Theory-laden experimentation

Samuel Schindler

Centre for Science Studies, Aarhus University, Denmark

samuel.schindler@ivs.au.dk

Abstract

The thesis of theory-ladenness of observations, in its various guises, is widely considered as either ill-conceived or harmless to the rationality of science. The latter view rests partly on the work of the proponents of New Experimentalism who have argued, among other things, that experimental practices are efficient in guarding against any epistemological threat posed by theory-ladenness. In this paper I show that one can generate a thesis of theory-ladenness for experimental practices from an influential New Experimentalist account. The notion I introduce for this purpose is the concept of 'theory-driven data reliability judgments' (TDR), according to which theories which are sought to be tested with a particular set of data guide reliability judgments about those very same data. I provide various prominent historical examples (among others, the confirmation of Einstein's prediction of star light bending in 1919) to show that TDRs are used by scientists to resolve data conflicts. I argue that the rationality of the practices which employ TDRs can be saved if the independent support of the theories driving TDRs is construed in a particular way.

Key words: theory-ladenness, experiment, data conflicts, independent support, novelty, theoretical virtues, star light bending

1 Introduction

The thesis of theory-ladenness of observations, roughly, is the idea that observations are affected by theoretical presuppositions. One can distinguish between at least three versions of theory-ladenness of observations (cf. Bogen 2010): (i) theories impact on perceptual processes so that 'what we see' is partially determined by our theoretical presuppositions; (ii) observations cannot be *described* in a theory-neutral way and the *meaning* of observational terms is determined by theoretical presuppositions; (iii) theories make certain observations *more salient* than others because some observations are just more interesting from a certain theoretical perspective than others. An example for (i) is the anomalous playing card experiment, famously used by Kuhn (1996), in which subjects "see" anomalous playing cards (such as the black four of hearts) as normal (namely, as the red four of hearts), because their conceptual categories are adjusted to the latter, but not the former. An example for (ii) is a different meaning of the term 'temperature' that the caloric theorist and the modern physicist assign to it, and an example for (iii) is the idea that *different* parts of a pendulum will be observationally salient for the Aristotelian and the Galilean, due to their respective concepts of motion.

The first version of the thesis of theory-ladenness of observations has been denied by various authors by appealing to the cognitive impenetrability of perceptions (Fodor 1984; Raftopoulos 2009).[1] The part of the second version that has it that there are hardly any pure observation reports is well accepted and is one of the reasons that led to the demise of logical positivism (Suppe 1977; van Fraassen 1980). However the other part of the second version has also been denied. Most philosophers reject a holist conception of meaning as defended by Quine (1951), and as implicitly held by Kuhn (1996). The third version of the thesis of theory-ladenness has received considerably less attention than the former two versions, possibly because a theory rendering certain evidence more salient than other evidence is "neither inevitable nor irremediable" (Bogen 2010).

Not always is theory-ladenness (in whatever version) epistemologically problematic. But in at least three scenarios it is. First, if observations are theory-laden with the assumptions that these observations are supposed to test, then the rationality of theory-testing is put in jeopardy on pain of circularity. Second, when observations are theory-laden by two logically incompatible theories, adherents of either theory may not reach agreement on which theory to adopt on the basis of these observations. Third, observations may be theory-laden in such a way that scientists perceive only confirming, but no disconfirming evidence.

An important rebuttal of the epistemologically problematic forms of theory-ladenness comes from so-called New Experimentalism. The New Experimentalism (a term coined by Ackermann 1989) promoted a turn away from the theory-observation dichotomy towards the experimental practices of science. It is often characterised in terms of Ian Hacking's famous slogan of experiments having "a life of their own" independently of theory. Mayo (1994, 270-1) has given three readings of this slogan. In the first reading, "experimental inquiry may be quite independent of testing, confirming or filling out some theory". This reading, I take it, says that at least some parts of experimental practice are not

[1] There has been a recent critique of this defense Lyons (2011).

directed by attempts to confirm or reject theories. Steinle (1997) aptly dubbed experiments that are not theory-driven and that are conducted in order to systematically explore new phenomenological realms “exploratory experiments”. In the second reading, “experimental data may be *justified* [sic] independently of theory” [2] and “need not be theory-laden in any way that invalidates its role in grounding experimental arguments”. In the third and last reading, “experimental knowledge may be retained despite theory change” (ibid., p. 270-1). The most interesting thesis of New Experimentalism for present purposes clearly is Mayo’s second reading of Hacking’s slogan.

The New Experimentalists and their sympathizers have sought to rebuke the thesis of theory-ladenness in several ways. A strategy highlighted by Hacking (1983) and defended in detail by e.g. Culp (1995) is the so-called argument from coincidence or robustness: it would be a preposterous coincidence if the data produced by various experimental methods or instruments were to coincide and not be reliable. Even though the design and use of experimental methods or instruments presupposes the truth of certain theoretical assumptions, and even though some instruments may be laden with the same theoretical assumptions that are sought to be tested, the theory-ladenness of the *individual* measurement procedures is debilitated by the fact that a *number* of methods converge on the same results. Another way of rebuking the thesis of theory-ladenness is to deny its relevance. As Bogen and Woodward (1988) have claimed, theories are not tested on the basis of observations but rather on the basis of “typically unobservable” phenomena, which are inferred from the data through statistical and experimental methods. Since phenomena are normally not observable, they cannot possibly be theory-laden (cf. ibid., 342-7). Mayo (1996) has sought to erect a philosophy of experiment on the basis of her notion of severity of test and on the basis of error statistics. Experimental activities like ruling out artifacts, distinguishing signal from noise etc., according to Mayo, “receive structure from statistical methods and arguments” (1994, 272). In Mayo’s account there is usually no risk of theory-ladenness since local experimental and statistical methods are distinct from the theory that is being tested.

