FOUR

Theory Confirmation and Novel Evidence

Error, Tests, and Theory Confirmation

John Worrall

In this chapter I address what seems to be a sharp difference of opinion between myself and Mayo concerning a fundamental problem in the theory of confirmation.[1] Not surprisingly, I argue that I am right and she is (interestingly) wrong. But first I need to outline the background carefully – because seeing clearly what the problem is (and what it is not) takes us a good way towards its correct solution.

1 The Duhem Problem and the "UN" Charter

So far as the issue about confirmation that I want to raise here is concerned: in the beginning was the "Duhem problem." But this problem has often been misrepresented. No sensible argument exists in Duhem (or elsewhere) to the effect that the "whole of our knowledge" is involved in any attempt to test any part of our knowledge. Indeed, I doubt that that claim makes any sense. No sensible argument exists in Duhem (or elsewhere) to the effect that we can never test any particular part of some overall theory or theoretical system, only the "whole" of it. If, for example, a theory falls "naturally" into five axioms, then there is – and can be – no reason why it should be impossible that some directly testable consequence follows from, say, four of those axioms – in which case only those four axioms and not the whole of the theory are what is tested.

What Duhem *did* successfully argue is that if we take what is normally considered a "single" scientific theory – such as Newton's theory (of mechanics plus gravitation) or Maxwell's theory of electromagnetism or the wave theory of light – and carefully analyze any attempt to test it empirically by deducing from it some directly empirically checkable consequence, then

[1] See, e.g., Worrall (2006) and Mayo (1996).

the inference is revealed to be valid only *modulo* some further set of auxiliary theories (theories about the circumstances of the experiment, about the instruments used and so on). For example, as became clear during the famous dispute between Newton and Flamsteed, to deduce from Newton's theory of gravitation a consequence that can be directly tested against telescopic sightings of planetary positions, we need to invoke an assumption about the amount of refraction that a light beam undergoes in passing into the Earth's atmosphere.

Moreover, Duhem pointed out that at least in many cases the "single" theory that we test itself involves a "central" claim together with some set of more specific assumptions; and in such cases, so long as the central claim is retained, we tend to describe changes in the specific assumptions as producing "different versions of the same theory" rather than a new, different theory. An example of this, analyzed of course at some length by Duhem himself, is "the" (classical) wave theory of light. This "theory" was in fact an evolving entity with a central or core assumption – that light is some sort of wave in some sort of mechanical medium – an assumption that remained fixed throughout, with a changing set of more specific assumptions about, for example, the kind of wave and the kind of medium through which the waves travel. For example, one celebrated (and relatively large) change was that effected by Fresnel when he abandoned the idea that the "luminiferous ether" that carries the light waves is a highly attenuated fluid and the waves, therefore, longitudinal, and hypothesized instead that the ether is an (of course still highly attenuated) elastic solid that transmits *transverse* waves.

These facts about the deductive structure of tests of "single" scientific theories of course have the trivial consequence that no experimental result can refute such a theory. Even assuming that we can unproblematically and directly establish the truth value of some observation statement *O* on the basis of experience, if this observation statement follows, not from the core theory *T* alone but instead only from that core, plus specific assumptions, plus auxiliaries, then if *O* turns out in fact to be false, all that follows deductively is that *at least one* of the assumptions in the "theoretical system" involving core, plus specific, plus auxiliary assumptions is false.

Kuhn's account of "scientific revolutions" is to a large extent a rediscovery – of course unwitting – of Duhem's point (along with a great number of historical examples).[2] Perhaps the claim in Kuhn that most strikingly challenged the idea that theory change in science is a rational process is that in "revolutions" the old-guard (or "hold-outs" as he calls them) were no

[2] This is argued in Worrall (2003).

less rational than the "revolutionaries" – there being "some good reasons for each possible choice" (sticking to the older theory or accepting the newer one).[3] This claim in turn is at least largely based on Kuhn's observation that the evidence that the "revolutionaries" regard as crucial extra empirical support for their new paradigm-forming theory can in fact also be "shoved into the box provided by the older paradigm."[4] Exactly as Duhem's analysis of theory-testing shows, it is always logically possible to hold onto the basic (central or core) idea of the older theoretical framework by rejecting some other – either "specific" or auxiliary – assumption.

For example, results such as that of the two-slit experiment that were certainly correctly predicted by the wave theory of light are often cited by later accounts as crucial experiments that unambiguously refuted that theory's corpuscular rival. But in fact plenty of suggestions existed within the early nineteenth-century corpuscularist literature for how to accommodate those experimental results. Some corpuscularists, for example, conjectured that, alongside the reflecting and refracting forces to which they were already committed, results such as that of the two-slit experiment showed that there was also a "diffracting force" that emanates from the edges of "ordinary, gross" opaque matter and affects the paths of the light particles as they pass. Those corpuscularists laid down the project of working out the details of this diffracting force on the basis of the "interference" results. (Of course, because this force needs to be taken to pull some particles into the geometrical shadow and push others away from places outside the shadow that they would otherwise reach, the corpuscularists thereby denied that the fringe phenomena are in fact the result of interference.)

Similarly, and as is well known, Copernicus (and following him Kepler and Galileo) was especially impressed by his theory's "natural" account of the phenomenon of planetary stations and retrogressions – despite the fact that it had long been recognized by Copernicus' time that this phenomenon could be accommodated within the Ptolemaic geostatic system: although they are certainly inconsistent with the simplest Ptolemaic theory, which has all planets describing simple circular orbits around the Earth, stations and retrogressions could be accommodated within the Ptolemaic framework by adding epicycles and making suitable assumptions about their sizes and about how quickly the planet moved around the epicycle compared to how quickly the center of that epicycle moved around the basic "deferent" circle.

[3] Kuhn (1962, pp. 151–2).
[4] Kuhn (1977, p. 328).

Kuhn seems to presume that the fact that such phenomena can be accommodated within the older "paradigm" means that the phenomena cannot unambiguously be regarded as providing extra support for the newer theoretical framework and as therefore providing part of the reason why the theory shift that occurred was rationally justified. But this presumption is surely wrong. It is, instead, an important part of any acceptable account of theory confirmation that merely "accommodating" some phenomenon within a given theoretical framework in an ad hoc way does *not* balance the evidential scales: the theory underlying the framework that predicted the phenomenon continues to receive greater empirical support from it, even if it can be accommodated within the older system (as Duhem's analysis shows is always possible). The wave theory continues to derive more support from the result of the two-slit experiment even once it is conceded that it is *possible* to give an account of the phenomenon, though in an entirely post hoc way, within the corpuscular framework. Planetary stations and retrogressions give more (rational) support to the Copernican theory even though the Ptolemaic theory can accommodate them (indeed even though the Ptolemaic theory had, of course, long *pre*-accommodated them).

A suspicion of ad hoc explanations has guided science from its beginning and is widely held and deeply felt. Take another (this time noncomparative) example. Immanuel Velikovsky conjectured that in biblical times a giant comet had somehow or other broken away from the planet Jupiter and somehow or other made three separate series of orbits close to the Earth (before eventually settling down to a quieter life as the planet Venus). It was these "close encounters" that were responsible for such biblically reported "phenomena" as the fall of the walls of Jericho and the parting of the Red Sea. Velikovsky recognized that, if his theory were correct, it is entirely implausible that such cataclysmic events would have been restricted to the particular part of the Middle East that concerned the authors of the Bible. He accordingly set about looking for records of similar events from other record-keeping cultures of that time. He found records from *some* cultures that, so he (rather loosely) argued, fit the bill, but he also found some embarrassing gaps: the apparently fairly full records we have inherited from some other cultures make no mention of appropriately dated events that were even remotely on a par with the ones alleged to have occurred in the Bible. Velikovsky – completely in line with Duhem's point – held on to his favoured central theory (there really had been these close encounters and widespread associated cataclysms) and rejected instead an auxiliary assumption. Suppose that similar cataclysms *had* in fact occurred in the homeland of culture *C*. To predict that *C*'s scribes would have recorded such events (which were, after all, one would have thought, well worth a line in anyone's diary!) it must of course be assumed that those scribes were able to bear accurate witness. But what if, in some cultures, the events associated with the close encounters with this "incredible chunk" proved *so* cataclysmic that all of the culture's scribes were afflicted by "collective amnesia?" Velikovsky conjectured that collective amnesia had indeed afflicted certain cultures and proceeded to read off which exact cultures those were from the (lack of) records. Those cultures that recorded cataclysms that he had been able to argue were analogous to the biblical ones did *not* suffer from this unpleasant complaint; those cultures that would otherwise have been expected to but did not in fact record any remotely comparable events *did* suffer from it. Clearly although this modified theory now entails correctly which cultures would and which would not have appropriate records, this can hardly be said to supply any empirical support to Velikovsky's cometary hypothesis – a hypothesis that has been augmented exactly so as to yield the already known data.

These intuitive judgments need to be underwritten by some general principle that in turn will underwrite the rejection of Kuhn's implicit claim that the fact that evidence can be forced into the "box provided by the older paradigm" means that that evidence cannot be significant extra support for the newer theoretical framework. It was this idea that led some of us to sign up to the "UN Charter."[5] This phrase, in slogan form, has been interpreted as ruling that "you can't use the same fact twice, once in the construction of a theory and then again in its support." According to this "use novelty criterion" or "no-double-use rule" as it has generally been understood, theories are empirically supported by phenomena that they correctly *predict* (where prediction is understood, as it invariably is in science, not in the temporal sense but in the sense of "falling out" of the theory without having had to be worked into that theory "by hand")[6] and *not* by phenomena that have to

[5] For history and references see Worrall (2002).

[6] Notice that, despite the fact that UN stands for "use novelty," the UN charter in fact gives no role to novelty of evidence in itself. Those who had argued that evidence that was, as a matter of historical fact, discovered only as a result of its being predicted by some theory carried greater confirmational weight were missing the real issue. This issue is one of accommodation versus prediction, where the latter is used in the proper sense, just meaning "not accommodated." Some but not all predictions are of hitherto unknown phenomena (although all accommodations must of course have been of known phenomena). This sense of prediction is accurately reflected in the following passage from French's textbook, *Newtonian Mechanics*: "[L]ike every other good theory in physics, [the theory of universal gravitation] had predictive value; that is, it could be applied to situations besides the ones from which it was deduced [i.e., the phenomena that had been deliberately accommodated within it]. Investigating the predictions of a theory may involve looking

be "accommodated within," or "written into" the theory post hoc. Thus, it yields the judgment that the (amended) corpuscular theory gets no support from the two-slit experiment because the details of the "diffracting force" had to be "read off" from already-given experimental results such as that of the two-slit experiment itself, whereas the wave theory, which predicted this experimental outcome in a way that is entirely independent of that outcome, *does* get support from the result. Similarly Velikovsky's (amended) theory gets no support from the empirical fact that no records of cataclysms in culture *C* have been preserved, because the facts about which cultures have or have not left records of appropriately timed cataclysms were used in constructing the particular form of his overall theory that he defended.

It is a central purpose of this chapter to clarify further and defend the UN rule under a somewhat different interpretation than the one it has often been given and in a way that clashes with Mayo's (partial) defence of that rule. Many philosophers have, however, claimed that the view is indefensible in any form – a crucial part of the clarification consists in showing how exactly these opponents of the view go astray.

2 "Refutations" of the UN Rule

Allan Franklin once gave a seminar at the London School of Economics under the title "*Ad Hoc* Is Not a Four Letter Word." Beneath the (multiple) surface literal correctness of this title is a substantive claim that is undeniably correct; namely, it is entirely normal scientific procedure to use particular data in the construction of theories, without any hint of this being in any way scientifically questionable, let alone outright intellectually reprehensible.

