting to assign relative frequencies to hypotheses e.g., as Reichenbach and Salmon often suggested. Even, for example, if hypothesis H has been randomly selected from an urn of hypotheses, p% of which are true, it is completely wrong-headed to suppose that the probability this particular H is true is equal to p. (I call this the fallacy of instantiating probabilities, Mayo 2003 , 2004 and 2005). The severe testing intuition with which we began is at home with a very different use of probability, namely to characterize the probativeness of the testing process itself.[9]

It is time to move on. We can begin to ameliorate the current crisis by (a) rejecting the erroneous conceptions of inductive or 'evidence-transcending' inference upon which their skeptical slide is based, and (b) showing how to develop an account with the resources to define and apply 'severe or reliable error probes'. Here, I can only sketch ingredients of a full severe-testing account developed elsewhere. To avoid confusion with probability statements, among other reasons, throughout Part II, I will replace Musgrave's P for 'proposition' with H for 'hypothesis', with qualifications to be noted.

6.1 The Common Sense Notion of Severe Testing: Isaac

Let us go back to the primitive intuition about severe testing with which we began. Ordinary considerations about testing will do.

---

[9] Probability, in this inferential philosophy, may still be ascribed to outcomes or events, or in formal statistical modeling, to the event that a random variable takes a given value. By testing and severely passing a statistical model, one can then use it to assign probabilities to the possible outcomes.

Consider a student, Isaac. If we are testing how well Isaac has mastered high school material so as to be considered sufficiently ready for work in a four-year college, then a test that covered work from 11th and 12th grade science, history, included mathematical problems (in geometry, algebra, trigonometry, and pre-calculus) required writing a critical essay, and so on, is obviously *more difficult to pass* than one which only required showing minimal proficiency in these subjects at a 6th or 7th grade level: it would be regarded as more searching, more probing, and more severe. The understanding behind this commonplace judgment is roughly this: Achieving a passing or high score is easier and more likely to have come about with the less severe test than the more severe one, *even among students who have not mastered the bulk of high-school material, and hence are not 'college-ready'*. In other words, *before regarding a passing result as genuine evidence for the correctness of a given claim or hypothesis H, it does not suffice to merely survive a test, such survival must be something that is very difficult to achieve if in fact H deviates from what is truly the case*.[10]

By the same token, if the test *is* sufficiently stringent, such that it is practically impossible for students who have not mastered at least p% of high-school material to achieve a score as high as Isaac's, then we regard his passing grade as evidence that he has mastered at least this much. The same reasoning abounds in science and statistics.

Note again the important distinction between highly probed and highly probable even in a so-called frequentist sense: Suppose Isaac had been randomly selected from a wealthy suburb in which, say, 95% of high school students are 'college ready'. Given this high (.95) 'prior' probability to H (i.e., Isaac is college-ready), even a low exam score can result in a fairly high posterior probability to H (Mayo 1997, pp. 326-29; 2004; 2005). Nevertheless, or so our severity intuitions tell us, the high posterior is not good evidence that H has withstood a severe test. In fact, we would wish to ask, What is the probability that

[10] An extremely common fallacy in other notions of severe tests is deliberately avoided in my account. At first blush, a test T that (a) regards a successful prediction as evidence for H, even though (b) a failed prediction would not have counted as disconfirming H, is typically thought 'to be about as blatant a violation of the Popperian commandment as you could commit' (Meehl 1967/1970). But in fact it might be that (a') P(test T passes H;H false) is very low, and yet (b') P(test T fails H; H true) is not low. (a) warrants (a') and (b) warrants (b'): thus there is no violation of the severity requirement. For discussion of this, see Chalmers 1999, Chapter 13 appendix.

the posterior probability would be high even if Isaac is not ready? (i.e., H false). This is an *error probability*, and if it is high, we deny we have good evidence for Isaac's readiness (H).

6.2 The (Error Probabilistic) Severity Principle

We can substantiate, finally, the intuitive severity principle without baiting and switching:

SP: **x** is evidence for H iff, or just to the extent that, **x** constitutes evidence that H has survived a severe test

while demanding, quite unlike the critical rationalist, that a test method be shown to be a reliable error probe.

Test Method M: H is classified as having withstood a severe test T to the extent that H would not have survived (or survived so well), were H false (i.e., a specified flaw in H is present).

Probability may be appealed to here in characterizing the capacity of the error probe: the test that H passed would not be severe if such a passing result is fairly probable, even if H is false.

Test Method M (probabilistic): H is classified as having withstood a severe test T to the extent that H would, *very probably*, not have survived (so well), were H false (i.e., a specified discrepancy from H is present).

Except for formal statistical contexts, 'probability' here may serve merely to pay obeisance to the fact that all empirical claims are strictly fallible, even if a counterexample is never to be actually instantiated in the whole course of human history of the world. Even in technical areas, such as in engineering, it is common to work without a well-specified probability model for catastrophic events, and yet the same requirement about evidence holds. Modifying the above definition for such contexts, the engineer, Yakov Ben-Haim suggests, 'We are subjecting a proposition to a severe test if an erroneous inference concerning the truth of the proposition can result only under

extraordinary circumstances.' (Ben-Haim, 2001, p.214).[11] The kind of inference here might be H: metal buckling of more than a specified amount will not occur under conditions **x**.