Although the New Experimentalists, in order to escape the thesis of theory-ladenness, have been keen to stress those procedures of experimental practice that are

[2] Note that it is slightly awkward to speak of the justification of data rather than the justification of statements describing those data.

unaffected by the theories at stake, interestingly, the perhaps most extensive analysis of data reliability in science, i.e. Allan Franklin's account of 'epistemological strategies', does grant *some* role to those theories in data analysis. According to Franklin, such theories can guide judgments about data reliability if there is independent support for those theories (Section 2). Although *prima facie* epistemologically unproblematic, the problem of theory-ladenness does crop up again if independent support is not construed in a particular way—or so I shall argue. Before presenting an argument to this effect, I shall first argue for the positive role of theories in guiding data reliability judgments. In particular I shall argue that guidance by theories is particularly valuable when it comes to data conflicts (Section 3). In Section 4 I will then argue for a particular form of independent support in order to evade the threat of theory-ladenness. The more abstract discussion will be proffered by a number of examples from scientific practice in Section 5. Section 6 will draw attention to the diversity of attitudes when it comes to the use of theories as guides to making judgments about data reliability. Section 7 will conclude this paper.

2 Theory-driven data reliability judgments

In the parlance of scientists data can be 'good' or 'bad', reliable or unreliable[3]. What scientists mean by these predicates is that the data produced by an experiment either was or was not produced by the causal factors of interest. If an experiment produces data that result from some artifact or confounding factor that was not controlled for, then the data are 'spurious', not 'trustworthy', or simply unreliable. But data do not carry on their sleeves to what kind they belong. Scientists must therefore not only take various precautions when producing data, but they must also carry out sophisticated error estimates for which the raw data then have to be corrected. Only after these procedures have been carried out to a satisfactory degree we can speak of 'hard facts' against which we should want to test our theories.

In numerous publications, Allan Franklin has argued that scientists have a whole battery of 'epistemological strategies' at their disposal for ensuring data reliability (before, during, and after data production). These strategies, Franklin emphasizes, are not infallible rules but

[3] Strictly speaking, it is not correct to speak of the reliability of data, since reliability is an attribute of a process. I nevertheless want to stick to scientists' jargon in this paper.

rather heuristic guides. As such they can steer researchers into the right direction, but they can also fail (e.g. Franklin 2002, 6). Among those strategies, Franklin has listed the following:

1. Experimental checks and calibration, in which the experimental apparatus reproduces known phenomena;
2. Reproducing artifacts that are known in advance to be present;
3. Elimination of plausible sources of error and alternative explanations of the result (the Sherlock Holmes strategy);
4. Using the results themselves to argue for their validity;
5. Using an independently well-corroborated theory of the phenomena to explain the results;
6. Using an apparatus based on a well-corroborated theory;
7. Using statistical arguments.

As can easily be seen, most of Franklin's strategies are very much in the spirit of the second tenet of New Experimentalism. Interestingly, however, Franklin does indeed reserve a place for theories in efforts to establish data reliability (strategy 5). This strategy contains the seeds for what we are going to be concerned with in the remainder of this essay. Rather than further describing strategy 5, Franklin illustrates it with the example of the discovery of the W± bosons in 1983, as predicted by the Salam-Weinberg-Glashow model more than ten years earlier:

> I believe that the agreement of the observations with the theoretical predictions of the particle properties *helped to validate the experimental results*. In this case the particle candidates were observed in events that contained an electron with high transverse momentum and in which there were no particle jets, *just as predicted by the theory*. In addition, the measured particle mass of 81 ± 5 GeV/c2 and 80+10-6, GeV/c2, found in the two experiments (note the independent confirmation also), was *in good agreement with the theoretical prediction* of 82 ± 2.4 GeV/c2. (Franklin 2002, 5, added emphasis)

Let us first of all note that Franklin discusses only the positive uses of strategy 5 (and of the other strategies, for that matter). That is, although he claims that "the agreement of the observations with the theoretical predictions […] helped to *validate* the experimental results", he never mentions that theories could help to *invalidate* experimental results. But it is not clear on which grounds one would not want to allow for such a possibility. In fact, as we shall see later in Section 3, this possibility is not merely a logical possibility. Scientists actually do make use of it. It is also interesting to note that all the examples Franklin cites for strategy 5 concern cases in which the theories guiding data reliability judgments are the very same theories which the data in question are supposed to test. Of course this need not be so, but this class of data reliability judgments is certainly the most interesting class when it comes to concerns about theory-ladenness.

Allowing for both the positive and the negative guidance of theories in questions about data reliability, and assuming that the theory guiding judgments about data reliability is the very same theory the truth of which is at stake, I want to introduce the notion of *theory-driven data reliability judgments* (TDR):

> **TDR:** An independently well supported theory T guides researchers in their assessment of the data by giving them good reasons for trusting that data implied by T are reliable and for doubting the reliability of those data that are not implied by T.

Theories that possess good inductive support induce the expectation that their predictions will be correct not only for the data at hand but also for other data. Barring the problem of induction, this expectation is reasonably rational. The inductive support of a theory may motivate the application of TDRs, i.e., it may guide scientists in their judgments about whether other data are reliable or not. Think of the recent report that neutrinos travel fast than the speed of light (Agafonova et al. 2010). The immediate response of the physics community was disbelief and suspicion that some experimental error would account for the result. Given that Einstein's theory of relativity, which is built on the assumption that nothing travels faster than the speed of light, has had such tremendous empirical success in the past, it would have been everything *but* reasonable not to question the reliability of the reported results.[4]

TDR might raise concerns about theory-ladenness. If the theory guiding data reliability judgments about a set of data that is then used to test that very same theory, the concern is that the theory guides the judgments about data reliability in such a way that the question of whether or not the data are reliable will *automatically* be answered in the theory's favor. But such a concern would only be appropriate if the theory in question were to *determine* whether or not the data were reliable. Because then the reliability of the data would depend on the truth of the assumptions made by the theory in question, when it was

[4] In a weak sense, the TDR thesis presupposes the truth of the Duhem-Quine thesis. According to that weak sense, a negative experimental result need not imply the falsity of the theory at stake; it may instead be due to a false auxiliary or background assumption. However the TDR thesis does *not* imply a stronger sense of the Duhem-Quine thesis, according to which we cannot know which part of the 'net' of our beliefs is mistaken. On the contrary, TDRs give the scientist clear directions of where to look for errors. See the next section for details.

the truth of those assumptions that we wanted to find out about on the basis of the data in question. But of course TDR does *not* say that the reliability of the data is determined by the theory in question; all it says is that reliability judgments about a certain set of data can be *guided* by a theory whose truth is sought to be tested with those data. But how is the guidance of a theory in data reliability judgments to be understood?