Suppose, for (a multiply realized) example, that a scientist is facing a general theory in which theoretical considerations leave the value of some parameter free; the theory does, however, entail that the parameter value is a function of some set of observable quantities. A particular example of this kind that I like to use is that of the wave theory of light, which leaves as an open question, so far as basic theoretical considerations are concerned, what is the wavelength of the light from any particular monochromatic source? Because the theory provides no account of the atomic vibrations within luminous objects that produce the light, it does not dictate from first principles the wavelength of light from a particular source. The theory

for hitherto unsuspected phenomena, or it may involve recognising that an already existing phenomenon must fit into the new framework. In either case the theory is subjected to searching tests, by which it must stand or fall" (French, 1971, pp. 5–6).

does, however, entail that that value, whatever it is, is a function of the slit distances, the distance from the slits to the screen and the fringe distances in the two-slit experiment. In such a situation, the scientist will *of course* not make "bold conjectures" about the value of the wavelength of light from some particular source and then test those conjectures. Instead, she will "measure the wavelength;" that is, she will perform the experiment, record the appropriate observable values, and infer the wavelength of light from that particular source from the formula entailed by the theory. She has then *deduced* a particular, more powerful theory (wave theory complete with a specific value of this particular theoretical parameter) from her general (parameter-free) theory plus observational results.

Clearly using observational data as a premise in the deduction of some particular version of a theory is a paradigmatic example of "using data in the construction of a theory." And yet this is an entirely sensible, entirely kosher scientific procedure. Moreover, if asked why she holds the particular version of the theory that she does – that is, if she is asked why, given that she holds the general wave theory of light, she also attributes this particular wavelength to light from this particular source – she will surely cite the observations that she has used in that deduction. What then of the "rule" that you can't use the same fact twice, once in the construction of a theory and then again in its support?

Nor do the apparent problems for the UN rule end there. Colin Howson, for example, likes to emphasize a different general case – standard statistical examples such as the following (see Howson, 1990). We are given that an urn contains only red and white balls though in an unknown (but fixed) proportion; we are prevented from looking inside the urn but can draw balls one at a time from it. Suppose that a sample of size n has been taken (with replacement) and k of the balls have been found to be white. Standard statistical estimation theory then suggests the hypothesis that the proportion of white balls in the urn is $k/n \pm \varepsilon$, where ε is calculated as a function of n by standard confidence-interval techniques. The sample evidence is the basis here of the construction of the particular hypothesis and surely, Howson suggests, also supports that particular hypothesis at least to some degree – the (initial) evidence for the hypothesis just *is* that a proportion k/n of the balls drawn were white. For this reason (and others) Howson dismisses the UN rule as "entirely bogus."[7]

Mayo cites and analyzes in more detail similar statistical cases that seem to count against the "no-double-use idea" and also cites the following "trivial

[7] See Howson (1990).

but instructive example" (1996, p. 271). Suppose one wanted to arrive at what she characterizes as a "hypothesis H" about the average SAT score of the students in her logic class. She points out that the "obvious" (in fact uniquely sensible) way to arrive at H is by summing all the individual scores of the *N* students in the class and dividing that sum by *N*. The "hypothesis" arrived at in this way would clearly be "use constructed." Suppose the constructed "hypothesis" is that the average SAT score for these students is 1121. It would clearly be madness to suppose that the data used in the construction of the "hypothesis" that the average SAT score is 1121 fail to support that hypothesis. On the contrary, as she writes,

> Surely the data on my students are excellent grounds for my hypothesis about their average SAT scores. It would be absurd to suppose that further tests would give better support. (1996, p. 271)

Exactly so: the data provide not just excellent, but, short of some trivial error, entirely *conclusive* grounds for the "hypothesis"– further "tests" are irrelevant. (This is precisely why it seems extremely odd to talk of a "hypothesis" at all in these circumstances – a point to which I return in my criticism of Mayo's views.)

How in the light of apparently straightforward counterexamples such as these can I continue to defend (a version of) the UN "rule"? Well, first we need to get a clearer picture of the underlying nature of all these "counterexamples." They all are (more or less clear-cut) instances of an inference pattern sometimes called "demonstrative induction" or, better, "deduction from the phenomena." The importance of this inference pattern to science was emphasized long ago by Newton and, after some years of neglect, has been increasingly reappreciated in recent philosophy of science.[8]

Of course, general theories are invariably logically stronger than any finite set of observational data and so a "deduction from the phenomena," if it is to be valid, must in fact implicitly involve extra premises. The idea is that certain very general principles are, somehow or other, legitimately taken for granted (as "background knowledge") and some more specific theory is deduced from those general principles plus experimental and observational data. Newton, in a complicated way that involves generalizing from models known to be (strictly) inaccurate, deduced his theory of universal gravitation from Kepler's "phenomena" plus background assumptions that included conservation of momentum. The statistical case cited by Howson, exactly because it is statistical, does not of course exactly fit the pattern – but

[8] See Worrall (2000) and the references to the literature therein.

something very similar applies. We are somehow given (or it seems reasonable to assume) that drawing balls from an urn (with replacement) is a Bernoulli process, with a fixed probability p – we then "quasi-deduce" from the fact that the sample of draws has produced a proportion k/n of white balls that the population frequency is $k/n \pm \varepsilon$. In Mayo's case, we deduce her "hypothesis" about the average SAT score of her logic students from background principles (basically the analytic principles that specify what an average is) plus the "observed" individual student scores. (The fact that the background principles in this last case are analytic is another reflection of the oddness of characterising the resultant claim as a "hypothesis.")

Of the cases cited, the most direct instance of the type of reasoning that is at issue (and that plays an important role in physics) is the wave theory case. The scientist starts with a theory, $T(\lambda)$, in which the theoretical parameter (in this case wavelength) is left free. However, the theory entails that λ is a determinate function of quantities that are measurable. Here the wave theory, for example, entails (subject to a couple of idealisations) that, in the case of the famous two-slit experiment performed using light from a monochromatic source – say, a sodium arc – the (observable) distance X from the fringe at the center of the pattern to the first fringe on either side is related to (theoretical) wavelength λ, via the equation $X/(X^2 + D^2)^{1/2} = \lambda/d$ (where d is the distance between the two slits and D the distance from the two-slit screen to the observation screen – both, of course, observable quantities). It follows analytically that $\lambda = dX/(X^2 + D^2)^{1/2}$. But all the terms on the right-hand side of this last equation are measurable. Hence, particular observed values e' determine the wavelength of the light (within some small margin of experimental error, of course) and so determine the more specific theory $T' = T(\lambda_0)$, with the parameter that had been free in T now given a definite value, λ_0 – again within a margin of error.

As always, "deduction from the phenomena" here really means "deduction from the phenomena plus general 'background' principles." In this case, the general wave theory with free parameter is given, and we proceed, against that given background, to deduce the more specific version with the parameter value fixed from the experimental data.

This case is clear and illustrative but rather mundane. More impressive cases exist such as Newton's deduction of his theory of universal gravitation from the phenomena or the much more recent attempt, outlined by Will and analysed by Earman,[9] to deduce a relativistic account of gravitation from phenomena. These cases involve background principles of extreme

[9] See Earman (1992, pp. 173–80); see also the discussion in Mayo (1996) and this volume.

generality that seem natural (even arguably "unavoidable"). These general principles delineate a space of possible – more specific – theories. Taking those principles as implicit premises, the data, by a process that can be characterized either as "deduction from the phenomena" or equivalently as "demonstrative induction," gradually cut down that possibility space until, it is hoped, just one possible general theory remains. Taking the simple case where the background principles specify a finite list of alternatives $T_1, \ldots, T_n$, each piece of data falsifies some T_i until we are left with just one theory, T_j – which, because the inference from (T_1 v T_2 v . . . v T_n) and $\neg T_1$, $\neg T_{j-1}$,$\neg T_{j+1}$, . . . , $\neg T_n$ to T_j is of course deductively valid – is thus "deduced from the phenomena."

Clearly such a deduction, if available, is very powerful – it shows, if fully successful, that *the* representative of the very general background assumptions at issue is dictated by data to be one general but particular theory T_j. The data *e* in such a case therefore provide powerful support for T_j in a very clear and significant sense: the data *dictate* that if any theory that satisfies these natural assumptions can work then it must be T_j.

In the less exciting but more straightforward wave theory case, the data from the two-slit experiment uniquely pick out (modulo some small error interval) the more particular theory T' (with precise value of λ_0 for the wavelength of light from the sodium arc) as the more specific representative of the general wave theory. If you hold the *general* wave theory already, then data dictate that you hold T'.

In Mayo's still simpler case the general background principles are analytic – stating in effect just what the notion of an average *means*. And, hence, the data from her students *dictate* that the average SAT score is 1121 and, therefore (in a very stretched sense), support (maximally, of course) the "hypothesis" that the SAT average is 1121.

Again, because of its statistical character, Howson's standard statistical estimation case does not *quite* fit, but essentially the same situation holds. The basic model is again treated – or so we suppose – as a given: it is taken that this is a Bernoulli process with fixed probability *p*. Of course, in this case the interval estimate for the proportion of white balls to red balls is not *deduced* from the data provided by the sample, but it might be said that it is "quasi-deduced" in line with standard statistical procedure.

In all these cases, then, a clear sense exists in which the theory is "deduced from the phenomena *e*" and yet is given strong support by *e*. In the wave theory case, for example, the result of the two-slit experiment using light from the sodium arc deductively entails $T(\lambda_0)$, the specific version of the wave theory with the wavelength of that light fixed, and what better support or confirmation could there be than deductive entailment? The "no-double-use rule" seems therefore to be entirely refuted.

3 Two Qualitatively Distinct Kinds of "Confirmation" or "Empirical Support": How to Get the Best of Both Worlds

The "UN" or "no-double-use" rule is not, in fact, refuted by the support judgments elicited in the cases discussed in Section 2; instead it simply needs a little elaboration. The principal step towards seeing this is to recognize just how conditional (and *ineliminably* conditional) the support at issue is in all these cases.

In the wave theory case, for example, the judgment that the result of the two-slit experiment with sodium light strongly supports the specific version of the theory $T(\lambda_0)$ is entirely dependent on the prior acceptance of the general wave theory $T(\lambda)$. Insofar as we already have good empirical reason to "accept" that general theory (whatever exactly that means!), the deduction from the phenomena outlined in Section 2 shows that we have exactly the same reason to accept the more specific theory, $T(\lambda_0)$. The "deduction from the phenomena" *transfers* whatever empirical support the general theory already had to the more specific theory that is the conclusion from that deduction. But it surely does not add anything to the support for the more general theory – which was not in any sense tested by this experiment. Of course, so long as the experimental results (that is, the slit and fringe distances) satisfy the general functional formula entailed by that general theory, then *any* particular outcome – any distance between the central bright fringe and the first dark fringe to either side, for example – is consistent with the general theory. A set of fringe distances different from those actually observed (assuming again that the set had the same functional features c – central bright band, symmetrically placed dark bands on either side of that central band, etc.) would not, of course, have led to the rejection of $T(\lambda)$ but simply to the construction or deduction of a *different* specific version – say, $T(\lambda_1)$ – of that same general theory. The fact, then, that $T(\lambda_0)$ entails the correct fringe, slit, and screen distances in the two-slit experiment with sodium light from which it was constructed provides no *extra* empirical reason at all for holding the general theory $T(\lambda)$.

The conditional nature of this sort of empirical support for some relatively specific theory – conditional, that is, on independent empirical support for its underlying general theory being already present – is further underlined by the fact that the sort of theoretical maneuvers that give *ad hocness* a bad name fit the model of "deduction from the phenomena." Consider, for example, the Velikovsky dodge outlined earlier. We can readily reconstruct Velikovsky's overall general theoretical framework (involving not just his assumptions about Jupiter but also about how the (alleged) subsequent terrestrial cataclysms would be reported by appropriate scribes) as employing a free (functional) parameter indicating whether or not the scribes in society *S* were afflicted by collective amnesia. And then his more specific theory involving claims about which particular societies were, and which were not, afflicted by collective amnesia follows deductively from his general theory plus the "phenomena" (here of course the records, or lack thereof, of appropriate cataclysms). And the deduction proceeds in exactly the same way – both the wave theory and the Velikovsky cases are then instances of "parameter-fixing."

