Can Scientific Theories Be Warranted with Severity? Exchanges with Alan Chalmers

Deborah G. Mayo

Reacting to Alan Chalmers's most serious challenges to the account of theories I have put forward provides an excellent springboard for dealing with general questions of how to characterize and justify inferences beyond the data (ampliative or inductive inference) and how to view the role of theory appraisal in characterizing scientific progress:

1. *Experimental Reasoning and Reliability*: Can generalizations and theoretical claims ever be warranted with severity? What is an argument from coincidence? Do experimental data so underdetermine general claims that warranted inferences are limited only to the specific confines in which the data have been collected?
2. *Objectivity and Rationality*: Must scientific progress and rationality be framed in terms of large-scale theory change? Does a negative answer entail a heuristic role for theories?
3. *Metaphilosophical Themes*: How do philosophical assumptions influence the interpretation of historical cases?

Although Chalmers appreciates that "the new experimentalism has brought philosophy of science down to earth in a valuable way" (1999), his decade-long call for us to supplement "the life of experiment" with a "life of theory," he argues, remains inadequately answered. In Section 9 ("Progress?"), however, I note how we may now be moving toward agreement on the central point at issue in this essay.

1 Argument from Coincidence

Chalmers denies that the account of the roles and appraisals of theories that emerges in the error-statistical account is adequate, because he thinks that

requiring a theory to be severely tested (i.e., to pass severely) in my sense is too severe. According to Chalmers, scientists must invariably "accept" theories T as warranted "to the extent that they are borne out by a wide variety of otherwise unconnected phenomena" even though such "arguments from coincidence" fail to warrant T with severity (this volume, pp. 67–8). This acceptance, moreover, must be understood as accepting T as fully true; it must not be regarded as taking any of the weaker stances toward T that I consider, such as regarding T as a solution to key experimental problems, or employing T in pursuit of "learning" or "understanding," or in discovering rivals. Such stances, Chalmers thinks, regard theories as mere heuristics to be tossed away after being used to attain experimental knowledge.

Although all of T is not warranted with severity, all of T may be warranted by dint of an "argument from coincidence," wherein T correctly fits or predicts a variety of experimental results, for example, $E_1, E_2, \ldots, E_n$. Would it not be an amazing coincidence if theory T got it right, and in detail, about such a wide variety of phenomena, if T were false? We are not told how to determine when the agreement is appropriately amazing were T false, except to insist on a version of the novelty requirement: "The sense in which a theory is borne out by the phenomena needs to be demanding. It is no coincidence that a theory fits the phenomena if the details of the theory are adjusted in the light of the phenomena to bring about the fit" (p. 68). It is hard to see how this argument differs from an inference to a severely tested theory in my sense; if it is really an amazing coincidence that theory T has passed diverse tests $E_1, E_2, \ldots, E_n$, "if it were false," then we have T passing with severity. If Chalmers's "demanding" requirement is to be truly demanding, it would require the passing results to be very difficult to achieve if T were false. Oddly, Chalmers insists that an argument from coincidence can never be sufficiently strong to warrant an inference to a theory, or even to a generalization beyond the evidence!

2 Unreliable Methods: Rigged Hypotheses and Arguments from Conspiracy

According to Chalmers, "If scientific claims are warranted only by dint of surviving severe tests then theories cannot be warranted" (this volume, p. 60). That is because "the content of theories goes beyond the empirical evidence in their support and, in doing so, opens up various ways in which theories could be wrong" (p. 61). But high severity never demanded infallibility, and to charge as Chalmers does that I seek error-free knowledge is

inconsistent with the centrality of statistical arguments in my account. But imagining such a charge to be a slip, we can consider his argument for denying that a general hypothesis H can ever pass severely on the basis of passing results $E_1, E_2, \ldots, E_n$. According to Chalmers, if we ask of a theory that has passed some experimental test "in what ways could it have passed this test if it were false," then one can answer, "if the theory got it right about the experiment in question but wrong in other contexts, or if some alternative theory, which says the same about the experimental claim in question, is true" (p. 61). Fortunately one's freedom to say this scarcely precludes arguing that it is very improbable for such passing results to occur, were H false.