Here is a way of thinking about it. TDRs guide the researcher in that they suggest to her an increase of efforts in her search for errors in those experiments which produced results that threatened the theory in question. In these efforts the researcher may then also perform new experiments in which she will attempt to tighten her control over those potential error sources. Conversely, if an experiment conforms to the predictions of a theory, the researcher will not be driven to perform error checks that go *beyond* the checks already carried out. In other words, TDRs may guide the researcher in her decision when to stop and when to continue in her error searches. Whether or not the researcher will find errors in the experiments is, at least in principle, an entirely contingent matter. Still, one may be worried about the following. Suppose an experiment E1 has been checked for errors. Because it agrees with T, error searches are not intensified. As a matter of fact, however, if there had been further error searches, scientists would have found further errors, shedding doubt on the trustworthiness of the original results. Since TDRs do not motivate scientists to probe E1 even harder than they already have, this error is never found. However, if this is a serious concern then it is not specific to TDRs. The risk of ending error searches prematurely is always a possibility, regardless of the application of TDRs.

Before turning our attention to epistemologically problematic forms of TDRs, let us ask why it might be attractive for scientists to apply TDRs in the first place.

2.1 TDRs and data conflicts

In his recent book *Selectivity and Discord*, Franklin (2002) highlights the problem of data conflicts as a significant problem for his account of epistemological strategies:

> […] it is a fact of life in empirical science that experiments often give discordant results. The occurrence of such discordant results [after the application of strategies] casts doubt on my epistemology of experiment and on the reasonable use of experimental results in science. If, as is the case, each of the experiments involved applied the epistemology of experiment, how can they produce discordant results? (Franklin 2002, p. 35; emphasis added)

Franklin claims that "[t]he resolution [of data conflicts] *must* proceed by demonstrating that, at least in some experiments, the strategies have been applied incorrectly" (p. 162). "The perhaps most important method of invalidating a result", Franklin claims, "is to show that the Sherlock Holmes strategy has been incorrectly applied" (ibid.). More concretely, "one can argue … that a plausible source of error (e.g. a background that might either mask or mimic the correct result) has been overlooked" in one of the conflicting data sets (ibid.). There are two ways of interpreting this sentence. Either scientists *negligently* overlooked an error source, or they were *limited in their means* in detecting an error at a particular point in time. Although the fact that Franklin speaks of an *incorrect* application of the Sherlock Holmes strategy (rather than of the *correct* but limited application of this strategy) seems to imply the former sense, it is plausible to assume that the failure to detect an error is more often due to the limited means of scientists at a particular point in time. Note that the latter type of error is just one form of the general fallibility of Franklin's strategies:

> [Epistemological strategies] provide us with good reasons for belief in experimental results. They do not, however, guarantee that the results are correct. There are many experiments in which these strategies are [successfully] applied, but whose results are later shown to be incorrect. (p. 6; added emphasis)

Thus, data may conflict even after Franklin's strategies have successfully been applied simply *because* strategies may return a reliability verdict even though the data are not reliable. But note the following tension: if data conflicts occur frequently even after the strategies have been applied (see first quote), and if the fallibility of the strategy is to blame for these occurrences (rather than the scientists' negligence), then this would mean that the degree of fallibility of the strategies is high. But this is undesirable for Franklin for he regards the strategies as *reliable* guides for negotiating data reliability. There are two ways for Franklin to escape this dilemma. Either he retracts from his claim that data conflicts occur frequently even after the application of his strategies, or he does blame the scientist's negligence. The latter is certainly undesirable if one trusts the reliability and rationality of experimental practices, as Franklin does. However, with regard to the first horn of the dilemma, it would seem rather awkward for someone like Franklin, i.e., someone with knowledge about experimental practices as intricate as hardly any other philosopher, to have made such a terribly mistaken descriptive mistake. Whichever way we turn it, we seem to have reached a *cul-de-sac* in Franklin's account.

Franklin’s account presents us with another puzzle. Franklin makes no distinctions between his strategies when it comes to the degree of fallibility; they all seem to be equally fallible. But if they are so, how do we figure out *which* strategies have let us down when data conflicts occur and when strategies have been applied to both data sets successfully? Franklin suggests that data conflicts can be resolved when attempts to replicate either of the conflicting data sets fail (p. 162). That is, one may infer that, despite first appearances, the reliability verdict on a particular set of data given by particular strategies is false if the data are not replicable. And yet, data can be unreliable and indeed be replicable[5]. In the 1950s and 1960s so-called bacterial mesosomes (i.e. membranous invaginations) were considered to be real and biological parts of certain bacteria (Rasmussen 1993). Over several years numerous replications were carried out in which mesosomes would show up. In other words, mesosomes were successfully replicated, even though they were not real. Yet in the mid-1970s biologists decided that mesosomes are indeed artifacts of the cytoprotection and fixation techniques. Consider another example. Before J.J. Thomson managed to deflect cathode rays with an electric field (in order to show that cathode rays consist of electrically charged particles, namely electrons) and 14 years after Heinrich Hertz’s first attempts in 1883, it was not known that cathode rays would ionise the remaining gas in the cathode ray tubes (Hon 1987). These ions would then neutralise the electric field and thereby prevent the deflection of cathode rays. Without any knowledge about this neutralising effect, about which Thomson had speculated (Falconer 1985), this false null result could have been replicated ad nausea. Consider this last example. Over a period of over ten years now, stable data *for* and stable data *against* the hypothesis of antidepressants reducing the effectiveness of breast cancer drugs has been produced (Holzman 2009). Thus, unreliable experimental results not only can be replicable but they also can be *stably* replicable.