The difference between the wave theory and Velikovsky cases is simply that, in the former but not the latter, *independent support* existed for the general theory ahead of the deduction from the phenomena. But that aside, the logic is identical: in both cases the deduction does no more and no less than *transfer* the empirical support enjoyed by the general theory to the specific deduced theory; it is just that in the Velikovsky case there is no such empirical support for the general theory that could be transferred.[10]

Again there is no question of the underlying theory getting any support from the data at issue and for exactly the same reason as in the wave theory case. The data of records from some cultures, and lack of them from others, do nothing to support the general idea of cataclysms associated with close encounters with the alleged massive comet, because that general theory (once equipped with a "collective amnesia parameter") is not tested by any such data – different data would not have led to the rejection of Velikovsky's general theory but instead simply to a specific version different from the one that Velikovsky actually endorsed given the actual data he had. (This different version would, of course, have simply had a different series of values for the "collective amnesia parameter.")

The sort of confirmation or empirical support involved in these cases is what might be called "purely intra-framework" or "purely intra-research program support." The lack of records in cultures $C_1, \ldots, C_n$ and their (arguable) presence in $C'_1, \ldots, C'_m$ gives very good reason for holding the

[10] I am assuming throughout this discussion that the *only way* in which Velikovsky could reconcile the lack of records of suitable cataclysms from some record-keeping cultures within his general theory was via the collective amnesia dodge. Because of the relative laxity of his theory, this is of course far from true. I am therefore idealizing somewhat to make it a crisp case of deduction from the phenomena (it is not really as good as that!). But I believe that all the methodological points stand in spite of this slightly idealizing move.

specific collective-amnesia version of Velikovsky's theory that he proposed *if* you already hold Velikovsky's general theory, *but* (and this is where the initial UN intuitions were aimed) those data give you absolutely no reason at all for holding that general theory in the first place. (Although there might, of course, have been other empirical reasons for doing so, it is just that as a matter of fact in this case there were not.) The data from the two-slit experiment give you very good (in fact, to all intents and purposes, *conclusive*) reason to hold the specific version of the wave theory with the particular value of the wavelength for light from a sodium arc *if* you already hold the general wave theory, *but* the data give you absolutely no reason at all for holding that general theory in the first place (although of course there may have been – and in this case actually were – other empirical reasons for doing so).

Not all empirical confirmation or support can have this ineliminably conditional and ineliminably intra-program character. After all, as we just saw, it seems clear that the difference between the general wave theory and the general Velikovskian theory is that the former has empirical support, which the latter lacks. *Some* general theories – the wave theory of light, but not the general Velikovsky theory – have independent empirical support; that is, empirical reasons exist for holding those general theories ahead of the sort of conditional confirmation (or demonstration) of some particular version of them from data. How can this be, especially in view of the fact that the Duhem thesis implies that all confirmation is of general theories plus extra assumptions? The answer must be that cases exist in which, in contrast to the cases of confirmation we have just considered, confirmation "spreads" from the theoretical framework (central theory plus specific assumptions) to the central theory of the framework – that is, some empirical results must exist which – rather than giving us good reason to accept some specific version of a general theory, given that we have already accepted the general theory – in fact give us good reason to accept the underlying general theory itself (and this despite the fact that the result, in line with Duhem's point, only follows deductively from some specific version of the theory *plus* auxiliaries).

Two kinds of case seem to exist where this occurs. The first is easy to describe. Having used data to fix the value of some parameter in a general theory – that new specific theory complete with parameter value, as well of course as giving you back what you gave to it by entailing the "used" data – may go on to make further predictions that are *independent* of the used data. Thus, for example, the general wave theory entails not only a general functional relationship between wavelengths and quantities measurable in the two-slit experiment but also another general functional relationship between wavelengths and quantities measurable in other experiments – for example, the one-slit diffraction experiment. Thus, having gone from $T(\lambda)$ with free parameter λ plus evidence e about slit separations and fringe distances in the two-slit experiment to the "specific" theory $T(\lambda_0)$, $T(\lambda_0)$ not only entails the original two-slit data e (of course it does!), it also makes an independently testable prediction about the fringe distances produced by light from the same source in the entirely different one-slit experiment. Moreover, this prediction turns out (of course, entirely nontrivially) to be correct. Similarly – in another much-discussed case – Adams and Leverrier, having used evidence e about the Uranian "irregularities" to deduce the existence of a further planet produced a modified Newtonian framework that not only gets e – that is, Uranus's orbit – correct (of course it is bound to) but also makes independent predictions e' about the existence and orbit of Neptune, predictions that again turn out to be correct.

The independent evidence e' – the one-slit result in the case of the wave theory and the observations of Neptune in the Adams–Leverrier case – surely gives *unconditional* support to the general underlying theory: not just support for the wave theory made more specific by fixing parameter λ conditional on the general theory that light consists of waves through a medium, but support for that general theory itself; not just support to the Newtonian system that is committed to a particular assumption about the number of planets, conditional on the basic Newtonian theory, but to the fundamental Newtonian theory itself. So alongside the conditional intra-research program confirmation that is obtained in all the cases discussed in Section 2, a second, more-powerful kind of confirmation exists that provides support for the general theory, or research program, itself. What the UN rule was saying all along, and saying correctly, is that this unconditional kind of support for the underlying general theory involved cannot (of course!) be obtained when the evidence concerned was used in the construction of the specific theory out of that general framework.

Given that in both wave theory and Newtonian cases, the specific theory constructed using evidence e turns out to be independently tested and confirmed by evidence e' (in contrast, of course, to the Velikovsky case of no independent testability), it might seem reasonable to count the used evidence as itself supportive. Given that Adams–Leverrier-amended Newtonian theory makes correct predictions about Neptune, the evidence about Uranus's orbit from which it itself was "deduced" can count as evidence for it, too; given that the wave theory complete with wavelength for sodium light deduced from the two-slit result is independently confirmed by its prediction of the one-slit result with light from the same source, the two-slit result can also count as (unconditional) support for the general wave theory. But this seems to me prejudicial as well as unnecessary and misleading. If Velikovsky is to get only conditional support from the lack of records in culture *C*, then, because the logic is exactly the same, so should the amended Newton theory from the evidence concerning Uranus. The difference between the two is simply, to repeat, that Newton's theory garnered lots of the unconditional kind of support, whereas Velikovskian-specific theories have *only* support conditional on a framework that itself has no support. Two quite different sorts of scientific reasoning seem to be involved after all – obtaining support for a general theory from data and *using* data to construct specific versions of that general theory.

There is at least one respect in which matters are sometimes slightly more complicated. Not perhaps invariably, but certainly quite often, the value of a parameter within a powerful general theory is *overdetermined* by the data. Indeed this is bound to be held whenever, as in the wave theory case discussed earlier, the fixing of a parameter via one experimental result leads to a theory that is (successfully) independently testable via a further experimental result. The general wave theory entails not just one but a *number* of functional relationships between wavelength – clearly a theoretical parameter – and measurable quantities in a range of *different* experiments. So, for example, a mid-nineteenth-century wave theorist could just as well have used the results from the *one-slit* diffraction experiment to fix the value of the wavelength of monochromatic light from some particular source and then have gone on to predict the outcome of the two-slit experiment performed using that same light. This, in the end, would be equivalent to the converse process that I just described in which the theorist uses the results from the two-slit experiment to fix the parameter and then goes on to predict the one-slit result. In general, there may be a series of experimental results $e_1, \ldots, e_n$, any (proper) subset of which of some size r can be used to fix parameter values; then the underlying general theory with these fixed parameter values predicts the remaining $n - r$ pieces of evidence. There is clearly no a priori guarantee that the set of data $e_1, \ldots, e_n$ admits any consistent assignment of values for the theoretical parameter at issue – it will do so if and only if the results of the $(n - r)$ independent tests of the theory once the parameter has been measured using r of the results are positive.

Clearly, in cases where this does indeed happen, the data set $e_1, \ldots, e_n$ tells us something positive about the underlying theory. It would not seem unreasonable to say, as I believe Mayo would, that this data set is *both* used in the construction of the theory *and at the same time* "severely" tests it. And this judgement would again seem to be in clear conflict with the "no-double-use rule." However, this judgment is surely coarse-grained. What really (and, once you think about it, pretty obviously) ought to be said is that *part* of the evidence set fixes parameters in the underlying general theory and then *part* of that set tests the resulting, more specific version of the theory. It is just that in such a case it does not matter which particular subset of size r you think of as doing the parameter-fixing and which remaining subset of size $n - r$ you think of as doing the testing. Nonetheless, two separate things *are* going on that are dependent on different bits of data: genuine *tests* of a theory and *application* of a theory to data to produce more specific theoretical claims.

This may seem an unnecessary quibble – why not just agree that the "no-double-use rule" fails in such cases? The evidence set is used *both* in the construction of the specific theory involved *and* in its (unconditional) support. One reason is as follows: suppose we had two theories, T and T', one of which, say T, has no relevant free parameters and entails $e_1, \ldots, e_n$ straight off, whereas T' involves parameters that are left free by theoretical considerations and need to be fixed using r of the evidential results e_i. Surely we would want to say in such a circumstance that the evidential set $e_1, \ldots,$ e_n supports T *more* than it does T'? If so, then there must be some confirmational "discount" for parameter-fixing: speaking intuitively, in such a case, T gets n lots of (unconditional) confirmation from the data set, whereas T' gets only $n - r$ lots. How much of the data set is needed to fix parameters plays a role in the judgment of how much (unconditional) support the theory gets from the data set. And this condition holds even when the choice is arbitrary of which particular subset (of a certain size) is used to fix parameters and which is used to genuinely test and, hence, (possibly) supply genuine "unconditional" support. In this sense, although the set as a whole, if you like, both fixes parameter values and (unconditionally) supports, *no particular element of the data set does both.*

I said that two kinds of case exist where support is unconditional – two kinds of case in which support "spreads" from the specific theory that entails the evidence to the underlying general theory. The first of these is the case of independent testability that we have just considered. The second type is equally important though somewhat trickier to describe precisely. This sort of confirmation (again, of the general underlying theory, rather than of some specific theory, *given* the general underlying theory) is provided in cases in which, roughly speaking, some prediction "drops out of the basic idea" of the theory. Here is an example.

The explanation of the phenomena of planetary stations and retrogressions within the Ptolemaic geocentric theory is often cited as a classic case of an ad hoc move. The initial geocentric model of a planet, say Mars, travelling on a single circular orbit around a stationary Earth, predicts that we will observe constant eastward motion of the planet around the sky (superimposed, of course, on a constant apparent diurnal westward rotation with the fixed stars); this prediction is directly refuted by the fact that the generally eastward (apparent) motion of Mars is periodically interrupted by occasions when it gradually slows to a momentary halt and then begins briefly to move "backwards" in a westward direction, before again slowing and turning back towards the east (remember that it never moves or even seems to move backwards on any particular night because the diurnal movement is always superimposed). The introduction of an epicycle of suitable size and the assumption that Mars moves around the center of that epicycle at a suitable velocity while the whole epicycle itself is carried around the main circular orbit (now called the deferent) leads to the correct prediction that Mars will exhibit these stations and retrogressions. Although not as straightforward as normally thought, this case surely is one that fits our first, entirely conditional, kind of confirmation – if you *already accept* the general geocentric view, then the phenomena of stations and retrogressions give you very good reason to accept (and in that sense they strongly confirm) the particular version of geocentricism involving the epicycles.[11] However, the fact that stations and retrogressions are "predicted" (or better, entailed) by the specific version of geocentricism with suitable epicyclic assumptions gives absolutely no further reason to accept (and so no support for, or confirmation of) the underlying basic geocentric (geostatic) claim.

The situation with Copernican heliocentric (or rather heliostatic) theory and planetary stations and retrogressions is, I suggest, entirely different.[12] According to the Copernican theory we are, of course, making our observations from a moving observatory. As the Earth and Mars

[11] This is often thought of as the archetypically ad hoc move (epicycles are almost synonymous with *ad hoccery*). However, the Ptolemaic move does produce an independent test (and indeed an independent confirmation) but not one that, so far as I can tell, was ever recognized by any Ptolemaist. It follows from the epicycle-deferent construction that the planet must be at the "bottom" of its epicycle and, hence, at its closest point to the Earth exactly at retrogression. But this, along with other natural assumptions, entails that the planet will be at its brightest at retrogression – a real fact that can be reasonably confirmed for some planets with the naked eye. (Of course, even had it been recognized, this test would not have been reason to continue to prefer Ptolemy over Copernicus because, as will immediately become apparent, the Copernican theory also entails, in an entirely non–ad hoc way, that the planet is at its nearest point to the Earth at retrogression.)