But it is important to see that even without being able to argue in a particular case that H has passed severely, our error-statistical tester most certainly can condemn Chalmers's general underdetermination gambit: Although one can always claim that the passing results hold only for the observed cases, or that all the results are as if H is true but in fact it is false, such a ruse would be highly unreliable. No true hypothesis could be corroborated! One can always argue that any hypothesis H, however well probed, is actually false and some unknown (and unnamed) rival is responsible for our repeated ability to generate results that H passes!

Note, too, that this would also preclude Chalmers's arguments from coincidence: One can always deny that an agreement of results in the wide variety of cases suggested by Chalmers would be a coincidence were H false; thus, Chalmers's own argument against severity would preclude the argument from coincidence he wishes to uphold.

Chalmers's gambit comes under the heading of a "rigged alternative" (Mayo, 1996, p. 205):

> *Rigged hypothesis R*: A (primary) alternative to H that, by definition, would be found to agree with any experimental evidence taken to pass H.

The problem is that even where H had repeatedly passed highly severe probes into the ways H could err, this general procedure would always sanction the argument that all existing experiments were affected in such a way as to systematically mask the falsity of H. Thus, such an argument has high if not maximal probability of erroneously failing to discern the correctness of H, even where H is true. Whenever it can be shown that such a stratagem is being used, it is discounted by the error statistician.

Thus it is unclear why Chalmers maintains that "it is not merely high-level theories such as Newton's that cannot be severely tested, Mayo style. Low-level experimental laws cannot be severely tested either" (p. 61). Aside from

the argument from "rigging," Chalmers points to "one of Mayo's own examples":

> Mayo rightly points out that, because a range of theories besides Einstein's predicts Einstein's law of gravity, testing the law of gravity by the eclipse experiments does not serve as a test of Einstein's theory as opposed to those alternatives. (Chalmers, this volume, p. 61)

True. However, the fact that Eddington's experiments failed to pass all of GTR with severity scarcely shows that no theory passes with severity. We have belied such skepticism, both in qualitative examples and in formal statistical ones. Having inferred a deflection effect in radio astronomy, experimental relativists can argue with severity that the deflection effect is approximately $L \pm \varepsilon$. Why? Because if this inference were false, then with high probability they would not have continually been able to reliably produce the results they did. Analogously, having studied the effects on blood vessels of high-level exposure to radioactive materials in well-designed studies, we argue with severity that there is evidence of such radioactive effects to any human similarly exposed. Although the inductive claim depends upon future experiments contingent on certain experimental conditions holding approximately, the fact that we can check those conditions enables the inference to pass. We could thereby discredit any attempt to dismiss the relevance of those studies for future cases as engaging in a highly unreliable, and hence unscientific, method of inference.

To hold as Chalmers does that "if we are to extract from science only those claims that have survived Mayo's version of a severe test, then we will be left with only some very low-level, and extremely qualified, statements" (this volume, p. 62) about H being consistent with past results. If we accept this position, it would be an extraordinary mystery that we have worked so hard at experimental design, and at developing procedures to avoid errors in generalizing beyond observed data. Why bother with elaborate experimental controls, randomized treatment control studies, and so on, if all one is ever able to do is announce the data already observed!

According to Chalmers, "the point that theories cannot be severely tested... is one that Mayo herself accepts. It is precisely because this is so that she herself has seen it necessary to take up the question of the role of theory in science" (p. 62).

I take up the question of the role of theory because I find that the error-statistical philosophy of science transforms the entire task of constructing an adequate philosophical account of the roles of high-level theories in science. The main point of my contribution was to show how progress is

made in coming to learn more about theories by deliberately exploiting the knowledge of errors not yet ruled out and by building on aspects that are severely probed. It is incorrect to suppose that I deny theories can pass severely simply because I deny that passing the comparativist test suffices.