Although replicability undoubtedly is an important strategy for ensuring data reliability and for disambiguating data conflicts, scientists have a more principled means at their disposal, namely TDRs. Whereas it is possible for Franklin’s other strategies and for replication to result in reliability verdicts about conflicting data sets, this cannot possibly happen when TDRs are applied. That is, if the theory driving TDRs is not inconsistent, TDRs *must* output that one data set is reliable and the other one isn’t. Furthermore TDRs always

[5] In fact Franklin himself admits that “incorrect results [in science] have been replicated” (p. 241). The example that Franklin mentions concern experiments on low-mass electron-positron states.

indicate *which* data set might be reliable and which one might not. Of course, and this needs emphasizing, whether the data set singled out as unreliable actually is unreliable, is an *entirely contingent matter*. TDRs cannot determine the reliability of data. When applying TDRs the researcher only has inductive reasons for thinking that a particular data set is reliable or not. But the consequences entailed by the theory in question may of course be wrong (despite inductive support) and the relevant TDRs therefore false. Nevertheless, whether correct or not, a theory's consequences do give scientists a clear guidance as to where to increase error searches, indeed a much clearer guidance than the scientist can expect to receive from any of Franklin's other strategies.

One must guard against further misunderstandings. First, the TDR notion is no stand-alone notion. It needs to be filled with the life of experimental practices. That is, TDRs must be combined with any of Franklin's or with other experimental strategies in order to probe data reliability. TDRs are therefore perhaps best described as a sort of 'background constraint' on error searches and reliability judgments. Relatedly, the notion of TDR is indifferent about the type of error to which the error searches triggered by TDRs ultimately lead to.[6] Second, the TDR thesis is no universal thesis. That is, it may well be that TDRs are used only in some parts of scientific enquiry but not in others. Relatedly, it is no part of my thesis that TDRs are used by *all* researchers of a particular scientific community (in one particular corner of research). As we shall see in Section 3, however, there are historical examples of data reliability judgments which are best explained by the majority of the relevant scientific community applying TDRs. Third, scientists need not be aware of the fact that they do apply TDRs. That is, they may be aware only of the experimental strategies that are guided by TDRs without being aware of the fact that they are so guided.

2.2 TDRs, undetected error sources, and theory-ladenness

As we saw earlier, the TDR thesis *per se* is not epistemologically problematic. However, the situation is considerably complicated by the observation of scientists postulating *undetected error sources*. That is, when TDRs indicate the reliability of a certain experimental result (because the theory's predictions are inconsistent with it), and error searches are intensified as a result of that but turn out unsuccessful, it does happen that scientists, on the basis of

[6] Hon (1987) distinguishes between several types of error sources. Although important, these distinctions will be of no concern in this paper. Again, my focus is on the motivations for performing error searches rather than on what types of errors these searches ultimately bring to light.

TDRs, deem those experiments unreliable despite being unable to identify an error source (we will discuss examples in Section 3). Equivalently, scientists may *speculate* about possible error sources without being able to establish the actual presence of any one of them, and still deem the experiment in question unreliable. The postulation of *undetected error sources* is not especially surprising. Many experiments are so complicated that potential sources of error can be exceedingly difficult to identify[7]. However what should be surprising is *that* scientists agree to deem experiments unreliable without being able to identify an error. This, I believe, is best explained by scientists applying TDRs.

The postulation of an undetected error source can be *tentative* or *conclusive*. Whereas in the former case, the scientific community withholds judgment as to whether the theory in question is confirmed or not (e.g. by waiting for more unequivocal evidence), in the latter case, the scientific community makes a conclusive judgment about this question. Certainly in the latter case, theory-ladenness must be a real concern. The theory driving data reliability judgments does seem to *determine* the reliability of the data relevant to its confirmation after all. But, as discussed above, for that to be the case the theory whose assumptions are to be tested would be presumed to be true, on pain of circularity. And yet, the theory does not determine the reliability of *all* relevant data but only those which are inconsistent with the theory's predictions. Does the theory receive support from those data that are predicted by the theory? The answer to this question seems to be strongly context-dependent. First consider the clear cases. In situations in which there is only positive evidence, the evidence does of course support the theory. In situations in which there is only negative evidence, it would be absurd for scientists to accept the theory as being confirmed by postulating undetected error sources for those. But consider situations in which there is positive *and* negative evidence. First assume that there are *unequal* amounts of positive and negative evidence. In contexts where the *negative* data constitute the *vast majority* of evidence, the theory should receive only little or no support at all from the positive evidence when the negative evidence is dismissed by the postulation of undetected error sources. Conversely, in contexts in which it is the *positive* data that constitute the vast majority of evidence, it seems, the theory should indeed receive support from the positive evidence, even when the negative evidence were to be disqualified by postulating undetected error

[7] A cell biologist from Cambridge has recently been quoted as saying that "Things are different in different labs for very subtle reasons. The water can be different. We're about to move labs, and my group is very concerned that delicate cells might hate something in the new pipes" (Giles 2006, 345).

sources. Now consider the most interesting kinds of situations, namely situations in which there are roughly *equal* amounts of positive and negative evidence. In those situations the theory's being supported by the positive evidence would appear to be conditional on the negative evidence being unreliable. If in those situations the negative evidence were to be dismissed by the postulation of undetected error sources, not only the negative evidence being unreliable but also the *positive* evidence supporting the theory would seem to be conditional on the theory's being (approx.) true. In other words, the *whole* testing procedure would be circular. The negative *and* the positive evidence would be theory-laden in an epistemologically problematic way even though only the reliability of the negative evidence would be determined by the theory in question. But note that, contrary to first appearances, even in these situations it would not necessarily be irrational to adopt the theory in question. It would not be irrational in cases in which the theory in question had already garnered independent empirical support. The theory's independent support, in turn, would give the scientists *inductive* reasons for questioning the reliability of the negative data and for treating the positive data as support for the theory in question. Of course, this independent support better be *strong* independent support for the inductive reasons to be good reasons.

2.3 Strong and weak independent support

What does it mean for a theory to be *strongly* supported? The right answer to this question may look trivial: the more evidence there is in a theory's favor, the stronger its support. But this is so only *ceteris paribus*. It is well-accepted that a theory that accommodates evidence in an *ad hoc* way should receive no support from that evidence. And it is accepted that this is so independently of the amount of evidence. *Prima facie*, a theory's best insurance against *ad hoc* accommodations is it to predict, rather than to accommodate, the facts. That is, a certain piece of evidence E can be accommodated by a theory T if E was known at the time T was proposed. But T could not have possibly accommodated E if E was not known at the time T was proposed. This gives us the following criteria for weak and strong empirical support.