7. INDUCTION AS SEVERE TESTING: THROWING OFF THE CRITICAL RATIONALIST'S SHACKLES

In the current view, data **x** provide good evidence for inferring H only if they result from a method which, *taken as a whole*, constitutes H having passed a severe test—that is, a method which would have (at least with very high probability) unearthed any error or flaw in the inference to H. This simple idea, once unpacked thoroughly, lets us shake off the fears and inhibitions that lead critical rationalists to ban ordinary talk of 'justification' and 'induction'. Warranted evidence-transcending inferences—i.e., justified inductive inferences—are to be regarded as cases of inferences from severe testing. A methodology for induction, accordingly, is a methodology for arriving at severe tests, and for scrutinizing inferences by considering the severity with which they have passed tests. Far from wishing to justify the familiar inductive rule from an observed correlation between A and B to an inference that all or most A's are B's (or the next A will be a B), we can see that such a rule would license inferences that had not passed severe tests: it would be a highly unreliable method. An induction following this pattern will be unwarranted, I claim, unless the inference has successfully passed a severe test. Nor, on this account, does H merit any brownie points by dint of being the least poorly tested in a crop of poorly tested hypotheses.

7.1 Taking Seriously the Need to Rule Out Errors Into Which Simple (Enumerative) Induction May Lead

Of course, critical rationalists recognize the errors into which enumerative induction may lead: that is the springboard for their skepticism about induction. Such errors, Musgrave rightly notes, are problems for 'adherents of inductive logic' (p. 346) insofar as an inductive logic is supposed to be formal and context-free. If it is a contingent matter whether given errors are sufficiently well ruled out to infer, say, from correlational data to causal claims, one cannot look to a purely formal

[11] Ben-Haim makes this notion rigorous by means of a definition based on convex sets, which I do not understand sufficiently to explicate.

inductive logic for evidence-transcending inferences. What he, and so many other philosophers of science, fail to see, is that the bankruptcy of the 'logicist' program for inductive logic in no way robs us from having a rich bank account of methods for reaching and warranting inductive inferences! Moreover, showing the bankruptcy of enumerative induction itself depends on being able to substantiate claims about how, in given contexts, blindly following enumerative induction readily leads one astray. In fact, 'the person of common sense', says Musgrave, is fairly savvy in avoiding such familiar foibles.

'People of sense do not argue that the more times your joke has made a person laugh the more likely it is to raise a laugh the next time you tell it.' (presumably to the same person). What warrants this assertion of Musgrave? It is not that he notices his sensible friends do not in fact make such claims, it is rather that Musgrave knows that this is one of those cases where the outcome of trials are *negatively dependent* on previous outcomes. How can Musgrave substantiate this? Is it not because there are well-known errors that would need to be ruled out before supposing the repeatability of an effect (e.g., diminishing returns)? He, like other persons of sense, are perfectly well capable of generalizing about the kinds of cases wherein the more A's that have been B's in the past, the *less likely* the next A will be a B.

Other errors that would need to be ruled out are similarly codified in good statistical reasoning as in conscientious informal critical thinking. But such 'context-dependent' tools appear as much out of reach of the critical rationalist as the inductive logician. Why else would these philosophers of science persist in overlooking the general epistemic justification for such critical tools? Of course, after our severe tester arrives at such reliable rules Musgrave can say, as I suspect he would, that he too would embrace such a method as 'best tested'. Unfortunately, this will only be a matter of *after-the-fact* reconstruction. The critical rationalist denies he is commending *forward-looking* inductive methods that are reliable. What Musgrave and other critical rationalists fail to realize, or fail to capitalize on, is that the basis of *their* criticisms of rudimentary induction rests on having general knowledge of types of situations wherein applying simple enumerative induction would readily lead to erroneous inferences.

It should come as no surprise to critical rationalists, except that they seem not to have heard the news: scientists, like people of sense, have deliberately developed models and methods for (a) checking whether

this kind of temporal dependency holds in a given case, and (b) capitalising on knowledge of dependencies to develop reliable inductive rules for the kind of case at hand. This is the focus of the conglomeration of statistical methods, understood broadly as I do, to include methods of planning, collecting, modelling, and drawing evidence-transcending inferences on the basis of uncertain and limited data. An informal repertoire of day-to-day errors serves an analogous role for the 'person of sense'.

7.2 Severity in Statistical Testing

Statistical tests do not employ our notion of severity directly, but severity can be seen to provide a metastatistical concept and corresponding principles that direct the interpretation and justification of standard statistical methods (e.g., of testing and estimation). For details, see Mayo 1996, Mayo and Spanos 2006. We can encapsulate the severity requirement in statistical testing set-ups thus:

(7.2) Hypothesis H passes a severe test T with **x** if (and only if):

(i) **x** agrees with or 'fits' H (for a suitable notion of fit[12]), and

(ii) test T would (with very high probability) have produced a result that fits H less well than x does, if H were false or incorrect.

8. ERROR STATISTICS AS THE SEVERE TESTER'S THEORY OF INDUCTION

Given the importance with which Musgrave regards the 'wedge' (between reasons to accept and supplying probabilistic justification) for the critical rationalist, it is surprising that he does not take advantage of the distinct philosophical tradition that uses probability not to assign degrees of confirmation or support or belief to hypotheses, but rather to characterise a procedure's reliability in a series of (actual or hypothetical) experiments. Deliberately designed to reach conclusions about statistical hypotheses without invoking prior probabilities in hypotheses, indeed, explicitly denying the

[12] See note 18.

relevance or meaningfulness of posterior probabilities in hypotheses (as opposed to events), probability is used to quantify how *frequently* methods are capable of discriminating between alternative hypotheses and how *reliably* tests facilitate the detection of error. These probabilistic properties of statistical procedures are called *error frequencies* or *error probabilities*.[13] An account based on error probability criteria, whether formal or informal, I dub an *error statistical account* of inference.