Chalmers finds my position on theories deeply problematic, and I shall spend the remainder of this exchange examining why. Note that even granting my arguments in Section 1 – that our experimentalist does not deny that theories pass severely – we are left with the argument of Chalmers I want to consider now. For even if he were to allow that some theories and hypotheses may be severely warranted, here Chalmers is keen to argue that scientists require a weaker notion wherein inseverely warranted theories may be accepted. So to engage Chalmers's position, we need to consider an argument from coincidence to theory T that is insevere; that is, we need to consider cases where evidence licenses an argument from coincidence to T, but T fails to be severely passed by dint of this evidence. Such an argument only warrants, with severity, some weaker variant of the full theory T, or proper subsets of T. The most familiar and yet the weakest form warrants merely an inference to a "real" or nonchance effect among the phenomena T seeks to explain. Allowing the move from severely passing a subset of T to all of T is easy enough; the question is why one would think it a good idea, much less required, as Chalmers does, to permit such an inference. I focus on what Chalmers deems his strongest arguments: those that he takes to preclude my getting away with claiming that only the severely corroborated portions or variants of a theory are warranted. The arguments, however, appear to land us immediately in contradictions and also conflict with the historical episodes on which he draws.

3 The Argument from Needing to Avoid Unexplained Coincidences

Chalmers argues that if we do not allow insevere arguments of coincidence to warrant all of theory T_1, then we will be stuck with unexplained coincidences when T_1 is replaced by theory T_2. He claims this would violate what he calls the "general correspondence principle," wherein "successful theories – theories that have been borne out by arguments from coincidence – must live on as limiting cases of their successors" (this volume, p. 68).

But we see at once that Chalmers's argument, to be consistent, must be construed as claiming that the argument from coincidence warrants not all of T_1 (as he supposes), but at most the severely corroborated subportions or variants of T_1.

For example, suppose that all of T_1 is warranted when T_1 is replaced by an incompatible theory T_2. Then the scientist would be accepting incompatible theories. Theory T_1 is replaced by T_2 when T_1 is determined to give an erroneous explanation of the results. If any portions of T_1 are retained, they could not comprise all of T_1, understood to hold in all the domains circumscribed by T_1, but at most those aspects that were and remain well corroborated. Nor are we then without an explanation of the passing results, because the replacement theory T_2 explains them. Therefore, Chalmers's appeal to the correspondence principle works against his position and he is faced with this dilemma: He must either deny that all of T_1 is retained when replaced or deny that accepting T_1 by an insevere argument from coincidence can really mean to accept all of T_1. Several of his own remarks suggest he adopts the former horn: "Replaced theories continue to be reliably applicable in the domains in which they have been borne out by powerful arguments from coincidence" (this volume, p. 69). But that would mean that the domains in which they are borne out by his argument from coincidence cannot be the full domain of the theory but only a truncated domain determined after the full theory is replaced.

To sum up this part, Chalmers is at pains to show that T_1 must be warranted in all of its domains, not just in those where it has passed severely. Yet he is implicitly forced to shift to the claim that what is borne out by powerful arguments from coincidence is T_1 *restricted to the domains in which it continues to be reliably applicable*, or some such T_1 variant. Moreover, when we add the falsifying case E_{n+1} to $E_1, E_2, \ldots, E_n$ that lead to T_1 being replaced by T_2, the full set of data no longer offers an argument from coincidence, because that would require accordant results for T_1. It would not be an amazing coincidence to observe $E_1, E_2, \ldots, E_n$ as well as E_{n+1}, were T_1 false (and T_2 true) – insofar as T_2 accords with the available data. His favorite example is retaining Newton's theory when replaced by Einstein's, but what would be retained would be a variant wherein relativistic effects (that falsify Newton) are imperceptible. Retaining the nonfalsified variants of a theory is scarcely to retain the full theory.

The only coherent position is that what remains as warranted by the data, if anything, are the portions severely corroborated, as we recommend. Granted, at any point in time one might not know which portions these are, and I concede Chalmers's point that we may later need to reinterpret or correct our understanding of what we are entitled to infer; that is why one deliberately sets out to explore implications on which theories disagree. But to suppose that hypothetically assuming a theory T for purposes of

drawing out T's implications requires accepting T as true, leads to inconsistencies. This takes us to Chalmers's next argument.