[12] See the treatment in Lakatos and Zahar (1976).

both proceed steadily eastward around the sun, the Earth, moving relatively quickly around its smaller orbit, will periodically overtake Mars. At the point of overtaking, although both are in fact moving consistently eastward around the Sun, Mars will naturally *appear*, as observed from the Earth, to move backwards against the background of the fixed stars. Planetary stations and retrogressions, rather than needing to be explained *via* specially tailored assumptions (having to be "put in by hand" as scientists sometimes say), drop out naturally from the heliocentric hypothesis. Copernican theory, in my view, genuinely *predicts* stations and retrogressions even though the phenomena had been known for centuries before Copernicus developed his theory. (Here I am talking about the qualitative phenomenon, not the quantitative details which, as is well known, need to a large extent to be "put in by hand" by both theories – and courtesy of multiple epicycles in Copernicus no less than in Ptolemy.[13])

The way that Copernican theory yields stations and retrogressions may, indeed, seem to be *so* direct that it challenges Duhem's thesis: doesn't the basic heliocentric hypothesis on its own, "in isolation," entail those phenomena? This is a general feature of the sort of case I am trying to characterize: the way that the confirming phenomenon "drops out" of the basic theory appears to be so direct that scientists are inclined to talk of it as a direct test of just the basic theory, in contradiction to Duhem's thesis. But we can see that, however tempting this judgment might seem (and I *am*, remember, endorsing the view that especially direct or strong support is yielded in such cases), it cannot be literally correct.

First of all, there must be assumptions linking actual planetary positions (as alleged by the theory) to our observations of them – no less so, or not much less so, with naked-eye observations as with telescopic ones. (Remember that the Flamsteed–Newton dispute revealed the inevitable existence of an assumption about the amount of refraction undergone by the light reflected from any given planet as that light enters the Earth's atmosphere.) But even laying this aside, no theory T, taken "in isolation," can deductively entail any result e if there is an assumption A that is both self-consistent and consistent with T and yet which together with T entails not-e. Therefore, in the case we are considering, if the basic Copernican theory alone entailed stations and retrogressions, then there would have to be *no possible* assumption consistent with that basic heliocentric claim that, together with it, entailed that there would be no stations or retrogressions.

[13] See, for example, Kuhn (1957).

But such possible assumptions *do* exist. Suppose, for example, that the Earth and Mars are orbiting the Sun in accordance with Copernicus' basic theory. Mars happens, though, to "sit" on an epicycle but only starts to move around on that epicycle when the Earth is overtaking Mars and does so in such a way that exactly cancels out what would otherwise be the effects of the overtaking (that is, the station and retrogression). Of course, this assumption is a monstrous one, but it is both internally consistent and consistent with the basic heliocentric view. The existence of this assumption implies that, contrary perhaps to first impressions, Duhem's thesis is not challenged in this case; the heliocentric hypothesis *alone* does not entail the phenomena (even if we lay aside the dependence on assumptions linking planetary positions with our observations of them).

However, those first impressions and the monstrousness of the auxiliary necessary to "prevent" the entailment of stations and retrogressions both reflect just how "natural" the extra assumptions are that are necessary for heliocentricism to entail the phenomena. All that needs to be assumed, in addition to the basic idea that Mars and the Earth are both orbiting the sun, is that they both do so in relatively regular ways (no sudden pirouettes and the like) and that the Earth (which has an observably smaller average period) moves relatively quickly around its smaller orbit and, hence, periodically "laps" Mars.

Let me, then, sum up this section of the chapter. It seems obvious on reflection, or so I claim, that two quite different precise ways exist for using data in science, each of which falls under the vague notion of data providing "empirical support" for a theory. Using empirical data *e* to construct a specific theory *T′* within an already accepted general framework *T* leads to a *T′* that is indeed (generally maximally) supported by *e*; but *e* will not, in such a case, supply any support at all for the underlying general theory *T*. The second and stronger type of empirical support involves a genuine test of, and therefore the possibility of real confirmation for, not just the specific theory that entails some observational result *e* but also the underlying general theory. And, as we have just been seeing, there are in turn two separate ways in which this stronger kind of support can be achieved. The "UN" or "no-double-use" rule was aimed at distinguishing general theoretical frameworks or research program that are "degenerating" from those that are "progressive." At, in other words, systematically underwriting the intuitive judgment that, when some piece of evidence *e* is predicted by some specific theory within general program *P*, but only accommodated post hoc by some specific theory within general rival program *P′*, this does not, contrary to what Kuhn seemed to suppose, balance the evidential scales – *e* continues to provide *a* reason for preferring *P* to *P′* (of course, this fact does not rule out other reasons for the opposite preference). The defenders of the rule were therefore pointing (correctly) at the importance of the "second," "stronger" unconditional type of support described earlier and (correctly) emphasizing that the conditional type of confirmation provides no support at all that "spreads" to the underlying general theory. What those who thought that they were criticising the "UN" or "no-double-use" rule were really doing was pointing out that the same manoeuvre – of using data to fix parameter values or particular theories within a given general framework – that is correctly regarded with suspicion when performed as a defensive, "degenerating" move when two general frameworks are vying for acceptance is often also used positively within general theoretical frameworks. The manoeuvre will seem positive when the general framework that is being presupposed is supported independently of the particular data being used. And it will look more positive the more such independent empirical support exists for the general framework. But, however positive the manoeuvre looks, the evidence involved does not – cannot! – supply any further support for the general framework. Instead that evidence simply (though importantly) transfers the support enjoyed by the general framework theory to the particular theory thus deduced from that evidence plus the general theory.

Mayo challenged me to be more explicit about the underlying *justification* for the two-type confirmation theory that I defend here. Well it is, I trust, clear that the justification for the conditional type (where the "no-double-use" rule *allegedly* fails) is deductive (or a close substitute): we already (we assume) have good reasons for holding some general theory; the relevant data then, *within that context*, support the specific version by deductively entailing it. As for the "stronger" "unconditional" type of confirmation, the underlying justification is exactly the same as that cited by Mayo in favour of her own approach (see next section) – a theory *T* is supported in this sense by some evidence ***e*** only if (and to the extent that) ***e*** is the outcome (positive so far as *T* is concerned) of some severe test of *T*. This, in turn – as Popper resisted recognizing – is underpinned by the intuitions that are often taken to be captured by the "No Miracles argument": it seems in some clear but (I argue[14]) elimininably intuitive sense very unlikely that a theory would survive a severe test of it if the theory were not somehow "along the right lines."

[14] See Worrall (n.d.).

4 Mayo's Alternative: Confirmation Is All about "Severe Tests"

How do these views on confirmation compare with Mayo's influential and more highly developed views? Some striking similarities certainly exist. Deborah starts, just as I do, with the "UN rule" and by emphasising the fact that the rule delivers judgments that accord with intuition in many cases; and she insists, just as I do, that the rule also seems to contradict what seem to be clearly valid intuitive judgments about support in other cases. Unlike me, however, she sees the UN rule as definitely refuted by these latter judgments and therefore as needing to be replaced, rather than, as I have argued, clarified.

Mayo's bold and challenging idea is in fact that *all* cases, both those that satisfy the UN rule and those that seem to conflict with it, are captured by one single underlying notion that is at once simple and powerful: the notion of a *severe test*. Confirmation of a theory for her *always* results from that theory's surviving a severe test. Echoing Popper, of course, she holds that hypotheses gain empirical credit only from passing genuine tests; the more severe the test, the higher the confirmation or support, if the theory passes it. This simple idea, when analysed from her own distinctive perspective, reveals – so she argues – *both* the rationale for the UN rule in the cases where it does correctly apply *and* the reason why that rule delivers incorrect judgments in other cases.

The defenders of the use-novelty account hold in effect that evidence used in the construction of a hypothesis cannot provide a genuine test of it and, hence, cannot supply genuine confirmation. Underlying their view, on Mayo's analysis, is the initially plausible-sounding claim that a severe test is one that a theory has a high probability of failing. Hence, the UN rule must, it seems, be correct because evidence *e* used in the construction of *T* cannot possibly test *T*, because there is no chance of *T*'s failing the "test" whose outcome is *e* – that outcome was instead "written into" *T*. No matter how plausible this may sound, argues Mayo, it in fact misidentifies the probability that we should be concerned to maximize so that we might maximize severity and, hence, it misidentifies the real notion of a severe test. It is easier to understand her characterization if we accentuate the negative: a *non*severe test is *not* one that has a high probability of being passed by a theory (in the limit, of course, is a test *certain* to be passed by the theory), but rather one that has a high probability of being passed by the theory, *even though the theory is false*. As she puts it, "what matters is not whether passing is assured but whether erroneous passing is" (1996, pp. 274–5).

In cases where the "no-double-use" rule delivers the correct answer (she cites "gellerized hypotheses,"[15] but would surely accept the Velikovsky case cited earlier as identical in the relevant respects), the "test" at issue was indeed nonsevere: the modified Velikovsky theory would have a good chance of passing the test of no records of suitable cataclysms in culture C even though that theory were false. On the other hand, in the cases where the no-double-use rule goes wrong, such as her SAT score example, although admittedly there was no chance of the "hypothesis" that the average score of her class is 1121 not passing the "test" arrived at by adding the N individual scores and dividing by N, the "test" was nonetheless genuine and severe, indeed maximally severe, because there would have been no chance of the "hypothesis" passing the test *if it were false.* Similarly in standard statistical estimation cases, such as the one cited by Howson and developed in much more detail by Mayo, assuming that we have some reason to think that the general model being applied really does apply to the real situation, then using the observed result of k out of n balls drawn being white to construct the hypothesis that the proportion of white balls overall in the urn is $k/n \pm \varepsilon$ (where ε is calculated as a function of n and the chosen significance level by standard confidence-interval techniques) does *not* preclude the sample relative frequency (e) of k/n red balls being good evidence for our hypothesis. Even though e was thus used in the construction of h, e still constitutes a severe test of h because there was little chance of h passing the test resulting in e if it were false.

Despite being a colleague of Nancy Cartwright, there are few bigger fans of unity than I. And Mayo here offers a unified alternative to my "two kinds of confirmation" view – two kinds of confirmation do not exist, only one: that supplied by a theory's surviving a severe test. It would seem churlish of me to turn this offer down and thus reject the call to join the "error paradigm." Moreover, so Deborah assures me, were I to join then I could avail myself of precise characterisations of notions such as that of an empirical prediction "falling out" of a theory, which are important to my view of confirmation but

[15] Uri Geller asserted (indeed no doubt still asserts) that he has genuine psychokinetic powers; when, unbeknownst to him, professional magicians controlled the situation in which he was to exhibit these powers, for example, by bending spoons at a distance, he proved impotent; however, Geller responded by claiming that his "special" powers were very delicate and had been affected by the presence of skeptics in the audience. Obviously he could only identify whether or not skepticism was playing this obstructive role post hoc: if he was able to bend spoons by the "power of pure thought" then no skeptics were around (and also, of course, no hindrance to his employing standard magicians' tricks); if he was unable to bend them "supernaturally" then clearly there *was* skepticism in the air.

are left as merely suggestive notions within it (though I hope with clear-cut illustrations from particular scientific cases).

Despite these enticements, I must turn this kind offer down. I do so for two interrelated reasons:

1. There seem to be a number of unclarities in or outright significant difficulties with Mayo's position; and more fundamentally
2. It just seems to be true – and plainly true – that, as I explained in the preceding section, there are two quite separate uses of evidence within science: using evidence in the construction of a theory is a quite different matter from using evidence to test it by "probing for errors;" Mayo's attempt to construct a one-size-fits-all account where all (positive) uses of evidence in science are regarded as the passing of a severe test is itself an error. (Einstein is reported as having said that physics should be as simple as possible, but not more so! The same surely applies to meta-science.)[16]

I begin with the already much-discussed SAT score example. As remarked, it does seem extraordinary to call the assertion arrived at about the average SAT score of Mayo's students a "hypothesis" and at least equally extraordinary to call the process of adding the individual scores and dividing by the number of students a "test" of that claim. Of course, had someone made a "bold conjecture" about the average score, then one might talk of the systematic process of working out the real average as a test of that conjecture. But boldly conjecturing would clearly be a silly way to proceed in this case, and, as already remarked, not one that would ever be used in more realistic cases in science. The process of adding the individual scores and dividing by the number of students surely is a *demonstration that* the average score is 1121, not a "*test*" of the "hypothesis" that this is the average score.[17] We construct the "theory" by deducing it from data (indeed the "theory" just encapsulates a feature of the data).