8.1 Neyman and Popper: Finessing Induction

In Neyman-Pearson (N-P) testing methods, we see an illuminating example of the 'wedge' Musgrave lauds as the cornerstone of Popper's 'solving' the problem of induction: Neyman and Pearson ground their statistical test rules while, quite deliberately, denying 'inductive' evidence for the truth or probability of statistical hypotheses themselves. The basic rationale underlying N-P statistics was precisely to provide procedures that satisfy aims for rationally adopting an action (whether it be publishing a paper, deciding to believe, or something else) as distinct from supplying grounds for inferring the truth (or probability) of any claim or hypothesis. Neyman referred to such rules for testing as *rules of inductive behaviour* (1952; 1971).

Wishing to draw a stark contrast between this conception of tests and those of Fisher as well as Bayesians (i.e., Jeffreys), Neyman declared that the goal of tests is not to adjust our beliefs but rather to 'adjust our behavior' to limited amounts of data. Erich Lehmann (Neyman's first statistics' student at Berkeley, and eminent statistician in his own right) notes:

> It is remarkable that independently and nearly simultaneously [early 1930's] Neyman and Popper found a revolutionary way to finesse the issue [of the problem of induction] by replacing inductive reasoning with a deductive process of hypothesis testing (Lehmann 1995, p. 32).

Equally striking is that there is scant evidence of direct influences between the two.

---

[13] Embodying a frequentist notion of probability, while denying it is useful to consider the frequency with which hypotheses like H are true (in this or other possible worlds), probability assignments are restricted to random variables (or events) associated with a probabilistic model.

There is, however, one exceedingly important difference between their 'finessing': The N-P tester is required to show that the statistical test procedures actually satisfy the aim of low error probabilities! 'Self-sealing' appeals will not do. Indeed, the central value of tests as rules of behavior is that 'it may often be proved that if we behave according to such a rule...we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false' (Neyman and Pearson 1933, p.142).[14]

In simple statistical significance testing, for example, where hypothesis H might be the familiar 'null hypothesis' (e.g., of no effect or no discrepancy from a fixed parameter value, we obtain a formal exemplification of the following reasoning:

(1) If H were false (i.e., a specified flaw in H present), then (with high probability) the tests would yield evidence of a discordance of at least d (between H and **x**).

(2) There is evidence of a discordance less than *d*.

(3) Therefore, **x** is evidence that the specified flaw in H is absent.

By justifying and showing how to implement premises (1) and (2) of such arguments, one escapes the disappointing limitations of the critical rationalist's game (for a discussion of testing statistical assumptions see Mayo and Spanos 2004). Thus, these error statistical procedures have properties that would have been expected to be embraced by Popperians; and although many speak approvingly of Fisherian tests, e.g., Gillies, they have not utilized these methods to escape the most serious limitations of critical rationalism.[15]

Granted, there is still an important lacuna in the N-P test reasoning: The move from premises (1) and (2) to conclusion (3) is not deductive. Even once the work is done to accept these premises, there is a gap. What is missing is a link from the low long-run error

[14] Neyman regarded '"inductive Behavior" as a Basic concept of Philosophy of Science' to cite the title of a paper of his 1957a. 'Rather than speak of inductive reasoning,' Neyman remarks (1971, p. 1) 'I prefer to speak of inductive behavior'. This refers to the adjustment of our behavior to limited amounts of observation. This is an excellent example of the critical rationalist's "wedge".

[15] The differences between N-P and Fisherian tests, while important, will not concern us here: both are in the error probability tradition as I understand that term Mayo and Cox 2006.

probabilities of (1), to the specific inference in (3). It is true that following the test method is reliable in the sense that one will rarely commit errors in a long-run series of applications. But what does this say regarding the inference at hand? This is where the conception of induction as passing a severe test and the severity principle enter.

8.2 Does the Failure to Reject a Null Hypothesis Confirm It?

If we view induction as severe testing, as I propose, one has the basis for arguing from error probabilistic properties of tests to well-probed claims in the (so-called) 'single case'. In other words, it is not just low error rates in the long run that matter, it is that these may be used to attain probative tests, hence warranted inductions for the claims that withstand them. Things might have been very different if Neyman had not been so wedded to the behavioural-decision model of tests, with its low long-run error justification, once the N-P testing model got off the ground. For it turns out that there is ample evidence of reasoning in accord with our severity principle in little known early papers of Neyman (as well as in works of Pearson).

In one, wherein the striking title of this subsection is found[16], Neyman is addressing his remarks to none other than Carnap. 'In some sections of scientific literature the prevailing attitude is to consider that once a test, deemed to be reliable, fails to reject the hypothesis tested, then this means that the hypothesis is 'confirmed' (Neyman 1955).' Calling this 'a little rash' and 'dangerous', he claims 'a more cautious attitude would be to form one's intuitive opinion only after studying the power function of the test applied.' (p. 41).