4 The Argument from the Need to Explore New Domains

Chalmers's second argument is that scientists must assume T holds in domains other than those in which it has been probed with severity when they set out to explore new domains and phenomena – were I to deny this, my account "would render impossible the development of those theories to the stage where their limitations could be appreciated" (this volume, p. 67). For example, he claims, following Clifford Will, that experimental relativists assume GTR in probing areas beyond those for which it has been severely corroborated at any point in time. But do they? Such explorations fall into one of two types:

1. A theory is merely assumed hypothetically to derive predictions to find new data to discriminate between rivals that are thus far corroborated, or
2. Only aspects of T that have passed severely, or on which all viable rivals agree, are relied on to learn more about a phenomenon of interest.

The case of experimental relativity does provide examples of each of these, but in neither case is a theory regarded as warranted beyond the domains it has passed severely. Moreover, we seem forced to conclude once again that Chalmers's own arguments become incoherent if taken to warrant all of T.

5 Aims of Science: A Tension between Severity and Informativeness?

We may agree that a "tension" exists between severely passed and highly informative theories, but it is important to see that the aim of science for our error statistician is not severity but rather finding things out. One finds things out by means of experimental inquiries, which we define as inquiries where error probabilities or severity may be assessed, whether quantitatively or qualitatively. By distinguishing those hypotheses that have passed severely from those that have not, one gets ideas as to new hypotheses to test, and how to test them. It is the *learning goal* that drives us to consider implications where thus far well-corroborated theories and hypotheses may fail. Chalmers disparages such goals as merely heuristic uses of theories, but this seems to forfeit what is crucial to the most dynamic parts of the life of theory (see Glymour, Chapter 8, this volume).

5.1 The Contrast between Us

To make the contrast between us clear, for the error statistician, theories and hypotheses may be "used" in an inquiry so long as either (1) they have themselves been warranted with severity or, if not, (2) any inferences remain robust in the face of their falsity. Commonly, under condition 1, what is used is not the truth of theory T but a hypothesis of *how far from true T may be* with regard to some parameter (having been determined, say, as an upper or lower bound of a confidence interval or severity assessment). By contrast, Chalmers maintains that inseverely warranted aspects of T must be assumed in exploring T; thus, unless robustness could be shown separately, it would prevent warranted inferences about T – at odds with our account. If the validity of an inquiry depends on the truth or correctness of the theory being investigated, then the inquiry would be circular! By failing to consider condition 2, Chalmers views "using T" as assuming the truth of T. An inference to T where T was already assumed would earn minimal severity, and thus supply poor evidence for T. So if Chalmers's description of scientific practice were correct, it would invalidate our account of evidence as capturing what goes on in science. But if we are right to deny that scientists must circularly accept the theory they are appraising – the kind of position the Kuhnians advance – then Chalmers's account of the role of theory will not do.

6 The Case of GTR

One of the sins that besets philosophers is their tendency to take a historical episode, view it as exemplifying their preferred approach, and then regard our intuitive endorsement of that episode as an endorsement of that approach. To combat this tendency in the HPS literature was one of the goals of Laudan's 1983 project – the so-called VPI program[1] – to test philosophies of science naturalistically. In my contribution to that project, I proposed that severe testing principles be applied on the "metalevel" (Mayo, 1988). This approach enjoins us to examine carefully how historical episodes might "fit" the philosophical account while actually giving an erroneous construal of both what happened and its epistemological rationale. Such a metastatistical critique, alas, is rare; many today largely regard such HPS case studies as mere illustrations, however interesting in their own right.

[1] Laudan launched this program by means of a conference at Virginia Tech, then called Virginia Polytechnic Institute and State University (Donovan, et al., 1988, 1992).