More important, because we all agree that the evidence here is conclusive for the "hypothesis" and it might be felt that it does not really matter

[16] The problems involved in reason 1 are in fact, as I shall explain, all produced by the fact that reason 2 is true.

[17] In fact this and some of the other cases that Mayo analyzes – such as the identification of the car that hit her own car's fender or the technique of "genetic fingerprinting" – seem altogether more naturally categorized as *applications* of already accepted theories (or "theories" in the case of the average SAT score) to particular circumstances rather than as any sort of empirical support for theories. We *apply* our theories of genetics to work out the probability that the match we have observed between the crime scene blood and that of the defendant would have occurred if he or she were innocent.

how we choose to express it, the case seems to me to highlight a problem with applying Mayo's central justification for all confirmation judgments. In the circumstances (and assuming that both the data on the individual students and the arithmetic have been carefully checked) there is *no* chance that the average SAT score is *not* 1121. If, as seems natural, this claim is interpreted as one about a conditional probability – namely, p (T passes the test with outcome e/T is false) $= 0$ – we are being asked to make sense of a conditional probability where the conditioning event (the claim's being false) has probability zero. Indeed we are asked not only to make sense of it but to agree that the conditional probability at issue is itself zero. It is well known, however, that – at any rate in all standard systems – $p(A/B)$ is not defined when $p(B) = 0$. Perhaps we are meant to operate with some more "intuitive" sense of chance and probability in this context. But I confess that I have examined my intuitions minutely and still have no idea what it might mean in this case to imagine that the average score is *not* 1121, when the individual scores have been added and divided by N and the result *is* 1121!

In correspondence Mayo tells me that I *should have* such intuitions "because the next time you set out to use your estimation tool it may NOT BE 1121." Well, maybe she can give me more hints on how to develop better intuitions, but this one certainly doesn't work for me: surely – again short of making some trivial arithmetical error – applying the "estimation tool" would just *have to* yield 1121 again with this particular group of students; if you were to arrive at any other figure you simply would not be taking the average. And if she means that the average score might not be 1121 for some *different* group of students then of course this is (trivially) true but whatever number you arrived at (assuming again that you arrived at it correctly without trivial error) would still be the group average for that new group!

It could, perhaps, be argued that this is simply a problem for this admittedly extreme case. But there are other, related problems with Mayo's account of severity and the associated intuitive probability judgments that underpin it that surface in other more standard, scientific cases.

One way that she likes now to put the partial connection, and partial disconnection, between "no double use" and severity is something like this:

It would seem that if hypothesis H is use-constructed then a successful fit (between H and [data] x) is assured, *no matter what* (and, hence, that in the case of use-construction the data always represent a nonsevere test). However, the "no matter what" here may refer to two quite different conditions:

1. no matter what the *data are*, or
2. no matter whether H is true or false.

In cases where the “no-double-use” rule gives the *wrong* answer, condition 1 is true alright (as always with double-use cases), *but condition 2 is false.* In cases in which “no double use” *correctly* applies, on the other hand, condition 2 is also true – that is, *both* conditions 1 and 2 must be met if x is to *fail* to represent a severe test. The no-double-users have mistakenly held that if condition 1 alone is met then the test is automatically nonsevere.

But this way of putting things, while sounding very neat, in fact leads simply to a new way of putting the previous objection. I have already pointed out that it seems impossible to me to make sense of condition 2, at least in the SAT score case (and, as I will shortly argue, more generally). But problems exist with condition 1, too – problems that apply across the board. The collective-amnesia–ized version of Velikovsky, V', is “use-constructed” from the data e concerning the cultures from which we have or have not appropriate records of suitably dated “catastrophes.” This is surely a case where the “no-double-use” idea ought to apply in some way or another: V', because of its method of construction, is not really tested by data e and, hence, is not supported by it – not, at least, in any sense that makes it rationally more credible. Does Mayo’s account deliver this judgment of nonseverity? In particular, does her condition 1 apply? That is, is it true that a successful fit between V' and data x is “assured no matter what the data are?”

Well if the data were anything other than they in fact are, say they were e' (that is, Velikovksy faced a different list of otherwise record-keeping cultures who have left no records of suitable cataclysms), then V' (that is, the particular version of the general Velikovskian theory actually constructed from the real data e) would of course conflict with this supposed data e': some cultures alleged by V' to have suffered from collective amnesia would have records of catastrophes and/or some cultures having no such records would not be alleged by V' to have suffered collective amnesia. Had the data been different then the Velikovskian would not have been proposing V' but instead some rival V'' – a specific version of the same general “cometary” theory that evaluated the “collective-amnesia parameter” differently.

It seems, then, to be straightforwardly untrue that a successful fit between V' and e is assured no matter what e is. Nor can we rescue the situation by allowing (as of course Mayo explicitly and elaborately does) grades of severity and, hence, the intuitive “probabilities” discussed earlier. It again makes no sense to me to say that the test of V' that turns out to have outcome e is nonsevere because there is a high probability of V' fitting the data no matter what those data are. Within the context of the general Velikovsky theory with free “collective-amnesia parameter,” we can in effect derive the biconditional V' if and only if e: that is, V' would definitely not have fit the data had the data been different than they in fact were.

It is not the successful fit of a particular hypothesis with the data that is guaranteed in these sorts of case but rather the fit of *some particular* hypothesis developed within the "given" underlying general framework. We again need to recognize that, as my account entails, two separate issues exist – the "confirmation" of a theory within a general framework (*e* maximally confirms *V′* given *e and given V*) and "confirmation" of a specific theory within a general framework that "spreads" to the underlying general theory. This condition, as pointed out earlier, is not satisfied in the Velikovsky case and, hence, *e* gives no "unconditional" support to *V′* of the sort that would spread to the underlying *V*, - exactly because the general theory places no constraints on the relevant parameter, the value of which can be "read off" whatever the data turn out to be.

Mayo may reply that her account does yield the result that the evidence *e* does not support the general Velikovskian theory, *V*, because the latter is not tested (or, as she often says, "probed") by that evidence. Of course I agree with this judgment, but that is not the problem – her account of the support lent to *V′* is at issue. We surely want to say that *e* provides no good reason to take *V′* seriously. If she were to deliver this judgment directly, it would have to be that *e* fails to be any sort of severe test of *V′*, which requires that conditions 1 and 2 of her latest formulation of (non)severity be satisfied; but, as we just saw, condition 1 is *not* in fact satisfied. If Mayo were tempted, in response, to rule that a specific theory like *V′* is only "probed" by a test with outcome *e* if that same test also probes (severely tests) its general version *V*, then it would mean that she had failed to capture the alleged exceptions to the UN rule. These (apparent) exceptions, as explained earlier, all involve deductions from the phenomena that all presuppose, and therefore cannot "probe," the underlying theory. And the attempt to capture these alleged exceptions is of course an important part of the motivation for her overall account. Adding the SAT scores of her *N* logic students and dividing by *N* does not probe the underlying definition of an average score! Adams and Leverrier's use of the anomalous data from Uranus to construct a version of Newton's theory (complete with a postulated "new" planet) did not probe the underlying Newtonian theory (three laws plus the principle of universal gravitation) but, on the contrary, presupposed it. It is again surely clear why her account meets these difficulties. We are dealing, in accordance with my own account, with *two quite different uses of evidence* relative to theories; her attempt to cover these two different cases with one set of criteria is bound to fail.

To see that these difficulties for Mayo's account are not simply artifacts of the strange, clearly pseudoscientific case of Velikovsky's theory, nor are they restricted to problems with her first condition for nonseverity, let us return to the case of the wave theory that I discussed earlier. This example, although deliberately a very simple one, nonetheless exemplifies an important and recurrent pattern of reasoning in real science. The simplicity of the case allows us to concentrate on the pattern of reasoning and not become sidetracked by scientific details and complexities.

Remember, the case involves the general wave theory of light, call it W. Theory W leaves the wavelengths of light from particular monochromatic sources as free parameters; it does, however, entail (a series of) functional relationships between such wavelengths and experimentally measurable quantities. In particular, subject to a couple of idealisations (which nonetheless clearly approximate the real situation), W implies that, in the case of the two-slit experiment, the (observable) distance X from the fringe at the center of the pattern to the first fringe on either side is related to (theoretical) wavelength λ via the equation $X/(X^2 + D^2)^{1/2} = \lambda/d$ (where d is the distance between the two slits and D the distance from the two-slit screen to the observation screen – both observable quantities). It follows analytically that $\lambda = dX/(X^2 + D^2)^{1/2}$. But all the terms on the right-hand side of this last equation are measurable. Hence, particular observed values of these terms, call their conjunction e, will determine the wavelength (of course within some small margin of experimental error) and so determine the more specific theory W', with the parameter that had been free in W now given a definite value – again within a margin of error.

This scenario is a paradigmatic case of "deduction from the phenomena" – exactly the sort of case, so critics of the UN rule have alleged, in which that rule clashes with educated intuition. We do want to say that e "supports" W' in some quite strong sense; and yet clearly e was used in the construction of W' and, hence, W' was guaranteed to pass the "test" whose outcome was e. Mayo's claim here is that, whenever "no double use" goes astray, it is because condition 2 has been ignored. A test of theory T may be maximally severe even if T is guaranteed to pass it, so long as T is not guaranteed to pass it *even if it is false*. (Remember: "what matters is not whether passing is assured but whether *erroneous* passing is.") But in fact Mayo's condition 2 for *non*severity is *met* here: whether or not W' is true the fit with e is assured, because the value of λ specified by W' has been calculated precisely to yield e. It is true that, for exactly the same reasons as we saw in the case of Velikovsky, it can be argued that Mayo's condition 1 fails to hold in this case. In fact an ambiguity exists over what "it" is in the condition (for nonseverity, remember) – that "it" would have passed the test concerned whatever that test's outcome. The particular W' that was in fact constructed from data e would certainly *not* have passed the test of measuring the fringe distances and so forth in the two-slit experiment with sodium light had those measurements produced results other than those expressed in e. What was bound to pass again *is some version or other* of W, with some value or other for the wavelength of sodium light. It might be argued, therefore, on Mayo's behalf that because it is not true that both conditions for nonseverity hold in this case, the test may be regarded as at least somewhat severe. But clearly what Mayo intended was that condition 1 *should* in fact hold in "use-constructed" cases and that it is the failure of condition 2 to hold in certain particular cases (despite condition 1 holding) that explains why the UN rule delivers incorrect verdicts in those cases.

It seems, therefore, to be at best unclear whether Mayo's scheme, when analysed precisely, can explain the judgment that e does at least something positive concerning the credentials of W' in the case we are considering. On the other hand, this judgment *is* captured by my account: e definitely supports W' in the conditional sense in that it establishes W' as *the* representative of the general theory W if that theory is to work at all; hence, one might say, the construction transfers to W' all the unconditional empirical support that W had already accrued (and in this case there was plenty of such support).

Mayo has claimed (personal correspondence) that this analysis entirely misrepresents her real view. She would *not* in fact want to say in this wave-theory case that e tests W', because it does not "probe the underlying [W]." Of course this is indeed true (and importantly true), though it is unclear how this relates to the issue of whether W' and e satisfy condition 2 of her latest account. But even supposing we go along with this view of what her account entails here, how then does that account deliver the (conditional but nonetheless positive) verdict concerning the support that e lends to W' that intuition does seem to require? Moreover, this interpretation of her account takes us back to the problem mentioned earlier in connection with Velikovsky – namely, that it then seems hard to understand how it delivers the judgments that she highlights as "refutations" of UN. How, if "probing the underlying theory" is also required for a test of a specific theory to be severe, can it be that the data in the SAT score case severely test the hypothesis about the average score for her class? Or that estimates of some parameter (such as the proportion of red to white balls in Howson's urn case) arrived at *via* standard statistical techniques severely test the hypothesis about that parameter? In neither case is the underlying theory "probed," but it is instead taken for granted (indeed in the SAT course case no option exists but to take the underlying theory for granted because it is analytic). No result that you could get from averaging SAT scores could challenge the definition of an average; no sample relative frequency of red and white balls could challenge the idea that the urn contains some unknown but fixed ratio of such balls and that the draws are independent.