If a non-statistically significant result occurred with a test with low power to detect discrepancies of interest, Neyman is saying, then such a non-significant result should not be taken to rule out such departures from the null. Indeed, it is a well known fallacy to go from 'no evidence against' the null hypothesis to 'evidence for' the null, and it instantiates the severity demand.

More generally, if data **x** yield a test result that is not statistically significantly different from H_0 (the null of no effect), and yet the test

[16] It is striking because it contrasts sharply with Neyman's usual disdain for talking of inductive inference, insisting, instead, on his notion of inductive behavior.

has low probability to reject H_0, even when discrepancy δ exists, then **x** is not good evidence for ruling out discrepancy δ.
On the other hand, in statistics as in informal reasoning, if H has managed to survive so probing, searching or *severe* a test, then this is evidence that H is true (or at least that it does not deviate from the truth by more than a given amount). Let us set it out explicitly.

Severity in the Case of Statistically Insignificant Results:

> If data **x** are not statistically significantly different from H_0, and the probability of detecting effect δ is high (low), then **x** constitutes good (poor) evidence that the actual effect is no greater than δ.

8.3 Using Severity in Scrutinizing Non-Significant Results: An Example

A common example is to collect a sample of size *n*, $\mathbf{X} = (X_1, \ldots, X_n)$, where each X_i is an independent and identically distributed Normal variable, $(N(\mu,\sigma^2))$, and run a one-sided test of the hypothesis H_0: $\mu \leq \mu_0$ versus H_1: $\mu > \mu_0$. $\bar{X}$ is the observed sample mean, and a measure of 'fit' or distance is $Z: = (\bar{X}-\mu_0)/\sigma_x$ which is distributed Normally N(0,1)), allowing us to calculate the severity associated with different outcomes and inferences. Letting $\mu_0 = 0$, we have H_0: $\mu = 0$ and H_1: $\mu > 0$. For simplicity let $\sigma_x = 1$. Suppose the test will reject H_0 iff ($\bar{X} = 2$ – a result which would be statistically significant at around the .03 level – and we observe $\bar{X}=1.5$, so H_0 is not rejected. According to the above reasoning we can interpret this result as evidnce not that H_0 is exactly true, but that the discrepancy from 0 is less than δ, provided the test had sufficient high power to have detected a discrepancy this large, So, for example, consider $\delta = 1$. The power to reject H_0 given $\delta = 1$ is only .16, so this does not warrant inferring $\mu < 1$; by contrast, the power against $\mu = 4$ is high, around .97 and thus, following our testing method, the result is good evidence that $\mu < 4$.[17]
These formal error probabilities parallel the informal qualitative assessments that are behnd the plausibility of the idea that evidence

[17] That is, the power against $\mu = 1$ is P(reject H_0; $\mu = 1$) = P(Z > 1) =1−Φ(1) = .16, where Φ is the cumulative distribution function of the standard Normal distribution, and P denotes probability. The power against $\mu = 4$ is P(reject H_0; $\mu = 4$) = P(Z > -2) =1−Φ(−2) = .97.

for H is a matter of H's surviving a severe test.[18] What prevented Popper from uncovering this key, I conjecture, is his failure to take what might be called 'the error probability turn'.[19]

9. THE SEVERITY PRINCIPLE: WHAT WE LEARN FROM ERROR STATISTICS

Whether severity is understood quantitatively or qualitatively, in terms of probability or in terms of non-probabilistic notions, the overarching principle of evidence remains, and may best be expressed as:

Severity Principle: Data **x** (produced by process G) provides a good indication or evidence for hypothesis H (just) to the extent that test T severely passes H with **x**.

By expressing it this way, it is emphasised that H is regarded (or modelled) as a claim about some aspect of the process that generated the data, G. According to the severity principle, when hypothesis H has passed a highly severe test (something that may require several individual tests taken together), we can regard data **x** as evidence for inferring H because it supplies good grounds that we have ruled out the ways it can be a mistake to regard **x** as having been generated by the procedure described by H.

---

[18] This use of power, while reasonable when the outcome just misses rejecting the null, is too coarse, and the severity assessment gets around this. Rather than construe 'a miss as good as a mile', the severity assessment depends on the actual non-statistically significant outcome. That is, we replace the usual calculation of power against μ':

(1) $P(Z > Z_\alpha ; \mu=\mu')$,

with:

(2) $P(Z > Z_p ; \mu=\mu')$, where Z_p is the observed (non-statistically significant) result Z, with corresponding p-value.

(2), quantifies the *severity* with which the test passes $\mu< \mu'$.

To illustrate, compare observing (a) $\bar{X} = 1.5$ and observing (b) $\bar{X} = .1$. Both outcomes fail to reject H_0 with our test, but intuitively we would like to reflect the fact that the latter is so much close to 0. While we saw that the power against $\mu = 1$ is low, and power does not change with the actual outcome, severity does. The severity associated with $\mu < 1$ in case (a) is $P(Z > 1.5 - 1;) = .3$ whereas in case (b) it is $P(Z > .1 - 1) = P(Z > -.9) = .8$ (all numbers are approximate here.) So in case (b), unlike case (a), the inference $\mu < 1$ is warranted with fairly high severity, .8. Note that in the case of rejecting H, high power corresponds to low severity whereas with accepting H it is the reverse. However, whatever form H takes, we can talk of the severity of a test to have uncovered that H is in error. For a detailed discussion see Mayo and Spanos 2006.