Alluding to the example of experimental gravitation given by Clifford Will, Chalmers claims that the "ways in which the development of that theory as construed by Will fits closely with my picture and poses problems for Mayo" (this volume, p. 69). But Chalmers's construal fails to square with that episode in much the same way that the comparativist account fails. The fact that GTR accorded with a variety of results (deflection effect, Mercury, redshift – the three classical tests) would warrant inferring all of GTR for Chalmers; but far from taking this as warranting all of GTR, scientists instead wished to build the PPN parameterization, whose rationale was precisely to avoid accepting GTR prematurely. The PPN framework houses a host of theoretical models and parameters, but the secret to its successful use was that it allowed hypotheses about gravitation to be tested without having to assume any one full gravitational theory, as we already discussed at length (Chapter 1, this volume). Will calls this a "gravitation theory-free" approach, by which he means we do not have to assume any particular theory of gravity to proceed in probing those theories. Because this stage of testing GTR has already been discussed, I turn to Chalmers's appeals to the more contemporary arena, wherein experimental gravitation physicists test beyond the arenas in which viable gravity theories have thus far been severely tested (e.g., in learning about quasars, binary pulsars, and gravity waves).

Here Chalmers's position seems to find support in Will's remarks that "when complex astrophysical systems [are involved] a gravitation-theory independent approach is not useful. Instead, a more appropriate approach would be to assume, one by one, that individual theories are correct, then use the observations to make statements about the possible compatible physics underlying the system. The viability of a theory would then be called into question if the resulting 'available physics space' were squeezed into untenable, unreasonable, or ad hoc positions." (Will, 1993, p. 303)

However, rival theories T_1 and T_2 are not accepted as true when used conditionally in order to derive consequences; else in considering "one by one" the predictions of rival gravity theories, the scientist would be forced to regard as warranted both T_1 and T_2. So unless we are to imagine that Chalmers is endorsing arguments from coincidence to mutually inconsistent rivals, he cannot really mean to say, as he does, that in moving to unexplored domains scientists take the theory as warranted in the untested domain. Instead they deliberately use the gaps in existing tests to *stress* theories further – to identify hurdles for lopping off some that have thus far survived.

Clifford Will brings out a very important insight into the roles of background theories in testing a primary theory T; namely, where we enter domains involving uncertain physics we may not be able to test "cleanly"

in his sense, or with severity in mine. What clean (severe) tests enable us to do is to *detach inferences* (in this case about gravity) and thereby shrink the possible alternative theories of gravity – in what he calls a "theory-independent way. The use of the PPN formalism was a clear example of this approach. The result was to squeeze theory space" (Will, 1993, p. 303). In cases where we cannot do this, we may at most hypothetically assume now this theory and then the other, in the hope that the predictions from some of the viable theories will be so qualitatively off that even so imprecise a "test" enables some to be killed off (which does not prevent their being used again for purposes of learning).

Chalmers is correct to note that in some cases GTR is used as a tool for measuring astrophysical parameters in the binary pulsar – we become "applied relativists" in Will's terminology. However, he overlooks the careful arguments that ensure the robustness of the resulting inferences. In particular, although "gravitational theory" is "used," it is used in such a way that avoids invalidating any inferred measurements. Several strategies exist to achieve this end. For instance, in using relativistic gravity to estimate the masses of the binary pulsar and its twin, one may rely on aspects for which all viable gravity theories agree, or they may conservatively take the values of parameters that represent the furthest a gravitation theory can disagree with GTR values. "The discovery of PSR 1913 + 16 caused considerable excitement in the relativity community... because it was realized that the system could provide a new laboratory for studying relativistic gravity"; in general, "the system appeared to be a 'clean' laboratory, unaffected by complex astrophysical processes" (Will, 1993, p. 284). Here, "relativistic gravitational theory" – but no one theory within the viable set – is used as a tool to estimate statistically such parameters as the mass of the pulsar. By obtaining knowledge of relativistic effects without assuming the truth of any one relativistic theory, we may opportunistically use the severely passed relativistic hypotheses to increase knowledge of relativistic phenomena such as gravity waves.