If the Mayo account could be defended at all here, then it would have to be, so it seems to me, by reinterpreting her second condition for nonseverity. Her account would have to be understood as saying that a test of *T* is nonsevere if the test's outcome is bound to fail to refute *T* (condition 1) *and* if the general theory underlying *T* is not itself empirically supported by *other* tests. But this would be in effect just to rewrite my own account in something approximating Mayo's terms. Moreover, as I have suggested earlier more than once, by thus writing my analysis into one account of severe versus nonsevere tests, the important and qualitative difference between the two uses of evidence as related to theories would be obscured.

The attempt to see everything in terms of severe testing, and probing for error, seems to lead either to error or at best to a confusing reformulation of the view that I defended. It surely is just the case that science makes use of two separate roles for evidence: a role in the construction of theories ("observation as theory-development by other means" as I believe van Fraassen says somewhere) and a role in testing theories, in probing them for errors. The latter is of course a vastly important use of evidence in science but it is not, as Mayo has tried to suggest, everything.

References

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

French, A. (1971), *Newtonian Mechanics*, MIT Press, Cambridge, MA.

Hitchcock, C., and Sober, E. (2004), "Prediction Versus Accommodation and the Risk of Overfitting," *British Journal for the Philosophy of Science*, 55: 1–34.

Howson, C. (1990), "Fitting Theory to the Facts: Probably Not Such a Bad Idea After All," in C. Wade Savage (ed.), *Scientific Theories*, University of Minnesota Press, Minneapolis.

Kuhn, T.S. (1957), *The Copernican Revolution*, Princeton University Press, Princeton.

Kuhn, T.S. (1962), *The Structure of Scientific Revolutions* (2nd enlarged ed., 1970), University of Chicago Press, Chicago.

Kuhn, T.S,. (1977), *The Essential Tension*, University of Chicago Press, Chicago.

Lakatos, I., and Zahar, E.G. (1976) "Why Did Copernicus's Programme Supersede Ptolemy's," Chapter 4 in I. Lakatos (ed.), *The Methodology of Scientific Research Programmes*, Cambridge University Press (reprinted 1978).

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

Worrall, J. (2000), “The Scope, Limits and Distinctiveness of the Method of “Deduction from the Phenomena”: Some Lessons from Newton’s “Demonstrations” in Optics,” *British Journal for the Philosophy of Science*, 51: 45–80.

Worrall, J. (2002), “New Evidence for Old,” in P. Gardenførs et al. (eds.), *In the Scope of Logic, Methodology and Philosophy of Science*, Kluwer, Dordrecht.

Worrall, J. (2003), “Normal Science and Dogmatism, Paradigms and Progress: Kuhn “versus” Popper and Lakatos,” in T. Nickles (ed.), *Thomas Kuhn*, Cambridge University Press, Cambridge.

Worrall, J. (2006), “Theory Confirmation and History,” in C. Cheyne and J. Worrall (eds.), *Rationality and Reality*, Springer, Dordrecht.

Worrall, J. (n.d.) “Miracles, Pessimism and Scientific Realism,” forthcoming.

An Ad Hoc Save of a Theory of Adhocness? Exchanges with John Worrall

Deborah G. Mayo

In large part, the development of my concept of severity arose to deal with long-standing debates in philosophy of science about whether to require or prefer (and even how to define) novel evidence (Musgrave, 1974, 1989; Worrall 1989). Worrall's contribution represents the latest twists on our long-running exchange on the issue of novel evidence, beginning approximately twenty years ago; discussions with Musgrave around that time were also pivotal to my account.[1] I consider the following questions:

1. *Experimental Reasoning and Reliability*: Do distinct uses of data in science require distinct accounts of evidence, inference, or testing?
2. *Objectivity and Rationality*: Is it unscientific (ad hoc, degenerating) to use data in both constructing and testing hypotheses? Is double counting problematic only because and only when it leads to unreliable methods?
3. *Metaphilosophy*: How should we treat counterexamples in philosophical arguments?

I have argued that the actual rationale underlying preferring or requiring novel evidence is the intuition that it is too easy to arrive at an accordance between nonnovel data and a hypothesis (or model) even if H is false: in short the underlying rationale for requiring novelty is severity. Various impediments to severity do correlate with the double use of data, but this correlation is imperfect. As I put it in Mayo (1996): "Novelty and severity do not always go hand in hand: there are novel tests that are not severe and severe tests that are not novel. As such, criteria for good tests that are couched in terms of novelty wind up being either too weak or too strong, countenancing poor tests and condemning excellent ones" (p. 253). This

[1] Mayo (1991; 1996, p. xv).

malady is suffered by the version of novelty championed by Worrall, or so I argue. His chapter turns us directly to the problem of metamethodology with respect to a principle much debated in both philosophy of science and statistical practice.

The UN requirement – or, as he playfully calls it, the UN Charter – is this:

1.1 Use-Novelty Requirement (UN Charter): For data **x** to support hypothesis *H* (or for **x** to be a good test of *H*), *H* should not only agree with or "fit" the evidence **x**, **x** must itself *not have been used* in *H*'s construction.

For example, if we find data **x** anomalous for a theory or model, and we use **x** to arrive at a hypothesized explanation for the anomaly *H*(**x**), it would violate the UN Charter to also regard **x** as evidence for *H*(**x**). Much as with the rationale for varying the evidence (see also Chapter 2, Exchanges with Chalmers), use-novelty matters just to the extent that its violation inhibits or alters the reliability or stringency of the test in question. We can write the severity requirement in parallel to the UN Charter:

1.2 Severity Requirement: For data **x** to support hypothesis *H* (or for **x** to be a good test of *H*), *H* should not only agree with or "fit" the evidence **x**, *H* must have passed a stringent or severe test with **x**.

If UN violations alter a test's probativeness for the inference in question, the severity assessment must be adjusted accordingly.

However, as I have argued, some cases of "use-constructed" hypotheses succeed in being well tested by the same data used in their construction. To allude to a case discussed in Mayo (1996, p. 284), an apparent anomaly for the General Theory of Relativity (GTR) from the 1919 Sobral eclipse results was shown to be caused by a mirror distortion from the sun's heat. Although the eclipse data were used both to arrive at and to test *A*(**x**) – the hypothesized mirror distortion explanation – *A*(**x**) passed severely because it was constructed by a reliable or stringent rule. Because Worrall agrees with the severity goal, these cases stand as counterexamples to the UN requirement. Worrall's chapter discusses his recent attempts to accommodate these anomalies; surprisingly, Worrall is prepared to substantially adjust his account of confirmation to do so.

In particular, he allows my counterexamples to stand, but regards them as involving a distinct kind of support or corroboration, a kind he has developed to accomodate UN violations. For instance, if *A*(**x**) is use-constructed to account for **x** which is anomalous for theory *T*, then the inference to *A*(**x**) gets what Worrall calls "conditional support." By this he means that *A*(**x**) is legitimately inferred only conditional on already assuming *T*, the theory to be "saved." So for Worrall the UN requirement still stands as necessary (and sufficient) for full-bodied support, but data **x** may still count as evidence for use-constructed *A*(**x**) so long as we add "conditional on accepting an overarching theory *T*."

But I do not see how Worrall's attempt to save the UN Charter can adequately accommodate the counterexamples I have raised. It is false to suppose it is necessarily (or even commonly) the case that use-constructed hypotheses assume the truth of some large-scale theory, whether in the case of blocking an anomaly for theory *T* or in any of the other use-constructions I have delineated (Mayo, 1996, 2008). Certainly using the eclipse data to pinpoint the source of the GTR anomaly did not involve assuming the truth of GTR. In general, a use-constructed save of theory *T* takes the form of a hypothesis designed to block the alleged anomaly for *T*.

1.3 A use-constructed block of an anomaly for *T*:

A(**x**): the anomalous data **x** are due to factor *F*, not the falsity of *T*, or
A(**x**) explains why the data **x** did not accord with the predictions from *T*.

By lumping together all cases that follow this logical pattern, Worrall's account lacks the machinery to distinguish reliable from unreliable use-constructions. As a result, I argue, Worrall's account comes up short when it comes to real experimental inferences:

> Any philosophy of experimental testing adequate to real experiments must come to grips with the fact that the relationship between theory and experiment is not direct but is mediated along the lines of the hierarchy of models and theories.... At various stages of filling in the links, it is standard to utilize the same data to arrive at as well as warrant hypotheses.... As a matter of course, then, the inferences involved violate even the best construals of the novelty requirement. (Mayo, 1996, p. 253)

By maintaining that all such use-constructions are conditional on already assuming the truth of some overarching theory or "research program," Worrall's philosophy is redolent of the image of scientists as locked into "theory-laden" paradigms (Lakatos, Kuhn). Conversely, by regarding UN as sufficient for warranted inference, Worrall overlooks the fact "that there is as much opportunity for unreliability to arise in reporting or interpreting (novel) results given knowledge of theoretical predictions as there is... in arriving at hypotheses given knowledge of (non-novel) results" (ibid, p. 254). By recognizing that what matters is the overall severity with which a claim may be inferred, we have a desideratum that allows us to discriminate, on a case-by-case basis, whether UN violations matter and, if so, how we might correct for them.

I begin by discussing Worrall's treatment of use-construction in blocking anomalies and then turn to some confusions and errors that lead Worrall to forfeit a discriminatory tool that would seem to fulfill the (Popperian) testing philosophy that he himself endorses.

2 Use-Constructing in Blocking Anomalies: Must All of *T* Be Assumed?

We are all familiar with a variety of "rigging" procedures so as to accommodate data while protecting pet hypotheses rather than subjecting them to scrutiny. One of Worrall's favorite examples is Velikovsky's method for use-constructing hypotheses to save his theory when confronted with any anomalous data **x**:

> The lack of records in cultures $C_1, \ldots, C_n$ and their (arguable) presence in $C'_1, \ldots, C'_m$ gives very good reason for holding the specific collective-amnesia version of Velikovsky's theory that he proposed *if* you already hold Velikovsky's general theory, *but* (and this is where the initial UN intuitions were aimed) those data give you absolutely no reason at all for holding that general theory in the first place. (Worrall, this chapter, pp. 136–7)

But why suppose that the inference to blocking a *T* anomaly assumes all of *T*? We know that, even if it is warranted to deny there is evidence against *T*, this fact alone would not provide evidence *for T*, and there is no reason to saddle every use-constructed save with committing so flagrant a fallacy (circularity). Obviously any method that assumes *T* in order to save *T* is minimally severe, but it is false to suppose that in use-constructing $A(\mathbf{x})$, *T* is assumed. It is not even clear why accepting Velikovsky means that any lack of records counts as evidence for the amnesia hypothesis, unless it is given that no other explanation can exist for the anomaly, as I take it Worrall does (note 10, p. 136). But are we to always imagine this? I put this aside. Even a proponent of Velikovsky's dodge could thwart Worrall's charge as follows.

> *V-dodger*: I am not claiming that lack of records of the cataclysms described in my theory *T* is itself evidence for *T* (other records and considerations provide that); I am simply saying that I have a perfectly sound excuse, $A(\mathbf{x})$, for discounting the apparent anomaly for my theory.

Despite the ability to escape Worrall's charge, the flaw in the V-dodger's inference seems intuitively obvious. The severity account simply provides some systematic tools for the less obvious cases. We are directed to consider the use-construction rule R leading from $\mathbf{x}$ to the inference $A(\mathbf{x})$ and the associated threats of error that could render the inference unwarranted. Here, we can characterize the rule R in something like the following manner.