[19] In private communication Popper explained that he regretted never having had the chance to learn statistical methodology.

9.1 Dangerous Misunderstandings

Although a full understanding of how to calculate severity demands careful discussion beyond this paper, the central points I need to make require avoiding some common misunderstandings on which criticisms often rest.

9.1.1 A Severity Assessment is always Relative to the Hypothesis that 'Passes'

It is common to talk as if a severity assessment attaches to the test itself—as Popper does—but doing so leads to untoward results. One cannot answer the question: 'How severe is test T?' without including the particular hypothesis that is claimed to have passed, or about which one wishes to make an inference. The great advantage of relativising the assessment to the particular inference (and the particular data set) is that high severity is always what is wanted for evidence.[20] No problem occurs unless one forgets that a given test may severely pass one hypothesis and not another, even among the hypotheses under consideration. This confusion most readily takes the form of what might be called: 'The Criticism From Overly Sensitive Tests'.

Severity cannot be a sensible desiderata, so the criticism goes, because a test may be made so severe that even a trivially small departure from a hypothesis H will result in inferring H´– where H´is a rival to H, or an assertion about some anomaly or error in H. What this criticism overlooks is that the inference whose severity we would need to consider in that case is H´; but having put H to a stringent test is not to have stringently probed H´! The misunderstanding behind the criticism boils down to thinking that H´ has passed a severe test, as I am defining it, but in fact it is quite the opposite.

Consider our test for deficiencies in Isaac's college readiness and the hypothesis: H: Isaac is college-ready, as against, H´: Isaac is not college-ready. We can make the tests so hard, and the hurdle for regarding grades as evidence for H so high, that his scores are practically always going to lead to denying H and inferring H´ (he is deficient). However, H´ has passed a test with *very low* severity because it would very often lead to inferring H´, even if H´ is false

[20] This contrasts with the use of Type I and Type II error probabilities in Neyman-Pearson tests.

and actually H is true. How to arrive at assessments of the largest discrepancy warranted by the test is formalised in a 'rule for rejection' in a severity interpretation of statistical tests (Mayo 1996).

*9.1.2 Severity Condition (ii) Differs from Saying that **x** is Very Improbable Given Not-H*

In contrast to Popper's attempted definition of severity, as well as others (e.g., likelihoodists) the second severity condition, i.e., condition (ii) is not merely to assert that P(**x**; H is false) is low,[21] where 'P(**x**; H is false)' is to be read: 'the probability of **x** under the assumption that H is false'. This is called the likelihood of H given **x**. A familiar example shows why. H_1 might be that a coin is fair, and **x** the result of n flips. For any **x** one can construct a hypothesis H_2 that makes the data maximally likely, e.g., H_2 can assert that the probability of heads is 1 just on those tosses that yield heads, 0 otherwise. P(**x**;H_1) is very low and P(**x**;H_2) is high, however, H_2 has not passed a severe test because one can always construct some such maximally likely hypothesis *or other* to perfectly fit the data on coin tosses, even though it is false and the coin is perfectly fair (i.e., H_1 is true).[22] The test that H_2 passes has minimal severity. (This is a case of what I call 'gellerization'.)

In other words, what principally distinguishes the error probability account of tests is that whatever 'fit' measure is satisfied in showing H 'withstands' the test, the error statistician requires asking a question that is one level removed, as it were: How frequently would H withstand the test so well, even if H is false? (Mayo 1996, Mayo and Kruse 2001).

---

[21] The requirement of 'fit' in the severity definition, clause (i), may be defined as a requirement about likelihoods, in particular, it requires that P(x; H) be higher than P(x; H is false). It is important to see that this differs from a conditional probability; there is no assumption that a prior probability assignment to H exists or is meaningful. See following note.

[22] I am using ';' in writing P(x;H) in contrast to the notation typically used for a conditional probability P(x/H) in order to emphasize that severity does *not* use a conditional probability which, strictly speaking, requires the prior probabilities P(H_i) be well-defined, for an exhaustive set of hypotheses.

9.1.3 The Degree of Severity with which a Test Passes H is Not the Degree of Probability of H

Finding that a hypothesis H severely passes test T with data **x** does not license a posterior probability assignment to H, a notion which depends on having prior probability assignments to an exhaustive set of hypotheses. As already noted in Section 6, 'highly probed is not the same as highly probable'; but it bears repeating, given how flagrant is this misinterpretation. Such Bayesian calculations (from whatever school of Bayesianism one chooses) are at odds with the severity principle: high posterior probability is neither necessary nor sufficient for high severity, in any sense of probability.

9.2 What Statistical Testing Teaches Us About Severe Testing in General

9.2.1 The Need for Methods to Test the Reliability of Data Statements.

Thinking of formal error statistical testing alerts us at once that it is impossible to assess reliability or severity with just statements of data and hypotheses divorced from the experimental context in which they were generated, modeled, and selected for testing. For the critical rationalist, this recognition spells nothing but trouble. It is assumed, but not explained, how we justify the data on which the critical rationalists' claims of 'best tested H' depends, save perhaps when **x** is the most rudimentary kind of perception. That General Relativity, GTR, one of Popper's favorite examples, is best tested, for example, depends on already having an account for inductively inferring highly sophisticated hypotheses. Even assuming we knew which of many rivals 'pass' the most tests, we are offered no tools for adjudicating disagreements about what passing results actually show about the phenomenon in question. Such accounts remain irrelevant both for science and for philosophy.