In particular, experimental relativists were able to contrast the predictions regarding the effects of gravity waves on the pulsar's orbit (in time for the centenary of Einstein's birth in 1979). Hypothetically assuming alternative theories of gravitation, they discover that one theory, Rosen's bimetric theory, "faces a killing test" by yielding a prediction qualitatively different – the orbit should slow down rather than speed up (Will, 1993, p. 287; Mayo, 2000a). The orbital decay that is estimated is in sync with GTR, but this is not regarded as providing reliable evidence for GTR; at most it provides indirect evidence for the existence of gravity waves. The

adjustable parameter in Brans–Dicke theory prevents the binary results from discriminating between them: "the theoretical predictions are sufficiently close to those of general relativity, and the uncertainties in the physics still sufficiently large that the viability of the theory cannot be judged reliably" (Will, 2004, p. 307). In interpreting the results, in other words, there is a careful assessment to determine what is ruled out with severity. The techniques by which uncertainties are "subtracted out" are part of the day-to-day measurements in experimental gravity, and their properties need to be understood (and critically evaluated) by philosophers of science if they are to learn from the episode. Far from providing grounds that all of T must be accepted as true for this opportunistic learning, the implications of Chalmers's arguments reinforce our claim that "enough experimental knowledge will do" in making progress.

7 The Role of Theories and the Error-Statistical Perspective

The error-statistical conception of the roles of theory admittedly is at odds with some standard conceptions. These differences, or so our exchanges have seemed to show, may result in our meaning different things using the same terms. I consider three areas that still need to be developed:

1. *Not One by One Elimination.* In claiming large-scale theories may pass severely through piecemeal tests, theory testers (Laudan, Chalmers, Musgrave) suppose I mean we must plod through all the predictions and all the domains, and they conclude it cannot be done (Laudan, 1997: "there is little prospect of severity flowing up"). Some assume I would need an exhaustive account of all theories or all ways a given theory may fail and then eliminate them one by one. However, this process overlooks the ingenuity of experimental learning. By building up severely affirmed effects and employing robustness arguments, a single type of local inference – once severely corroborated – can yield the theory (along the lines of "A Big Shake-up Turns on a Small Result" in the case of Brownian motion; Mayo, 1996, p. 246). When Will asserts that "in fact, we conjecture that for a wide class of metric theories of gravity, the binary pulsar provides the *ultimate* test of relativistic gravity" (Will, 1993, p. 287), he is referring to the fact that a distinct type of gravity wave signature (reflecting its adherence to the strong equivalence principle), once found, could entail GTR. It matters not whether this is actually the case for this episode; what matters is the idea of large-scale appraisal turning on very local results.

2. *Understanding a Theory.* One often hears scientists make claims about having a correct or an incorrect *understanding* of a theory or of the way some processes of interest behave in some domain. This seems to be what is captured in talking about experimental knowledge in relation to theories. An example would be how experimental relativists claim that it is only when they began experimentally testing GTR that they really started to understand relativistic gravity. What is behind this type of claim is basically what I mean to capture in saying that what is learned with severity is "experimental knowledge" (certainly I did not mean to limit it to knowledge of "observables"). So even if it was somehow known in 1930 that GTR was true (despite the limited evidence), scientists could not be said to have correctly understood relativistic gravity – how it behaves in its variety of interactions and domains. (Work on explanation does not quite get at this goal; see Glymour, Chapter 8, this volume.)

Thus, what is learned with severity need not be well captured by writing down some hypotheses or equations that are a piece of the full theory, at least not as that is usually understood. It may be best captured by one or more hypotheses couched in one of the intermediate models linking data and theory. But even those hypotheses seem more like placeholders for the full comprehension of the mechanism or process involved.

The interest in deliberately probing the ways one may be wrong – deliberate criticism – stems from the goal of making progress in understanding and revealing how misunderstandings were concealed, which leads to my third point:

3. *Not Kuhnian Normal Science.* Those who champion accepting a large-scale theory, even knowing many domains have yet to be probed, often sound as if they see their task as providing the scientist with a valid excuse for holding onto T, as if scientists wished to have an alibi for adhering to the status quo. The Kuhnian "normal" scientist who will not rock the boat except when faced with a crisis of anomalies comes to mind.

