CHAPTER 5 – JOHN WORRALL:
Error, Tests and Theory-Confirmation

In this paper 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 will 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. There is no sensible argument 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. There is no sensible argument in Duhem (or elsewhere) to the effect that we can never test any particular part of some overall theory or theoretical system, but only the 'whole' of it. If, for example, a theory falls 'naturally' into 5 axioms, then there is - and can be - no reason why it should be impossible that some directly testable consequence follows from, say, 4 of those axioms – in which case only those 4 axioms and not the whole of the theory is what is tested.

What Duhem *did* successfully argue is that if we take what is normally considered to be 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 analyse any attempt to test it empirically by deducing from it some directly empirically checkable consequence, then the inference will be 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, in order to deduce from Newton's theory of gravitation a consequence that can be directly tested against telescopic sightings of planetary positions, we will 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, analysed 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, together 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' which carries the light waves is a highly attenuated fluid and the waves, therefore, longitudinal, and hypothesised 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 have of course 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 an – of course unwitting – rediscovery 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 himself calls them) were no less rational than the "revolutionaries" – there being 'some good reasons for each possible choice' (sticking to the older theory/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 on to 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 there were plenty of suggestions within the early 19th 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 since this force will need 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) were especially impressed by his theory’s ‘natural’ account of the phenomenon of planetary stations and retrogressions. This despite the fact that it had long been recognised 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 centre of that epicycle moved around the basic ‘deferent’ circle.

Kuhn seems to presume that the fact that phenomena such as these 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 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 always will be 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 non-comparative) 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 recognised 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, fitted 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. In order 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 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 will and which will not have appropriate records, this can hardly be said to supply any empirical support to Velikovsky's cometary hypothesis – that hypothesis has been augmented exactly so as to yield the already known data.

These intuitive judgments need to be underwritten by some general principle which will in turn 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 signed up to the 'UN Charter'.[5] This, 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 rather 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 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 already given experimental results such as that of the two-slit experiment itself; while 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, since 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 the present paper further to clarify 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 will consist 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 LSE under the title '*Ad hoc* is not a four letter word'. Beneath the (multiple) surface literal correctness of this title, there is a substantive claim that is undeniably correct. Namely that 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-realised) 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 it, so far as basic theoretical considerations are concerned, an open question what the wavelength of the light from any particular monochromatic source is. Since the theory provides no account of the atomic vibrations within luminous objects that produce the light it does not dictate from first principles what the wavelength of light from a particular source will be. The theory does however entail that that value, whatever it is, is a function of, for example, 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 from the formula entailed by the theory what the wavelength of light from that particular source is. She has then *deduced* a particular, more powerful theory (wave theory complete with a specific value of this particular theoretical parameter) from her general theory (with parameter free) 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 emphasise a different general case — standard statistical examples such as the following (see his 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 analyses in more detail similar statistical cases that seem to count against the 'no double use idea' and also cites the following 'trivial but instructive example' (1996, p. 271). Suppose one wanted to arrive at what she characterises 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 fails to support that hypothesis. On the contrary, as she writes (*ibid.*):

> "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."

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 shall return *below* 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 emphasised long ago by Newton and, after some years of neglect, has been increasingly re-appreciated 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 generalising from models known to be (strictly) inaccurate, deduced his theory of universal gravitation from Kepler's 'phenomena' plus background assumptions including conservation of momentum. The statistical case cited by Howson, exactly because it is statistical, does not of course exactly fit the pattern – but 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 of k/n 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 the 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 one. 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, a sodium arc, say, the (observable) distance X from the fringe at the centre 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′ will determine the wavelength of the light (within of course some small margin of experimental error), 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' really means here '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. There are more impressive cases 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 involve, rather than general but still particular theories, background principles of extreme generality that seem natural (even arguably 'unavoidable'). These general principles delineate a space of possible – again specific but general - theories. Taking those principles as implicit premises, the data, by a process that can either be characterised 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, since the inference from (T_1 v T_2 v… v T_n) and $\neg T_1, \neg T_{j-1}, \neg T_{j+1}, \ldots \neg T_n$ 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 provides 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 are supposing - 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 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, there is a clear sense 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**, but instead simply needs a little

elaboration. The principal step toward seeing this is to recognise 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 $T(\lambda)$. In so far 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 that deduction's conclusion. 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. So long of course 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 different set of fringe distances to 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 of course not have led to the rejection of $T(\lambda)$ but simply to the construction/deduction of a *different* specific version $T(\lambda_1)$, say, 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 there already being independent empirical support for its underlying general theory - is further underlined by the fact that the sort of theoretical manoeuvres that get *ad hocness* a bad name fit the model of 'deduction from the phenomena'. Consider, for example, the Velikovsky dodge outlined above. 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 of them) of appropriate cataclysms). And the deduction goes through in exactly the same way – both the wave theory and the Velikovsky cases then being instances of 'parameter-fixing'.