9.2.2 Beyond the 'Tower Image' of Data.

Philosophers, critical rationalists included, seem stuck in what might be called the 'tower image' of empirical data: evidence claims are only as reliable as are the intermediary inferences used in arriving at or inferring them. The minute that intermediary inferences are admitted to have their own assumptions, there is a knee-jerk reaction to concede an unacceptable 'regress' without bothering to question

whether this need be so. Were it so, then piling inference upon inference would leave us with increasingly less reliable results—but the opposite is true. Individual measurements, for example, may each have wide margins of error, whereas an inferred estimate (e.g., through averaging) may be highly reliable. What makes such inductive inferences about evidence work is that properly modelled and cleverly used, data can lead from less to more accurate and reliable claims: through interconnected checks of error and robustness results, garbage in need not be garbage out! This day-to-day truism of the 'person of common sense' seems overlooked by the critical rationalist!

9.2.3 Need to Partition: 'H is False' is Not the so-called Catchall Factor

The catchall factor is the disjunction of hypotheses other than H, including those not yet even thought of. 'H is false' in the definition of severity refers, instead, to a specific error that hypothesis H may be seen to be denying. Since the error probability assignments needed for formal statistical cases require hypotheses that exhaust the space of alternatives, the methods, as well as the motives, for splitting off questions and *partitioning* spaces of answers that are found in statistics are highly instructive for the broader aims of an error statistical philosophy.

This leads to an important criticism or challenge raised especially by philosophers in the Popperian tradition, whether or not they call themselves critical rationalists; namely, how does the error statistical account severely pass high level theories? (see Chalmers, Earman, Laudan, Mayo 2002a,b and forthcoming). A quick sketch must suffice.

10. SEVERE TESTING IN PROBING LARGE SCALE THEORIES

What enables this account of severity to work is that the hypothesis H under test by means of data **x** is designed to be a specific and local claim, e.g., about parameter values, about causes, about the reliability of an effect, or about experimental assumptions. 'H is false' is not a disjunction of all possible rival explanations of **x**, which would include those not yet known as just noted. This is true, even if H is part of some large scale theory T: the condition 'given H is false'

always means 'given H is false with respect to what it says about *this particular* effect or phenomenon'. We can abbreviate this claim as $_T$(H) to indicate H is a piece of $_T$. If a hypothesis $_T$(H) passes a severe test we can infer something positive: that the theory $_T$ gets it right about the specific claim H that severely passes.

The price of this localisation is that one is not entitled to regard global or large-scale theories as having passed severe tests so long as they contain hypotheses and predictions that have not been well-probed. If scientific progress is viewed as turning on appraising high-level theories, then this type of localized account of testing will be regarded as guilty of a serious omission, unless it is supplemented with an account of theory appraisal.

> [Her] argument for scientific laws and theories boils down to the claim that they have withstood severe tests better than any available competitor. The only difference between [her] and the Popperians is that she has a superior version of what counts as a severe test. (Chalmers, 1999, p. 208)

The truth is that I never intended to provide any kind of 'argument for scientific laws and theories'; and permitting the comparativist account that he, like Popper and others champion, would conflict with the aims of severity, and preclude the very features Chalmers endorses as superior to Popper's.

Whereas we *can* give guarantees about the reliability of the piecemeal experimental test, it is highly unreliable to follow the comparativist's method: If large-scale theory $_T$ is 'best-tested', regard all of $_T$ as having withstood severe testing, and thus accept or believe $_T$. Still, Chalmers, like Laudan (1997), suggests that in the case of high-level theory I should define severity comparatively:

> *The Comparativist's Suggested Definition*: A theory has been severely tested provided it has survived (severe?) tests its known rivals have failed to pass (and not vice versa).

(The question mark here is due to the fact that in Laudan's account, at least, it is not required that these lower level tests themselves be severe – or, at any rate, he does not make this point clear.) Laudan calls this the 'comparativist rescue' for my account; but no such rescue mission is required or desired. In my view, it is disingenuous to say that all of a theory has survived a good test when there are ways it can be wrong that have not been probed, that there are regions of

implication not checked at all. To embrace the comparativist account is to be thrown back to the critical rationalists' problem: being unable to say what is so good about the theory that (by historical accident) happens to be best-tested so far.

10.1 Learning from Tests that Fail to be Classified as Severe

Comparativist testing accounts, eager as they are to license the entire theory, ignore what for our severe tester is the central engine for making progress, for getting ideas for fruitful things to do next, to learn more; namely, by asking, how could we be wrong in supposing all of theory T has severely passed? Why are we *not* allowed to say that the entire theory is severely probed as a whole? —in all the arenas in which the effects in question may occur. Even without having alternatives we can ask *how could it be a mistake to regard the existing evidence as good evidence for all of the theory?*
Although we learn a lot about phenomena from hypotheses that pass or fail severe tests, we learn at least as much from finding that our test is inadequate as a severe error probe. I know of no account of testing that recognizes this explicitly, and yet it falls out immediately from the error statistical account. One way to unearth errors in taking a passing result as evidence for a given theory T is to construct a suitable alternative T* often using the known data **x** to ensure T* accords with **x** in the respects already tested. This may serve an important role in showing why T fails to have passed severly as a whole. This does not mean that T* has passed just as severely as T has (even with respect to the aspects probed). But it is instructive in finding out that aspects of T we might have thought were well probed by test T, were in fact not well probed. This is a crucial tool in discovering and constructing new theories. (For a discussion of how this strategy figured centrally in developing alternatives to GTR, see Mayo 2002a and forthcoming).