Ironically, while these theory testers are hardly fans of Kuhn, the picture they paint is far too redolent of the Kuhnian normal scientist, at least to my tastes. Popper rejected ("as dangerous") the Kuhnian conception of normal science for just the kind of complacency it fostered, advocating instead the view of the scientist as continually looking for flaws and fallibilities (see Mayo, 1996, ch. 2). One may grant that Kuhnian normal scientists are being

perfectly "rational" and perhaps earning their paychecks, while denying this mindset would spark the kind of probing and criticism that leads to extending the frontiers of knowledge. Why then does so much philosophical talk about theory appraisal (even among followers of Popper) seem to reflect the supposition that it is desirable to have things settled? Taking seriously the goal of "finding things out," the scientist in my view is driven to create novel effects and to compel the replacement of hypotheses and theories with ones that not only give a more correct understanding about more domains, but also teach us where our earlier understanding was in error.

Looking back, these three points hook up neatly with the experimental theses delineated in Chapter 1, Mayo (1996). Admittedly, I was unaware of these connections until I took up the challenge to supply the error-statistical account of experiment with an account of large-scale theories – thanks to Chalmers and our many exchanges.

8 A Word on the Task of Metamethodology

The error-statistical philosophy of science requires identification of the epistemological rationale behind strategies (e.g., varying the results, replication, novelty). It will not do to identify an impressive historical episode and note that it exemplifies the strategy. The method we supply for pinpointing the rationale is to consider how it might contribute to controlling error probabilities and making inferences severely. This task of "metamethodology" (Mayo, 1996, p. 455) requires "(a) articulating canonical models or paradigm cases of experimental arguments and errors, and (b) appraising and arriving at methodological rules by reference to these models. Historical cases, if handled correctly, provide a unique laboratory for these tasks... [however] the data we need do not consist of the full scientific episode, all finished and tidied up. The data we need are the experimental data that scientists have actually analyzed, debated, used, or discarded... Especially revealing are the processes and debates that take place before the case is settled and most of the learning is going on."

Chalmers's rule – require demanding arguments from coincidence – is clearly in the spirit of demanding severity, but by leaving it at a vague level we cannot determine its valid use or rationale. The most familiar canonical exemplar of an argument from coincidence is statistical null hypothesis testing for ruling out "mere chance" in inferring a "real" or reproducible effect. The null hypothesis, H_0, asserts that the effect or observed agreement is "due to coincidence" and the test is designed to ensure that if in fact H_0 can adequately account for experimental results, then with high probability,

outcomes consistent with H_0 would occur (i.e., there is a low probability that the null hypothesis H_0 would be found false erroneously). In effect the nonnull hypothesis – that the effect is real or not chance – is given a "hard time" before data are regarded as evidence for it (see Mayo and Cox, Cox and Mayo, Chapter 7, this volume). The argument in the formal statistical arena provides an exemplar for the argument from coincidence more generally. However, to apply it more generally to cover cases without any explicit probability model, we erect various strategies for giving H_0 a "hard time." One such strategy is to require that the effect be repeatable in a "wide variety" of cases.

But not just any kind of intuitively varied results constitute relevant variety. The different trials should check each other, so that whatever causes an error in experiment E_1 would not also be causing an error in E_2. Otherwise the variability does not strengthen the severity of the inference. Consider my highly informal example (Chapter 1) of inferring that George had not gained weight by checking him with a variety of scales with known calibrations. This is best classified as a strategy for checking errors in measuring instruments or in underlying assumptions of statistical models. For example, if I tested George's weight using only a single scale, it would not severely warrant the hypothesis of no weight gain because it might be some property of this scale that is responsible. Of course, we would really like to check directly that our instruments are working or that the underlying assumption holds in the case at hand; it is precisely in cases where such checking is not feasible that arguments from variety are important. If the identical effect is seen despite deliberately varying the backgrounds that could mar any one result, then we can subtract out the effects of flawed assumptions. However, this would not be relevant variability for ruling out other claims such as explanations for his weight maintenance (e.g., a pill that acts as a thermostat causing more fat to burn with increased consumption – how I wish!). That sort of theory would require different kinds of variable results to be inferred with severity. My general point is that, without a clear understanding of the epistemic value of a given strategy, interpretations from cases in HPS may be illicit.