Weak P-support: E is weak empirical support for T if E was not predicted but merely accommodated by T.

Strong P-support: E is strong empirical support for T if E was predicted by T.

In spite of the intuitive appeal of P-support, it has been pointed out that it is too strong a notion. John Worrall and others have shown in several important historical case studies that scientists do not appear to adopt it. Worrall (1989) for example showed that Fresnel's famous white spot prediction appears to have counted no more in the appraisal of Fresnel's theory than his accommodation of the already known straight edge diffraction phenomena (ibid.). In order to make sense of cases like these, Worrall has suggested that novel evidence instead be interpreted more broadly, namely as use-novel evidence: evidence E is use-novel with regard to a theory T that entails E, if E was not used in the construction of T. Use-novelty not only is in better accord with scientific practice but it also, Worrall claims, captures much better the intuition driving P-support: why should the bare time-order between T and E be of any significance for the confirmation of T by E? Worrall's account gives us the following support criteria:

Weak UN-support: E is weak (or no) empirical support for T if E was used in the construction of T.
Strong UN-support: E is strong empirical support for T if E was *not* used in the construction of T.

In the examples that I will consider in the next section, however, the independent support for the theories driving TDRs was only of the weak kind. By the lights of Weak UN-support, scientists applying TDRs in these cases had thus rather weak (if any) reasons to apply TDRs. It would then follow that in these cases scientists dismissed the data in question unjustifiably. But given the prominence of these examples I think we should resist this implication. Rather I believe we need to question the plausibility of *Weak* and *Strong UN-Support* as methodological imperatives[8]. In fact, there are well-rehearsed general criticisms of the plausibility of the UN criterion. The most damaging is perhaps this: why should it matter to the *communal* appraisal of a theory whether or not the individual constructing that theory

[8] Worrall himself recognizes that the unqualified UN criterion is too strong. He therefore introduces the further constraint that UN violations must result in independent support for T (e.g.Worrall 2002). Since independent support for T is what the Weak UN-support is supposed to establish in the context of the application of TDRs, appeal to independent support for T in order to establish independent support for T is circular.

sought to accommodate a particular fact or not. After all it seems overly optimistic to believe that this "biographical" information is always made transparent by those individuals and that it is always fully available to the community (cf. Gardner 1982)[9]. Hence, I want to suggest the following support criteria in the stead of UN-support:

***Weak V-support*:** E is weak (or no) empirical support for T if E is explained or predicted by T and T is a non-virtuous (i.e. an inconsistent, piece-meal, convoluted, incoherent, 'infertile') theory.

Strong V-support: E is strong empirical support for T if E is explained or predicted by T and T is a virtuous (i.e. a consistent, unifying, simple, coherent, fertile) theory.

What are the motivations for this proposal? It is clear enough that a theory that is consistent should receive more support from a particular set of evidence than a theory that is inconsistent. It is also clear that a theory that unifies the phenomena in a simple and coherent way should receive more support from those phenomena than a theory that explains those phenomena by invoking conceptually distinct explanations that are only loosely related. A case in point is the competition between Lorentz's and Einstein's proposal for how to make sense of the Michelson-Morley ether drift null result. Lorentz's explanation was effectively the amendment of the then standard ether theory with the notorious Lorentz-FitzGerald contraction hypothesis. Whereas in the amended aether theory there remained "a strict separation of ether and matter", Einstein's theory of relativity provided a coherent theory, in which the laws governing matter and fields received a *common* justification in terms of Minkowski spacetime (Janssen 2002).

How are the properties alluded to by V-support to be understood precisely? On some of them we appear to have a fairly good grasp. This is certainly so for the property of consistency and there have also been very influential accounts of unifying power (Kitcher 1981). Yet other theoretical virtues, such as simplicity and coherence, are notoriously hard to come by. Although there have been attempts to provide a rationale for the imperative of theories having to be simple (Forster and Sober 1994), it is not clear that these discussions have implications beyond the limited domain in which they have been developed[10]. There

[9] See (Schindler forthcoming) for more details.

[10] Forster and Sober (1994) develop their account in the context of model selection. If a model is tied too closely to a particular "training set", then its capacity to accommodate new ('unexpected') data sets will be

have been also attempts to spell out the property of coherence precisely (beyond the rough meaning of "things hanging well together") within the Bayesian framework (Bovens and Hartmann 2003), but it is far from clear that these discussions have picked out the property that we mean when we attribute coherence to theories like Einstein's theory of relativity[11]. The theoretical virtue of fertility was first described by Kuhn (1977) as the capacity to "disclose new phenomena or previously unnoted relationships among those already known" (p. 357). Kuhn was not more specific on this virtue than that, but, by others, fertility has been taken to imply more than the capacity to make novel predictions. McMullin (1976) for instance explains that a theory's fertility springs from its "imaginative resources" which "enable anomalies to be overcome and new and powerful extensions to be made" (p. 16).[12]

In whatever way the *precise* nature of these properties is going to be spelled out precisely in our final philosophy of science, one may, at the present point, more modestly, try to understand the *role* these properties should play in our philosophy of science[13]. Indeed, a major aim of this essay may be understood as an attempt to direct attention to these properties when it comes to the confirmation of theories, attention that these properties have hitherto received only to a rather limited extend. This I want to do by posing the following (potential) dilemma with regard to the examples to be discussed in the next section: either we adopt the V-support criterion, and therefore regard the theories in question to be independently supported, or these cases must be deemed examples for irrational scientific practice due to the implied theory-ladenness. Since these cases are regularly considered to be examples for some of our most impressive scientific achievements, the second horn of the dilemma must be avoided at all costs. To many, the first horn of the dilemma, due to the difficulties associated with specifying the nature of

diminished. It is important to note that a model's parameters that are fixed on the basis of a certain set of data are empirical parameters. It is not clear at all how their conclusion could be extrapolated to theories where simplicity does not refer to such empirical parameters but rather to much more abstract theoretical properties.