10.2 Reliability and Stability through Large Scale Theory Change

In addition to reliability, this account has another feature that is missing from the comparativist tester: stability and cumulativity. The severity for passing a lower level hypothesis remains the same even

through changing interpretations and through changing high level theories in which it might be embedded. Suppose, for example, that a hypothesis about parameter value *m* has passed severely. This severity evaluation is not altered by the existence of another theory that agrees with this hypothesized parameter value. More generally, severely passing what theory $_T$ says about H, $_T$(H), gives us knowledge about this aspect of theory $_T$, and this assessment remains even though the theory undergoes repeated improvements, revisions, and even reconceptions. By contrast, as soon as an alternative theory comes to light that does as well as theory $_T$ does on existing tests, the comparativist would be forced to change the assessment of how well $_T$ has been tested (it would have to 'give back the crown' as it were.) The error statistical tester is not precluded from talking of 'accepting' the theory, understood as accepting as severely passed some of its key hypotheses, or simply, regarding it as a fruitful basis from which to learn more and probe further the phenomenon of interest. It would be correct to regard it as a fruitful basis for learning if it allows us to say, even without having a clue about the correct large-scale theory, that any theory in the domain in question would have to include the severely tested effects.

11. CONCLUDING COMMENTS

The lack of progress in the neo-Popperian philosophy, I have argued in Part I of this paper, may be traced to its inability to characterize the severity principle (SP) underlying the epigraph of this paper:

> Observations or experiments can be accepted as supporting a theory (or a hypothesis, or a scientific assertion) only if these observations or experiments are severe tests of the theory—or in other words, only if they result from serious attempts to refute the theory". (Popper, 1994, p. 89)

Critical rationalists have failed to actually cash out what 'surviving serious criticism' demands, and why H's surviving the 'ordeal' makes it reasonable to accept H, regard H as supported, or as beliefworthy. The severity principle is at the heart of rationality in science – Popper was right about that – so long as 'H's surviving serious criticism' may be taken to mean that H has been put to a scrutiny that would have (or would very probably have) uncovered the falsity of (or errors in) H, and yet H emerged unscathed. That is to say, the epistemological force behind SP holds just to the extent that H has survived a highly reliable

probe of the ways in which H might be false. However, Popper's logical account deprived him of the resources to articulate 'reliable error probes'; and surprisingly, his current day followers have yet to remove their logical empiricist blinders.

In Musgrave's "promissory notes" (p. 323), he calls for "an account of which kinds of criticism are serious criticisms and which not." Building such "a theory of criticism" is welcome but it demands empirical not purely logical assessments of the error-probing capacities of tools. The comparativist principle CR that he endorses, moreover, is unreliable —claims can easily (frequently) be 'best surviving' with **x** (at time t), even though **x** provides little or no reassurance that errors have been ruled out or even probed.

In part II of this paper, I discussed an account of testing that captures the spirit behind Popper's intuitions about severity, while enabling it to be made operational. Here, probability is used to quantify how frequently methods are capable of discriminating between alternative hypotheses and how reliably tests facilitate the detection of error. These probabilistic properties of test procedures are called error probabilities, and an account based on error probability criteria, whether formal or informal, I dub an *error statistical account of inference*. The error statistical tester agrees with Musgrave's critical rationalist in rejecting the probabilists's view of justifying claims, while being able to provide tests that are genuinely reliable error probes. On the critical rationalist's own criteria of appraisal, therefore, the error statistical approach is to be preferred to the method CR: CR fails to withstand critical scrutiny.

I have sketched how the quantitative conception of severity arising in error statistics may be carried over into more qualitative arenas and in learning about high level theories. Although much work remains in developing such qualitative severe tests, it is a research program with the properties that have much to offer the critical rationalist. If Musgrave and other critical rationalists are serious about pushing forward the stalled Popperian research program beyond the "twelve or twenty" adherents, it is to be hoped that they will at least consider the avenue for progress offered by developing the methodology of error statistics and severe testing. Not only would this help salvage the brilliant gems in Popper, it would be relevant for foundational debates

in statistics as to what is really required for rational and objective scientific inquiry.

REFERENCES

Achinstein, P. (2001) *The Book of Evidence*, Oxford University Press.

Ben Haim, Y. (2001) *Information-Gap Decision Theory: Decisions Under Severe Uncertainty*, Academic Press: San Diego CA

Chalmers, A. F. (1999) *What Is This Thing Called Science?* 3rd ed., University of Queensland Press, Australia.

Earman, J. (1992) *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, MIT Press, Cambridge, MA.

Gillies, D. A. (1973) *An Objective Theory of Probability*, Methuem, London.

Grünbaum, A. (1978) 'Popper vs. Inductivism', in Radnitzky and Andersson (eds.), *Progress and Rationality in Science*, Boston Studies in the Philosophy of Science, vol. 58. Dordrecht, The Netherlands: Reidel, pp. 117-142.

Laudan, L. (1997) 'How about Bust? Factoring Explanatory Power Back into Theory Evaluation', *Philosophy of Science* 64(2): 306-16.

Lehmann, E. L. (1995) 'Neyman's Statistical Philosophy', *Probability and Mathematical Statistics*, 15: 29-36.