Rule *R* (Velikovsky's scotoma dodge): For each possible set of data $\mathbf{x}^i$ indicating that culture C^i has no records of the appropriate cataclysmic events, infer $A^i(\mathbf{x}^i)$: culture C^i had amnesia with regard to these events.

The blocking hypothesis $A^i(\mathbf{x}^i)$ is use-constructed to fit data $\mathbf{x}^i$ to save Velikovsky from anomaly.

Clearly, rule R prevents any observed anomaly of this form to threaten Velikovsky's theory, even if the culture in question had not suffered amnesia in the least. If one wanted to put this probabilistically, the probability of outputting a Velikovsky dodge in the face of anomaly is maximal, even if the *amnesia explanation* is false (a case of "gellerization") – therefore, severity is minimal. Because rule R scarcely guards against the threat of erroneously explaining away anomalies, we would say *of any particular output of rule R* that the observed fit fails to provide evidence for the truth of $A(\mathbf{x}_0^i)$.

Notice that one need not rule out legitimately finding evidence that a given culture had failed to record events that actually occurred, whether due to memory lapses, sloppy records, or perhaps enforced by political will. For example, we could discern that all the textbooks in a given era were rewritten to expunge a given event, whose occurrence we can independently check. But with Velikovky's rule R there is no chance that an erroneous attribution of scotoma (collective amnesia) would be detected; nothing has been done that could have revealed this fact, at least by dint of applying rule R.

Although we condemn inferences from tests that suffer from a low probability of uncovering errors, it is useful to have what I call "canonical errors" that stand as extreme cases for comparison (cases of zero severity). Velikovsky's case gives one. We utterly discredit any inference to $A(\mathbf{x})$ resulting from Velikovsky's use-construction rule, as seems proper. It is surprising, then, that Worrall's account appears to construe Velikovsky's gambit as no worse off than any other use-constructed saves, including those that we would consider altogether warranted.

The detailed data analysis of eclipse plates in 1919 warranted the inference that "the results of these (Sobral Astrographic) plates are due to systematic distortion by the sun and not to the deflection of light" (Mayo, 1996, p. 284). To warrant this explanation is to successfully block an interpretation of those data as anomalous for GTR. In Worrall's account, however, all use-constructed saves of theory *T* are conditional on assuming *T*; the only way they can avoid being treated identically to the case of Velikovsky's dodge is if *other*, independent support arises for accepting *T*.

As a matter of fact, however, the data-analytic methods, well-known even in 1919, did not assume the underlying theory, GTR, nor is it correct to imagine Eddington arguing that, provided you accept GTR, then the mirror distortion due to the Sun's heat explains why the 1919 Sobral eclipse results were in conflict with GTR's predicted deflection (and in agreement with the Newtonian prediction). GTR does not speak about mirror distortions. Nor were even the staunchest Newtonians unable to agree (not that it was immediately obvious) that the detailed data analysis showed that unequal expansion of the mirror caused the distortion. It was clear the plates, on which the purported GTR anomaly rested, were ruined; accepting GTR had nothing to do with it. Nor could one point to GTR's enjoying more independent support than Newton at the time – quite the opposite. (Two data sets from the same eclipse afforded highly imprecise accordance with GTR, whereas Newton enjoyed vast support.) Nor would it make sense to suppose that vouchsafing the mirror distortion depended on waiting decades until GTR was warranted, as Worrall would seem to require.

Thus, I remain perplexed by Worrall's claim that we need "to recognise just how conditional (and *ineliminably* conditional) the support at issue is in all these cases" (this volume, p. 135). By this, he means not that there are assumptions – because that is always true, and Worrall is quite clear he does not wish to label all cases as giving merely conditional support. He means, rather, that the entire underlying theory is assumed. We have seen this to be false.

My goal (e.g., in Mayo, 1996, sec. 8.6) was to illustrate these counterexamples to the UN requirement, at several stages of testing:

> The arguments and counterarguments [from 1919 to ∼1921] on both sides involved violating UN. What made the debate possible, and finally resolvable, was that all . . . were held to shared criteria for acceptable and unacceptable use-constructions. It was acceptable to use any evidence to construct and test a hypothesis . . . so long as it could be shown that the argument procedure was reliable or severe. (p. 289)

Although the inferences, on both sides of the debate, strictly violated UN, they were deliberately constrained to reflect what is correct, at least approximately, regarding the cause of the anomalous data.

These kinds of cases are what led me to abandon the UN Charter, and Worrall has yet to address them. Here the "same" data are used both to identify and to test the source of such things as a mirror distortion, a plane crash, skewed data, a DNA match, and so on – without threats from uncertain background theories ("clean tests"). In statistical contexts, the stringency of such rules may be quantitatively argued:

A Stringent Use-Construction Rule (*R*-*α*): The probability is very small, $1 - \alpha$, that rule R would output $H(\mathbf{x})$ unless $H(\mathbf{x})$ were true or approximately true of the procedure generating data $\mathbf{x}$. (Mayo, 1996, p. 276)

Once the construction rule is applied and a particular $H(\mathbf{x}_0)$ is in front of us, we evaluate the severity with which $H(\mathbf{x}_0)$ has passed by considering the stringency of the rule R by which it was constructed, taking into account the particular data achieved. What matters is not whether H was deliberately constructed to accommodate $\mathbf{x}$; what matters is how well the data, together with background information, rule out ways in which an inference to H can be in error.

2.1 Deducing a Version (or Instantiation) of a Theory

At several junctures, it appears that Worrall is taking as the exemplar of a UN violation "using observational data as a premise in the deduction of some particular version of a theory" (p. 131) so that there is virtually no threat of error. True, whenever one is in the context wherein all of the givens of Worrall's inference to "the representative or variant of the theory" are met, we have before us a maximally severe use-construction rule (α would equal 1). We can agree with his claim that "[w]e do want to say that $\mathbf{x}$ supports $T(\mathbf{x})$ in some quite strong sense," (see p. 151), where $T(\mathbf{x})$ is what he regards as the variant of theory T that would be instantiated from the data $\mathbf{x}$. Confronted with a particular $T(\mathbf{x}_0)$, it would receive maximal support – provided this is understood as inferring that $T(\mathbf{x}_0)$ is the variant of T that would result if T were accepted and $\mathbf{x}_0$ observed. Instantiating for the wave theory, W, Worrall asserts: "$\mathbf{x}_0$ definitely supports $W(\mathbf{x}_0)$ in the conditional sense in that it establishes $W(\mathbf{x}_0)$ as *the* representative of the general theory W if that theory is to work at all" (replacing e with $\mathbf{x}_0$).[2] (p. 152) Although this

[2] An example might be to take the results from one of the GTR experiments, fix the parameter of the Brans-Dicke theory, and infer something like: if one were to hold the B-D theory, then the adjustable constant would have to be such-and-such value, for example, $q = 500$. (See Chapter 1, Section 5.2.)

inference is not especially interesting, and I certainly did not have this in mind in waging the counterexamples for the UN Charter, handling them presents no difficulty. If the assumptions of the data are met, the "inference" to the instantiation or application of theory *T* is nearly tautological.

The question is why Worrall would take this activity as his exemplar for use-constructed inferences in science. Certainly I would never have bothered about it if that was the sort of example on which the debate turned. Nor would there be a long-running debate in methodological practice over when to disallow or make adjustments because of UN violations and why. Yet, by logical fiat – construing all UN violations as virtually error-free inferences that aspire to do no more than report a specific variant of a theory that would fit observed data – the debate is settled, if entirely trivialized. If philosophers of science are to have anything useful to say about such actual methodological debates, the first rule of order might be to avoid interpreting them so that they may be settled by a logical wand.

Worrall claims to have given us good reasons for accepting his account of confirmation in the face of anomalies – where the anomalies are counterexamples to his view that UN is necessary for full-bodied confirmation. We might concur, in a bit of teasing reflexivity, that Worrall has given reason to support his handling of anomalies if you already hold his account of conditional support! But I doubt he would welcome such self-affirmation as redounding to his credit. This point takes me to a cluster of issues I place under "metaphilosophy."

3 Metaphilosophy: The Philosophical Role of Counterexamples

To a large extent, "the dispute between those who do and those who do not accept some version of the novelty principle emerges as a dispute about whether severity – or, more generally, error characteristics of a testing process – matters" (Mayo, 1996, p. 254). If it is assumed that whether *H* is warranted by evidence is just a function of statements of evidence and hypotheses, then it is irrelevant how hypotheses are constructed or selected for testing (I call these evidential-relation accounts). What then about the disagreement even among philosophers who endorse something like the severity requirement (as in the case of Worrall)? Here the source of disagreement is less obvious, and is often hidden: to dig it up and bring it to the surface requires appealing to the philosopher's toolkit of counterexamples and logical analysis. However, "philosopher's examples" are anything but typical, so one needs to be careful not to take them out of their intended context – as counterexamples!

3.1 Counterexamples Should Not Be Considered Typical Examples: The SAT Test

Now Worrall agrees with the general severity rationale: "the underlying justification is exactly the same as that cited by Mayo in favour of her own approach . . . a theory *T* is supported in this [strong] sense by some evidence ***e*** only if (and to the extent that) ***e*** is the outcome (positive so far as *T* is concerned) of some severe test of *T*" (Worrall, this volume, p. 144). We concur that, for a passing result to count as severe, it must, first of all, *be a passing result*; that is, the data must fit or accord with hypothesis *H* (where *H* can be any claim). Although Worrall often states this fit requirement as entailment, he allows that statistical fits are also to be covered. The key difference regards *what more* is required to warrant the inference to *H*. Should it be Worrall's UN criterion, or my severity criterion?

To argue for the latter, my task is to show how UN could be violated while intuitively severity is satisfied. In this I turned to the usual weapon of the philosopher: counterexamples. Observe what happens in cases where it is intuitively, and blatantly, obvious that a use-constructed hypothesis is warranted: the method that uses the data to output $H(\mathbf{x})$ is constrained in such a way that $H(\mathbf{x})$ is a product of what is truly the case in bringing about data $\mathbf{x}$. Worrall and like-minded use-novelists often talk as if an accordance between data and hypothesis can be explained in three ways: it is due to (1) chance, (2) the "blueprint of the universe" (i.e., truth or approximate truth of *H*), or (3) the ingenuity of the constructor (Worrall, 1989, p. 155). That hypotheses can be use-constructed reliably is precisely what is overlooked. In my attempts to lead them to the "aha" moment, I – following the philosopher's craft – sought extreme cases that show how use-constructed hypotheses can pass with high or even maximal severity; hence, the highly artificial example of using the data on the SAT scores to arrive at the mean SAT score. As I made clear, "the extreme represented by my SAT example was just intended to set the mood for generating counterexamples" (Mayo, 1996, p. 272), after which I turn to several realistic examples. Worrall (who is not alone) focuses on the former and gives little or no attention to the latter, realistic cases.

Ironically, it was Musgrave's reaction long ago to such flagrant cases that convinced me the Popperians had erred in this manner: "An older debt recalled in developing the key concept of severe tests is to Alan Musgrave" (Mayo, 1996, p. xv). Actually, as Musgrave reminds me, the example that convinced him was the incident that first convinced me that UN is not necessary for a good test: using data on the dent in my Camaro to hunt for a car with a tail fin that would match the dent, to infer that "it is practically impossible for the dent to have the features it has unless it was created by a specific type of car tail fin" (p. 276). The point is that counterexamples serve as this kind of tool in philosophy, and no one would think the user of the counterexamples intended them as typical examples. Yet some charge that I must be regarding the SAT averaging as representative of scientific hypotheses, forgetting that it arises only in the service of getting past an apparent blind spot.