9 Progress?

In Chalmers's new book (2009), there are signs that our positions on theory testing are moving closer on a key point we have been discussing. Agreeing now that "theories can be partitioned into those parts that have been and those that have not been tested," at least in some cases, he is perhaps now prepared to concur that a theory is confirmed by an argument from coincidence only if "the successful tests cannot be accounted for by some

specified sub-set of the theory." It will be interesting to see whether a weaker notion of theory testing is still thought to be needed.

References

Chalmers, A. (1999), *What is This Thing Called Science?* 3rd ed., Open University Press, and University of Queensland Press.

Chalmers, A. (2009), *The Scientist's Atom and the Philosopher's Stone: How Science Succeeded and Philosophy Failed to Gain Knowledge of Atoms*, Springer, Dordrecht.

Donovan, A., Laudan, L., and Laudan, R. (1988), *Scrutinizing Science*, Kluwer, Dordrecht (reprinted by Johns Hopkins University Press, 1992).

Kuhn, T. (1962), *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago.

Laudan, L. (1997), "How About Bust? Factoring Explanatory Power Back into Theory Evaluation," *Philosophy of Science*, 64: 306–16.

Mayo, D. (1988), "Brownian Motion and the Appraisal of Theories," pp. 219–43 in A. Donovan, L. Laudan, and R. Laudan (eds.), *Scrutinizing Science*, Kluwer, Dordrecht (reprinted by Johns Hopkins University Press, 1992).

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

Mayo, D.G. (2000a), "Experimental Practice and an Error Statistical Account of Evidence," pp. S193–S207 in D. Howard (ed.), *Philosophy of Science*, 67 (Symposia Proceedings).

Will, C.M. (1993), *Theory and Experiment in Gravitational Physics*, Cambridge University Press, Cambridge (revised edition).

Will, C.M. (2004), "The Confrontation between General Relativity and Experiment," *Living Reviews in Relativity*, http://relativity.livingreviews.org/Articles/lrr-2001–4/title.html.

Related Exchanges

Chalmers, A. (2000), "'What Is This Thing Called Philosophy of Science?' Response to Reviewers of *What Is Thing Called Science?* 3rd Edition," *Metascience*, 9: 198–203.

Chalmers, A. (2002), "Experiment and the Growth of Scientific Knowledge," pp. 157–69 in P. Gardenfors, J. Wolenski, and K. Kijania-Placet (eds.), *In the Scope of Logic, Methodology and Philosophy of Science*, Vol. 1, Kluwer, Dordrecht.

Mayo, D.G. (2000b), "'What Is This Thing Called Philosophy of Science?' Review Symposium of A. Chalmers' *What Is This Thing Called Science?*" *Metascience*, 9: 179–88.

Mayo, D.G. (2002), "Theory Testing, Statistical Methodology, and the Growth of Experimental Knowledge," pp. 171–90 in P. Gardenfors, J.Wolenski, and K. Kijania-Placek (eds.), *In The Scope of Logic, Methodology and Philosophy of Science* (Vol. 1 of the 11th International Congress of Logic, Methodology, and Philosophy of Science, Cracow, August 1999), Kluwer, Dordrecht.

Staley, K. (2008). "Error-Statistical Elimination Of Alternative Hypotheses," *Synthese (Error and Methodology in Practice: Selected Papers from ERROR 2006)*, Vol. 163(3): 397–408.

Worrall, J. (2000), "'What Is This Thing Called Philosophy of Science?' Review Symposium of A. Chalmers' *What Is This Thing Called Science?*" *Metascience*, 9: 17–9.