Mayo, D. G (1991) 'Novel Evidence and Severe Tests,' *Philosophy of Science* 56:*523-552*.

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

Mayo, D. G. (1997a) 'Duhem's Problem, The Bayesian Way, and Error Statistics, or 'What's Belief Got To Do With It'?' and 'Response to Howson and Laudan,' *Philosophy of Science*, 64(2): 222–24 and 323–33.

Mayo, D. G (1997b) 'Error Statistics and Learning from Error: Making a Virtue of Necessity,' *Philosophyof Science* 64 (Proceedings) §195-212

Mayo, D. G. (2002a) 'Theory Testing, Statistical Methodology, and the Growth of Experimental Knowledge,' *Proceedings of the International Congress for Logic, Methodology, and Philosophy of Science*, Kluwer Press, pp. 171-190.

Mayo, D. G. (2002b) 'Severe Testing as a Guide for Inductive Learning,' in Kyburg, H. E. and M. Thalos, eds., *Probability is the Very Guide of Life*, Chicago: Open Court, pp. 89-117.

Mayo, D. G. (2003) 'Could Fisher, Jeffreys and Neyman Have Agreed? Commentary on J. Berger's Fisher Address,' *Statistical Science* 18: 19–24.

Mayo, D. G. (2004) 'An Error-Statistical Philosophy of Evidence', in M. Taper and S. Lele (eds.), *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Consideration*, Chicago: University of Chicago Press

Mayo, D. G. (2005) 'Evidence as Passing Severe Tests: Highly Probed vs. Highly Proved' in P. Achinstein (ed.) *Scientific Evidence*, Johns Hopkins University Press: 95-127.

Mayo, D. G. (2006) 'The Philosophy of Statistics', in S. Sarkar (ed.) *Routledge Encyclopedia of the Philosophy of Science*.

Mayo, D. G. (forthcoming) 'Using Low-Level Tests to Probe High-Level Theories?: Severity Without Comparativism'

Mayo, D. G. and D. Cox (2006) 'Frequentist Statistics as a Theory of Inductive Inference,' with D.R. Cox, *Proceedings of The Second Erich L. Lehmann Symposium*, Vol xx, Institute of Mathematical Statistics (IMS) Lecture Notes-Monograph Series.

Mayo, D. G. and M. Kruse (2001) 'Principles of Inference and their Consequences,' pp. 381–403 in D. Cornfield and J. Williamson (eds.), *Foundations of Bayesianism*, Kluwer Academic Publishers, Netherlands.

Mayo, D. G. and A. Spanos (2004) 'Methodology in Practice: Statistical Misspecification Testing', *Philosophy of Science* 71: 1007-1025.

Mayo D. G. and A. Spanos (2006) 'Severe Testing as a Basic Concept in the Neyman-Pearson Philosophy of Induction,' *British Journal of Philosophy of Science*.

Meehl, P. E. (1967/1970) 'Theory-Testing in Psychology and Physics: A Methodological Paradox,' In D. E. Morrison and R. E. Henkel (eds.), *The Significance Test Controversy* (1970), Aldine, Chicago.

Musgrave, A. (1974a) 'Logical versus historical theories of confirmation,' *British Journal for the Philosophy of Science* 25:1-23.
Musgrave, A. (1974b) 'The Objectivism of Popper's Epistemology', in P. A. Schilpp (ed.) *The Library of Living Philosophers*, LaSalle: Open court, pp. 560-596
Musgrave, A. (1989) 'Deductive Heuristics', in K. Gavroglu, Y. Goudaroulis and P. Nicolacopoulos (eds.) *Imre Lakatos and Theories of Scientific Change*, Dordrecht' Kluwer Academic, pp. 15-32
Musgrave, A. (1999) *Essays in Realism and Rationalism*, (Chapter 16) Amersterdam: Rodopi; Atlanta, GA.
Neyman, J. (1952) *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd ed. U.S. Department of Agriculture, Washington.
Neyman, J. (1955) 'The Problem of Inductive Inference,' *Communications on Pure and Applied Mathematics*, VIII, 13-46.
Neyman, J. (1957a) 'Inductive Behavior as a Basic Concept of Philosophy of Science,' *Revue Inst. Int. De Stat.*, 25: 7-22.
Neyman, J. (1957b) 'The Use of the Concept of Power in Agricultural Experimentation,' *Journal of the Indian Society of Agricultural Statistics*, IX: 9-17.
Neyman, J. (1971) 'Foundations of Behavioristic Statistics,' in V. P. Godambe and D. A. Sprott (eds.) *Foundations of Statistical Inference*, Toronto: Holt, Rinehart & Winston, pp.1-13 (comments and reply, pp. 14-19).
Neyman, J. and E. S. Pearson (1933) 'On the problem of the most efficient tests of statistical hypotheses', *Philosophical Transactions of the Royal Society*, A, 231: 289-337. Reprinted in Neyman, J. and E. S. Pearson (1966).
Popper, K. (1959) *The Logic of Scientific Discovery*, New York: Basic Books
Popper, K. (1962) *Conjectures and Refutations: The Growth of Scientific Knowledge*, New York: Basic Books.
Popper, K. (1976) 'A Note on Verisimilitude', The British Journal for the Philosophy of Science 27: 124-159.
Popper, K. (1994) *The Myth of the Framework: In Defence of Science and Rationality* (edited by N.A. Notturno). London: Routledge.