We should clear up a problem Worrall has with the probabilistic statement we make. He considers the example of deducing $H(\mathbf{x})$ from data $\mathbf{x}$ (e.g., deducing the average SAT score from data on their scores). The probability $H(\mathbf{x})$ would be constructed, if in fact the data came from a population where $H(\mathbf{x})$ is false, is zero. (Because this is true for any $\mathbf{x}$, it is also true for any instance $\mathbf{x}_0$). But Worrall claims it would be undefined because the denominator of a conditional probability of a false claim is zero. Now the correct way to view an error-probabilistic statement, for example,

$$\text{P}(\text{test } T \text{ outputs } H(\mathbf{x});\ H(\mathbf{x}) \text{ is false}),$$

is *not* as a conditional probability but rather a probability *calculated under the assumption that* $\mathbf{x}$ *came from a population where* $H(\mathbf{x})$ *is false.* The probability that a maximally severe use-construction rule outputs $H(\mathbf{x}_0)$, calculated under the assumption that $H(\mathbf{x}_0)$ is false, is zero – not undefined. Moreover, if we bar conditional probabilities on false hypotheses, then Bayesians could never get their favorite theorem going because they must exhaust the space of hypotheses.

3.2 Equivocations and Logical Flaws

If counterexamples will not suffice (in this case, to deny UN is necessary for severity), a second philosophical gambit is to identify flaws and equivocations responsible for leading astray even those who profess to share the goal (severity). *But one can never be sure one has exhausted the sources of confusions!* Worse is that the analytic labors carefully crafted to reveal the logical slip can give birth to yet new, unintended confusions. This seems to have happened here, and I hope to scotch it once and for all.

Everything starts out fine: Worrall correctly notes that I identify, as a possible explanation for the common supposition that UN is necessary for severity, a slippery slide from a true assertion – call it (a) – to a very different assertion (b), which need not be true:

(a) A use-constructed procedure is guaranteed to output an $H(\mathbf{x})$ that fits $\mathbf{x}$, "no matter what the data are."
(b) A use-constructed procedure is guaranteed to output an $H(\mathbf{x})$ that fits $\mathbf{x}$, "no matter whether the use-constructed $H(\mathbf{x})$ is true or false" (Mayo, 1996, p. 270; Worrall, this volume, p. 148).

Giere, for example, describes a scientist unwilling to consider any model that did not yield a prediction in sync with an observed effect $\mathbf{x}$. "Thus we know that the probability of any model he put forward yielding [the correct effect $\mathbf{x}$] was near unity, independently of the general correctness of that model" (Giere, 1983, p. 282). It is this type of multiply ambiguous statement, I argue, that leads many philosophers to erroneously suppose that use-constructed hypotheses violate severity. Pointing up the slide from (true) assertion (a) to (false) assertion (b) was intended to reveal the equivocation. Let me explain.

A use-constructed test procedure has the following skeletal form:

Use-Constructed Test Procedure: Construct $H(\mathbf{x})$ to fit data $\mathbf{x}$; infer that the accordance between $H(\mathbf{x})$ and $\mathbf{x}$ is evidence for inferring $H(\mathbf{x})$.

We write this with the variable $\mathbf{x}$, because we are stating its general characterization. So, *by definition*, insofar as a use-constructed procedure is successfully applied, it uses $\mathbf{x}_0$ to construct and infer $H(\mathbf{x}_0)$, where $\mathbf{x}_0$ fits $H(\mathbf{x}_0)$. This is captured in assertion (a). But assertion (a) alone need not yield the minimally severe test described in assertion (b); it need not even lead to one with low severity. The construction rule may ensure that false outputs are rare. We may know, for example, that anyone prosecuted for killing JonBenet Ramsey will have to have matched the DNA from the murder scene; but this is a reliable procedure for outputting claims of form:

The DNA belongs to Mr. X.

In any specific application it outputs $H(\mathbf{x}_0)$, which may be true or false about the source of the data, but the probability that it outputs false claims is low. The familiar argument that use-constructed tests are invariably minimally severe, I suggest, plays on a (fallacious) slide from assertion (a) to assertion (b).

Having gotten so used to hearing the Popperian call for falsification, it is sometimes forgotten that his call was, strictly speaking, for falsifying hypotheses, *if false.* Admittedly, Popper never adequately cashed out his severity idea, but I would surmise that, if he were here today, he would agree that some construction procedures, although guaranteed to output some

$H(\mathbf{x})$ *or other*, whatever the data, nevertheless ensure false outputs are rare or even impossible.

Worrall sets out my argument with admirable clarity. Then something goes wrong that numerous exchanges have been unable to resolve. His trouble is mainly as regards claim (a). Now claim (a) was intended to merely capture what is generally assumed (*by definition*) for any use-constructed procedure. So, by instantiation, claim (a) holds for the examples I give where a use-constructed procedure yields a nonsevere test. From this, Worrall supposes that claim (a) is necessary for nonseverity, but this makes no sense. Were claim (a) required for inseverity, then violations of claim (a) would automatically yield severity. Then hypotheses that do not even fit the data would automatically count as severe! But I put this error aside. More egregiously, for current purposes, he argues that claim (a) is false! Here is where Worrall's logic goes on holiday.

He considers Velikovsky's rule for blocking anomalies by inferring that the culture in question suffered amnesia $A(\mathbf{x})$. Worrall says, consider a specific example of a culture – to have a concrete name, suppose it is the Thoh culture – and suppose no records are found of Velikovsky-type events. Velikovsky conveniently infers that the apparent anomaly for his theory is explained by amnesia:

A(Thoh): Thoh culture suffered from amnesia (hence no records).

Says Worrall, "It seems, then, to be straightforwardly untrue that a successful fit between" A(Thoh) and $\mathbf{x}$ "is assured no matter what [$\mathbf{x}$] is" (p. 149). Quite so! (For instance, the procedure would not output A(Thoh), or any claims about the Thoh culture, if the observation was on some other culture). But this does not show that assertion (a) is false. It could only show that assertion (a) is false by an erroneous instantiation of the universal claim in assertion (a). The assertion in (a) is true because every anomalous outcome will fit *some Velikovsky dodge or other*. It does not assert that all anomalous cultures fit a *particular* instantiation of the Velikovsky dodge, e.g., A(Thoh).[3]

[3] Worrall seems to reason as follows:

1. According to assertion (a), for any data $\mathbf{x}$, if $\mathbf{x}$ is used to construct $A(\mathbf{x})$, then $\mathbf{x}$ fits $A(\mathbf{x})$.
2. But suppose the data from the Thoh culture is used to construct A(Thoh).
3. (From assertion (a) it follows that) all data $\mathbf{x}$ would fit A(Thoh) (i.e., a successful fit between A(Thoh) and $\mathbf{x}$ "is assured no matter what $\mathbf{x}$ is").

Then from the falsity of premise 3, Worrall reasons that premise (a) is false. But premise 3 is an invalid instantiation of the universal generalization in premise (a)! It is unclear whether Worrall also takes this supposed denial of assertion (a) as denying assertion (b), but to do so is to slip into the fallacy that my efforts were designed to avoid. (For further discussion of variations on this fallacy, e.g., in Hitchcock and Sober, 2004, see Mayo, 2008).

Sometimes a gambit that a philosopher is sure will reveal a logical flaw instead creates others. Pointing up the faulty slide from the truth of assertion (a) to that of assertion (b) was to have illuminated the (false but common) intuition that UN is necessary for severity. Instead we have been mired in Worrall's resistance to taking assertion (a) as true for use-constructed procedures – something I took to be a matter of mere definition, which just goes to show that one cannot always guess where the source of difficulties resides. Hopefully now no obstacles should remain to our agreement on this issue.

4 Concluding Comment on the Idea of a Single Account of Evidence (Remarks on Chapters 2, 3, and 4)

I do not claim that all of science involves collecting and drawing inferences from evidence, only that my account is focused on inference. As varied as are the claims that we may wish to infer, I do not see that we need more than one conception of what is required for evidence to warrant or corroborate a claim. Worrall berates me for holding a "one-size-fits-all" account of inference that always worries about how well a method has probed for the errors that threaten the inference. Similar sentiments are voiced by Chalmers and Musgrave. Granted data may be used in various ways, and we want hypotheses to be not just well tested, but also informative; however, if we are talking about the warrant to accord a given inference, then I stand guilty as charged.

I cannot really understand how anyone could be happy with their account of inference if it did not provide a unified requirement. In the context of this chapter, Worrall's introduction of "conditional evidence" was of no help in discriminating warranted from unwarranted use-constructions. The severity desideratum seems to be what matters. Similarly, Chalmers's "arguments from coincidence" and Musgrave's "inference to the best tested portion of an explanation" in Chapters 2 and 3, respectively, are all subsumed by the severity account. Different considerations arise in *applying* the severity definition, and different degrees of severity are demanded in different cases, but in all cases the underlying goal is the same. The whole point of the approach I take is to emphasize that what needs to have been probed are the threats of error in the case at hand. Even if one adds decision-theoretic criteria, which we will see leads Laudan to argue for different standards of evidence (Chapter 9), my point is that, *given the standards*, whether they are satisfied (by the data in question) does not change.

The deepest source of the disagreements raised by my critics, I see now, may be located in our attitudes toward solving classic problems of evidence, inference, and testing. The experimental account I favor was developed precisely in opposition to the philosophy of science that imagines all inferences to be paradigm-laden in the sense Kuhnians often espouse, wherein it is imagined scientists within paradigm T_1 circularly defend T_1 against anomaly and have trouble breaking out of their prisons. In this I am apparently on the side of Popper, whereas Musgrave, Chalmers, Worrall (and Laudan!) concede more to Lakatos and Kuhn. It is to be hoped that current-day Popperians move to a position that combines the best insights of Popper with the panoply of experimental tools and methods we now have available.

At the same time, let me emphasize, there are numerous gaps that need filling to build on the experimentalist approach associated with the error-statistical account. The example of use-novelty and double counting is an excellent case in point. Although in some cases, understanding the way formal error probabilities may be altered by double counting provides striking illumination for entirely informal examples, in other cases (unfortunately), it turns out that whether error probabilities are or should be altered, even in statistics, is unclear and requires philosophical–methodological insights into the goals of inference. This is typical of the "two-way street" we see throughout this volume. To help solve problems in practice, philosophers of science need to take seriously how they arise and are dealt with, and not be tempted to define them away. Conversely, in building the general experimentalist approach that I label the error-statistical philosophy of science, we may at least find a roomier framework for re-asking many philosophical problems about inductive inference, evidence and testing.

References

Giere, R.N. (1983), "Testing Theoretical Hypotheses," pp. 269–98 in J. Earman (ed.), *Testing Scientific Theories*, Minnesota Studies in the Philosophy of Science, vol. 10, University of Minnesota Press, Minneapolis.

Hitchcock, C., and Sober, E. (2004), "Prediction Versus Accommodation and the Risk of Overfitting," *British Journal for the Philosophy of Science*, 55: 1–34.

Mayo, D.G. (1991), "Novel Evidence and Severe Tests," *Philosophy of Science*, 58: 523–52.

Mayo, D.G. (1996), *Error and the Growth of Experimental Knowledge* (Chapters 8, 9, 10), University of Chicago Press, Chicago.

Mayo, D.G. (2008), "How to Discount Double Counting When It Counts," *British Journal for the Philosophy of Science*, 59: 857–79.

Musgrave, A. (1974), "Logical Versus Historical Theories of Confirmation," *British Journal for the Philosophy of Science*, 25: 1–23.

Musgrave, A. (1989), "Deductive Heuristics," pp. 15–32 in K. Gavroglu, Y. Goudaroulis, and P. Nicolacopoulos (eds.), *Imre Lakatos and Theories of Scientific Change*, Kluwer, Dordrecht.

Worrall, J. (1989), "Fresnel, Poisson, and the White Spot: The Role of Successful Prediction in the Acceptance of Scientific Theories," pp. 135–57 in D. Gooding, T. Pinch and S. Schaffer (eds.), The Uses of Experiment: Studies in the Natural Sciences, Cambridge University Press, Cambridge.

Related Exchanges

Musgrave, A.D. (2006), "Responses," pp. 301–4 in C. Cheyne and J. Worrall (eds.), *Rationality and Reality: Conversations with Alan Musgrave*, Kluwer Studies in the History and Philosophy of Science, Springer, Dordrecht, The Netherlands.

Worrall, J. (2002), "New Evidence for Old," in P. Gardenførs, 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.

Worrall, J. (2006), "History and Theory-Confirmation," pp. 31–61 in J. Worrall and C. Cheyne (eds.), *Rationality and Reality: Conversations with Alan Musgrave*, Springer, Dordrecht, The Netherlands.