[11] According to Bovens and Hartmann (2003), "[c]oherence is a property of an information set that boosts our confidence that its content is true *ceteris paribus* when we receive information from independent and partially reliable sources". However I believe this idea is much closer to the idea of robustness (see Introduction) than to the property of coherence.

[12] More recently it has been denied that fertility is a virtue in its own right (Nolan 1999). This in turn has been criticized (Segall 2008).

[13] Note that this is also the strategy of one of the most recent writings on theoretical virtues. Okasha (2011), for instance, applies Arrow's impossibility theorem—developed in the realm of social-choice theory—to the problem of theory-choice and draws a conclusion that is more pessimistic than the conclusions drawn by Kuhn (1977) in his classic piece on the topic: there are 'no algorithms' rather than 'too many algorithms' for theory-choice. Okasha simply adopts the five criteria of theory-choice identified by Kuhn without trying to further illuminate them.

theoretical virtues, will appear undesirable too (making the above a true dilemma). But the onus of making them a more integral part of our theories of confirmation, I believe, is certainly preferable to the other option.

2.4 General concerns

Before we can start to discuss examples for the application of TDRs and the postulation of undetected error sources in situations of data conflicts, we must briefly consider three general concerns.

First, if scientists hold different theories and disagree on which theory is better supported by the evidence, then they may be guided towards different reliability judgments. In fact, such a scenario would be an instantiation of the second form of theory-ladenness introduced at the beginning of this essay. Second, one may be concerned that, if the postulation of undetected error sources were methodologically legitimate, theories could no longer be disconfirmed. After all, all counter-evidence might simply be disqualified as unreliable by postulating undetected error sources. As to the first concern, it seems reasonable to think that scientists, even if disagreeing on the degree of inductive support of theories, should *mutatis mutandis* prefer those theories that happen to postulate fewer undetected error sources in experiments than others. Similar remarks apply to the second concern. Also here, one would think, scientists should abandon theories that keep postulating undetected error sources with little confirmatory evidence to show for them in return.

The third concern is that TDRs, when combined with the postulation of undetected error sources smack very much of theoretical bias, which is generally regarded as detrimental to objective scientific work. However one must distinguish between two kinds of theoretical bias. There is one kind theoretical bias which one may refer to as theoretical bias of a psychological kind. A theoretical bias of a psychological kind means that data analysis is selective in such a way that it favors the theory that the scientist in question *happens to hold*. The causes for theoretical bias of this kind have to do with the scientist in question having put all her bets on the theory in question, having heavily invested work in this theory, having defended this theory publicly as being true, etc. The cause for her theoretical bias is therefore entirely psychological. Although understandable, this kind of theoretical bias is usually detrimental to scientific objectivity. Psychological concerns simply should not influence data analysis and theory confirmation. But the concept of TDR has got nothing to do with this kind of bias. Contrary to theoretical bias of the psychological kind, the 'bias' implied by TDRs is entirely epistemic. Whereas there are no good reasons for a bias of the psychological kind, as judged from an epistemological perspective, there are indeed good reasons for applying TDRs. These reasons, as pointed out above, are of an inductive kind. When applying TDRs, one relies on the empirical support the theory has received from explaining another set of data. It is the independent support the theory driving TDRs has thus gained that gives one inductive reasons that the theory in question is correct also in the present case. Although these reasons are of course fallible, they are of an entirely different nature then the reasons one has when having a theoretical bias of the psychological kind. Again, the former may legitimately figure in data assessments, whereas the latter shouldn't.

That the examples to be discussed in the next section are examples for TDRs and not for psychological bias is indicated by the fact that the majority of the scientific community appears to have followed the guidance of TDRs. Although it is not impossible that the majority of the scientific community developed a theoretical bias of a *purely* psychological kind, to the adherent of science being a rational enterprise, this must be an utterly unlikely possibility.

3 Resolving data conflict with TDRs: examples

The first example I want to discuss concerns electron-atom collision experiments conducted by J. Franck and G.L. Hertz in 1914 (Hon 1989). In order to construct a "kinetic theory of electrons" and, in particular, in order to distinguish between elastic and inelastic collisions between electrons and atoms, Franck and Hertz tried to determine the kinetic energy of electrons transferred to atoms after their collisions with gas molecules. This they did "in an ingenious two-pronged attack" (Hon 2003) by measuring (i) the energy retained by the accelerated electrons after colliding with the gas molecules, and (ii) the frequency of light emitted by the molecules after bombardment. Frank and Hertz's results exhibited a regular rise and fall of the measured current caused by the electrons after colliding with the gas molecules as a function of the potential accelerating the electrons. The difference between the peaks of this plot gave the critical potential, which they measured at 4.9V and interpreted as ionization potential. On the basis of the quantum relation $h\nu = Ve$ they related this critical potential to the emitted wave length of λ = 2536.7 Å, which they measured when using an electron impact voltage of 4.9V. Their measurements were limited by their apparatus, which did not allow the determination of critical potentials much higher than 4.9V.

Franck and Hertz were not aware of the Bohr model at the time they planned and conducted their experiments, but it was quickly realized by others that their results were inconsistent with the Bohr model; it predicted an ionization energy at an impact voltage of 10.5V rather than 4.9V for mercury molecules. Interestingly, however, other experimenters first *confirmed* Franck and Hertz's alleged ionization results. Perhaps even more importantly, those subsequent experiments, which were less limited than Franck and Hertz's, were not able to measure the light emission lines where Bohr's theory predicted them, namely at an impact voltage of 10.5V. As Bohr and others pointed out, however, if the critical potentials were to be interpreted as excitation potentials rather than ionization potentials, the Bohr model would be consistent at least with the measured impact voltage of 4.9V. But probably because experimentalists were unable to detect the emission line associated with a 10.4V impact voltage, there was strong resistance against Bohr's re-interpretation of the 4.9V result. Indeed, some even concluded that "the theory is invalid". It took several years and sustained efforts from several experimenters until Bohr's interpretation could be sustained and 10.4V be confirmed as the ionization energy of mercury in 1917.