The difference between the wave theory and Velikovsky cases is simply that in the former but not the latter, there was *independent support* 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, lack of them from others, does nothing to support the general idea of cataclysms associated with the close encounters with the alleged massive comet, since 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 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 programme support'. The lack of records in cultures $C_1 \ldots C_n$ and their (arguable) presence in $C'_1 \ldots C'_m$ gives you 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 (though 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 weren't); 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* they 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 there actually were – other empirical reasons for doing so).

Not all empirical confirmation or support can have this ineliminably conditional and ineliminably intra-programme 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, there are empirical reasons 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 there are cases in which, in contrast to the cases of confirmation we have just been considering, confirmation 'spreads' from the theoretical framework (central theory plus specific assumptions) to the central theory of the framework – that is, there must be some empirical results 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 will, in line with Duhem's point, only follow deductively from some specific version of the theory *plus* auxiliaries.)

There seem in fact to be two kinds of case 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 *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, it also entails 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 then 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 non-trivially) 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 programme confirmation that is obtained in all the cases discussed in section **2**, there is a second, more-powerful kind of confirmation that provides support for the general theory, or research programme, 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 the 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 where there is no independent testability), it might seem reasonable to count the used evidence as itself supportive. *Given* that Adams-and-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 count as (unconditional) support for the general wave theory too. 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, since 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, while Velikovskian specific theories have *only* support conditional on a framework which itself has no support. There do after all seem to be two quite different sorts of scientific reasoning involved – 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 true whenever, as in the wave theory case discussed above, 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-19th 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 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 she 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 and then the underlying general theory with these fixed values of the parameters 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 but 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 then, 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 a case like this it doesn't 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 there *are* two separate things going on 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 *both* used in the construction of the specific theory involved *and* in its (unconditional) support? Well here's one reason: suppose we had two theories T and T′ one of which, T say, has no relevant free parameters and entails $e_1, \ldots e_n$, straight off, while 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, while 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 is so even when the choice of which particular subset (of a certain size) is used to fix parameters and which to genuinely test and hence (possibly) supply genuine 'unconditional' support is arbitrary. 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 there are two kinds of case 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 been considering. 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's 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, Mars say, 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 is directly refuted by the fact that Mars' generally eastward (apparent) motion 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 since the diurnal movement is always superimposed). The introduction of an epicycle of suitable size and the assumption that Mars moves around the centre 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' (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 both proceed steadily eastward around the sun, the Earth, moving relatively quickly round 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. (I am talking here 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 characterise: 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 there is especially direct or strong support in such cases), it cannot be literally correct.

First of all there have to 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 which is both self-consistent and consistent with T and yet which together with T entails not-e. So 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 and retrogressions. But there *are* such possible assumptions. Suppose for example that the earth and Mars are orbiting the Sun in accordance with Copernicus's 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 as exactly to cancel out what would otherwise be the effects of the overtaking (that is, the station and retrogression). Of course this is a monstrous assumption, 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 round its smaller orbit and hence periodically 'laps' Mars.

Let me, then, sum up this section of the paper. It seems obvious on reflection, or so I claim, that there are two quite different precise ways of using data in science each of which fall 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 will lead 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/ research programmes that are 'degenerating' from those that are 'progressive'; and at systematically underwriting the intuitive judgment that when some piece of evidence e is predicted by some specific theory within general programme P, but only accommodated *post hoc* by some specific theory within general rival programme 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 doesn't rule out there being 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 above; and (correctly) emphasising 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 this. They were 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 the more positive the more such independent empirical support there is 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 exacty the same as that cited by Mayo in favour 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 recognising – 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 it were not somehow 'along the right lines'.

4. Mayo's alternative: confirmation is all about 'severe tests'

How do these views on confirmation compare with the influential and more highly developed views of Mayo? There are certainly some striking similarities. 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 conflict what seem to be clearly valid intuitive judgements 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; and 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 since evidence e used in the construction of T cannot possibly test T, as there is no chance of T's failing the 'test' whose outcome is e- that outcome was instead 'written into' e. No matter how plausible this may sound, argues Mayo, it in fact misidentifies the probability that we should be concerned to maximise in order to maximise severity and hence it misidentifies the real notion of a severe test. It is easier to understand her characterisation 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 *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 above as identical in the relevant respects), the 'test' at issue was indeed non-severe: there would be a good chance of the modified Velikovsky theory passing the test of no records of suitable cataclysms in culture C even though that theory were false. On the other hand, in those cases where the 'no double use rule' goes wrong, such as her SAT score example, while 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, since 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's, there are few bigger fans of unity than I. And Mayo here offers a unified alternative to my 'two kinds of confirmation' view – there aren't two kinds of confirmation but 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 which are left as merely suggestive notions within it (though I hope with clear-cut illustrations from particular scientific cases).

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

1. There seem to me 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 last 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 an '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 importantly, since we all agree that the evidence here is conclusive for the 'hypothesis' and it might be felt that it doesn't really matter how we choose to express this, 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; and indeed 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' just would *have to* yield 1121 again with this particular

group of students; if you were to arrive at any other figure you simply wouldn't 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 represents a non-severe test.) However the 'no matter what' here may refer to two quite different conditions:

(a) No matter what *the data are*, or

(b) No matter *whether H is true or false*.

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

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 (b), at least in the SAT score case (and, as I will shortly argue, more generally). But there are problems with condition (a) too – and ones that apply across the board. The collective-amnesia-ised 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 non-severity? In particular, does her

condition (a) 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 will 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′ which turns out to have outcome e is non-severe 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 bi-conditional V′ iff 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, that is, need to recognise that, as my account entails, there are two separate issues – 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 however '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. This is 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 what is at issue. We surely want to say that e provides no good reason to take V′ seriously either. 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′; and this requires that conditions (a) and (b) of her latest formulation of (non-)severity be satisfied; but as we just saw condition (a) is *not* in fact satisfied. If Mayo were tempted, in response to this, 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 this would mean that she failed to capture the alleged exceptions to the UN-rule. These (apparent) exceptions, as explained earlier, all involve deductions from the phenomena which 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 of course 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 principle of universal gravitation), but on the contrary presupposed it. It is again surely clear why her account is meeting these difficulties. We are dealing, in accordance with 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 artefacts of the strange, clearly pseudoscientific case of Velikovsky's theory, and nor are they restricted to problems with her condition (a) for non-severity, let's 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.

The case involves, remember, the general wave theory of light, call it W. W leaves the wavelengths of light from particular monochromatic sources as free parameters. W 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 centre 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 of these terms, call their conjunction e, will determine the wavelength (within of course 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 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 (b) 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 (b) for *non*-severity is *met* here: whether or not W′ is true the fit with e is assured, since the value of λ specified by W′ has been calculated precisely so as 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 (a) fails to hold in this case. There is in fact an ambiguity over what 'it' is in the condition (for non-severity, 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 *etc* 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 since it is not true that both conditions for non-severity hold in this case, the test may be regarded as at least somewhat severe. But clearly what Mayo intended was that condition (a) *should* in fact hold in 'use-constructed' cases and that it is the failure of (b) to hold in

certain particular cases (despite (a) 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; and 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 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 (b) of her latest account. But even supposing we go along with this view as to what her account entails here, how then will 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 is instead taken for granted (indeed in the SAT course case there is no option but to take the underlying theory for granted since 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 there is some unknown but fixed ratio of such balls in the urn 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 condition (b) for non-severity. Her account would have to be understood as saying that a test of T is non-severe if the test's outcome is

bound to fail to refute T (condition (a)) *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 already suggested earlier more than once, by thus writing my analysis into one account of severe versus non-severe 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 a confusing reformulation of the view that I defended. It surely just is the case that there are two separate roles for evidence in science: 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,* Cambridge, MA: MIT Press.

French, A. (1971), *Newtonian Mechanics*. Cambridge, Mass: M.I.T. Press.

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*. Minneapolis: University of Minnesota Press.

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

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

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

Lakatos, I. and Zahar, E.G. (1976) 'Why did Copernicus's Programme Supersede Ptolemy's' reprinted as chapter 4 of I .Lakatos *The Methodology of Scientific Research* Programmes. Cambridge University Press, 1978.

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

Nickles, T (ed.) (2003), *Thomas Kuhn*, Cambridge University Press, Cambridge.

Worrall, J. (2000b), '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**, pp. 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 Thomas Nickles (ed): *Thomas Kuhn*, Cambridge University Press, Cambridge.

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

Worrall, J. (2007) 'Miracles, Pessimism and Scientific Realism', *British Journal for the Philosophy of Science*: forthcoming.

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

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

[3] Kuhn (1962), p.151-2

[4] Kuhn (1977), P.328

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

[6] Notice then 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 conformational weight were missing the real issue. This is the issue of accommodation vs. 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 on *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 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)

[7] See Howson (1990).

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

[9] See Earman (1992, pp. 173-180) see also the discussion in Mayo *op. cit*..

[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 obviously true. I am therefore idealising somewhat in order to make it a crisp case of deduction from the phenomena (it isn't really as good as that!). But I believe that all the methodological points stand in spite of this slightly idealising move.

[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 recognised 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 recognised, this test would not have been reason to continue to prefer Ptolemy over Copernicus, since, as will immediately become apparent, the the Copernican theory too 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).

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

[14] See Worrall (2007).

[15] Uri Geller asserted of course (indeed no doubt still asserts) that he has genuine psychokinetic powers; when, unbeknown to him, professional magicians controlled the situation in which he was to exhibit these powers by for example 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 sceptics in the audience. Obviously he could only identify whether or not scepticism was playing this obstructive role *post hoc*: if he was able to bend spoons by the 'power of pure thought' then no sceptics 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* scepticism in the air.

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

[17] In fact this as well as some of the other cases that Mayo analyses—such as the identification of the car that hit her own car's fender, or the technique of 'genetic fingerprinting'—seem altogether more naturally categorised 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 s/he were innocent.