Now let us ask: had these experimental efforts that ultimately led to the confirmation of the Bohr model been made if the Bohr model had not made predictions that first appeared to be contradicted by the experiments? They might of course have made for contingent reasons, but the Bohr model certainly gave the experimenters good epistemic and principled reasons for intensifying their experimental searches and error checks. Since this is so, we here seem to have a good case for the application of TDRs.[14] But was the use of TDRs justified? P- and UN-support are clearly violated. Bohr did use the known spectral line series to construct his theory, which is why on both P- and UN-support Bohr's theory should not have received any, or only little, support from those data. On the V-support criterion, however, known spectral line series did support Bohr's model (at least to some extent). After all, the Bohr model, at the time, was not only the only available explanation for emission and

[14] The TDR that was thus applied ultimately led to the conclusion that Franck and Hertz and many others had committed an error of *interpreting* the critical potential as ionization potential rather than excitation potential (Hon 1989). But in the present context, this is only of secondary interested. As mentioned above, the TDR thesis is indifferent about the *kind* of error that the error searches that are driven by the theories they seek to test ultimately lead to (cf. Section 2 and footnote 6).

absorption spectra of atoms, but it also did a good job at unifying all those data in a simple model. However even if there hadn't been any V-support, it seems, the use of TDRs would have been justified in this case. This is so, as far as I can see, since no undetected error sources were postulated in the judgment about Franck and Hertz's experiments being erroneous. But this was different in the next case we are going to consider: the discovery of the DNA structure in 1953.

In 1952 Rosalind Franklin had discovered that DNA could take two forms: a 'wet' and a 'dry' form (according to the amount of water molecules contained in it). It was pretty much uncontroversial that the x-ray crystallography evidence coming from the B form of DNA, i.e. the 'wet' form, was straightforward evidence for a helical DNA structure. On the other hand, the x-ray crystallographic evidence for the A form wasn't so unequivocal. Indeed, Franklin had produced a picture of the A form in 1952 that, to her and her colleagues at King's College, London, was clearly incompatible with a helical DNA structure[15]. Accordingly, Franklin tried to develop complex models that would accommodate both sorts of evidence through conformation changes. In contrast, Crick and Watson decided that Franklin's negative evidence had to be spurious—without even having seen it![16] After publication of Crick and Watson's double-helix model and Franklin's positive helical result, the scientific community accepted the DNA structure as being confirmed. The negative helical result that had been published earlier was ignored (see Schindler 2008, for details). Contrary to the Bohr example discussed above, we here have an example for the (implicit) postulation of an undetected error source resulting from the application of a TDR. But then, how can we make sense of Crick and Watson's judgment without accusing them of irrational and unscientific behaviour? As argued in Section 2.2, this must proceed by arguing that the theory (here: the model) driving the TDR was independently supported. The best candidate for independent support is Crick and Watson's explanation of E. Chargaff's discovery that in samples of DNA, the bases that constitute DNA (*A*denine, *T*hymine, *G*uanine, and *C*ytosine) exhibit one-to-one ratios (%A = %T and %G = %C). So in what sense were those data independent support for Crick and Watson's model? Was it support in the sense of UN support? Even though Crick

[15] Wilkins (2003, 182) stated recently that "Stokes and I could see no way round the conclusion that Rosalind had reached after months of careful work. It seemed, in spite of all previous indications, that the DNA molecule was lop-sided and not helical."

[16] In a later interview, Crick stated that "When she told us DNA couldn't be a helix, we said, 'Nonsense'. And when she said but her measurements showed that it couldn't, we said, 'Well, they're wrong'. You see, that was our sort of attitude" (Judson 1996, 118).

and Watson later denied the use of Chargaff's data (see Schindler 2008), these data were available and thus in principle use-able. As pointed out above, it is one of the short-comings of the use-novelty criterion that it is difficult to determine whether a particular piece of data was used or not. Are Crick and Watson saying the truth when they say they didn't use it? We can't say for sure. Does this information matter to whether or not their model was or was not a good model of the DNA structure? I think not. So here we then have a case where the UN-support criterion *may* have been violated. If it had been, then Crick and Watson's application of TDRs, resulting in the judgment that Franklin's negative result was unreliable would have been theory-laden (cf. Section 2.2). This sort of indeterminacy we do not face with the V-support criterion. Crick and Watson's DNA model was clearly explanatorily highly virtuous. It not only made perfect sense of Chargaff's *ratios* (which are directly entailed by the A-T and G-C base *pairings* in Crick and Watson's model), but it also suggested an elegant DNA replication mechanism. So not only did the model provide an explanation of already known evidence, but it also suggested further research avenues. Why should the model receive no (or hardly any) empirical confirmation from the former (as P support would have it, and as UN support *may* imply) and why would it not receive at least a fair amount of plausibility from the latter? If we accept that Crick and Watson's model was independently supported (as V-support suggests, but P-support would deny, and UN-support may not be able to determine), then Crick and Watson's application of a TDR and their postulation of an undetected error source may be deemed *reasonably* rational (cf. Section 2.2). In fact, given that Franklin's allegedly antihelical photograph had been publicised in an important journal before Crick and Watson's proposal, it seems fair to say that it was the *community* that applied a TDR when accepting Crick and Watson's model as beings supported by the evidence.[17]

In my third example, which I want to discuss in some length, concerns the confirmation of Einstein's general theory of relativity through the British solar eclipse expeditions in 1919, which is known as one of the most impressive confirmations of a theory

[17] It is of course no requirement of the TDR thesis that *all* scientists use theory to disambiguate the evidence in the way envisioned by the TDR thesis. R. Franklin and her colleagues, for instance, contrary to Crick and Watson refrained from applying TDR judgments. This indeed stymied them. As Franklin's colleague Wilkins later remarked with respect to the seemingly antihelical evidence: "*[o]ur main mistake was to pay too much attention to experimental evidence*. Nelson won the battle of Copenhagen by putting his blind eye to the telescope so that he did not see the signal to stop fighting" (Wilkins 2003, 166; my emphasis). See also the Millikan-Ehrenhaft debate (below).

in the history of science. However, in a remarkable paper that re-analysed this episode, Earman and Glymour (1980) conclude:

> The British results, taken at face value, were conflicting and could be held to confirm Einstein's [general] theory only if many of the measurements were ignored. (Earman and Glymour 1980, p. 51)

Two expeditions (one to Sobral in Brazil, one to the island Principe off the West African coast) produced three data sets that can be categorized according to the sorts of telescopes that were used: (1) the Sobral 4 inch, (2) the Sobral astrographic and (3) the Principe astrographic. Whereas data set 1 (1.98''±0.12'') and set 3 (1.61''±0.30'') were roughly consistent with the prediction of Einstein's theory (1.75''), data set 2 (0.93'') was inconsistent with Einstein's theory but very much in agreement with the prediction of Newton's theory (0.87''). In the light bending measurements, as in so many experiments in physics, the physical "signal" of interest had to be discerned from a background effect. In the present case this background effect was a "change of scale" in the recorded pictures of the starfields due to an instrumental artifact (in particular: an accidental change of focus) that could be caused by temperature differences. Since a change of scale would have manifested itself in a change of focus in the recorded images, and since such a change of focus (or at least a blurring; see below) was indeed what the images of data set 2 exhibited, it was this type of error that Dyson et al. speculated was responsible for the low light bending result in data set 2 (p. 74). Dyson et al. (1920) therefore discounted this result in their published report in the *Philosophical Transactions of the Royal* and did not even mention it in their presentation at the "Joint eclipse meeting of the Royal Society and the Royal Astronomical Society" in 1919[18]. Earman and Glymour have objected to this discounting of data set 2 for the following reason. Since data set 3 was "the worst of all" (p. 74), Earman and Glymour argue, "it is hard to see decisive grounds for dismissing one set [i.e. set 2] but not the other [i.e. set 3]" (p. 75). And although Earman and Glymour deem data set 1 as "much more impressive" than 2 and 3, they remark with respect to Einstein's theory that "the mean value [of data set 1] is too high and the dispersion too small" (ibid.). Effectively, Earman and Glymour point out, data set 1 was no better evidence for Einstein's theory than data set 3 was for Newton's theory. In other words, interpreting data 3 in favour of Einstein's theory (rather than Newton's theory) was to some extent arbitrary. Furthermore, even though the

[18] Cf. Thomson (1919).

value for star light deflection determined from data set 1 was significantly higher than the one predicted by Einstein, it was interpreted as confirming Einstein's theory. Earman and Glymour infer that physicists presupposed a "trichotomy of possible results" that they never argued for: the results would either indicate no starlight deflection, a deflection consistent with Newton's theory, or a deflection with Einstein's theory. Only then, and under the supposition of data set 2 being somehow erroneous, "the results had to be viewed as confirmation of Einstein's prediction" (pp. 79-80). Earman and Glymour reach a similarly gloomy verdict on Einstein's redshift prediction. They conclude: "If one were willing to throw out most of the data, one could argue that Einstein's prediction was confirmed" (p. 51).

Recently, Earman and Glymour's conclusions have been challenged by Kennefick (2009). Kennefick is concerned in particular with the idea that data set 2 was thrown out because the physicists analysing the data were biased toward Einstein's theory of relativity. Kennefick defends the British physicists with three arguments. First, Kennefick acknowledges that Eddington took a "theory-centric approach to data analysis", but he deems this irrelevant to the question of whether or not the analysis of data set 2 was biased, since he contends that it was Dyson, not Eddington, who analyzed data set 2.[19] But Dyson, Kennefick argues, was "moderately skeptical" about Einstein's theory (2009, p. 40). The fact that Dyson was not positively inclined towards Einstein's theory, Kennefick takes to be at least *prima facie* evidence against the idea that the data analysis was carried out in such a way that it was unduly favorable to Einstein's theory. Kennefick's second argument is his most sophisticated. It runs as follows. Kennefick stresses that the British physicists were hesitant as to whether the "effects of the sun's heat on the mirror" caused a "real change of scale or merely blurred the images" (Dyson et al. 1919, p. 309). As mentioned above, a change of scale would have implied a significant experimental error, which then would have to be corrected for. The British physicists in fact considered both possibilities in their paper. Whereas they obtained the aforementioned low value when assuming a change of scale, they calculated a value of 1.56'' when assuming that the instruments were working fine. The latter value, Kennefick claims, would have been "not far off Eddington's Principe result" (Kennefick 2009, p. 42). Kennefick concludes that

> Support for the Newtonian theory was thus, in some sense, logically incompatible with the instruments having behaved in the intended manner. I

[19] There is no unequivocal evidence for this claim, but Kennefick provides persuasive circumstantial evidence.

> suspect that line of argument strongly influenced the Greenwich team's decision to exclude the astrographic data from their final report (ibid.).

And because Eddington, too, had come to the conclusion that no real change of scale occurred in *his* eclipse recordings at Principe under very similar climatic conditions, the assumption of "no significant change of scale" at Sobral was a plausible assumption. Nevertheless, Dyson et al. dismissed data set 2 because they were not able to determine beyond doubt whether the plausible assumption of "no change of scale" was *actually* correct. Third, Kennefick seeks to rehabilitate the British physicists for their "throwing out" of data set 2 by citing a re-analysis of the eclipse experiments in 1979, which seems to retrospectively vindicate the assumption that the loss of focus in the images of data set 2 did *not* cause a change of scale but mere blurring, for the re-analysis reproduced a value of 1.55''±0.34'', very close to the value estimated by Dyson et al. Hence Dyson et al. were correct to suspect that a change of scale had not occurred; their decision to exclude data set 2 on the suspicion that it *might* support Einstein's theory was therefore reasonable.

I find Kennefick's arguments unconvincing. With regard to Kennefick's first argument, it must again be emphasized that neither Eddington, *nor* Dyson mentioned data set 2 when presenting the results of their expeditions to the physics community (cf. Earman and Glymour 1980, p. 77). And they *both* mentioned but dismissed data set 2 in their detailed published report (Dyson et al. 1920). So the question of whether or not Dyson was sceptical towards Einstein's theory when *analysing* the data seems of little importance to whether or not he took a charitable stance towards Einstein's theory, when it really mattered, namely in the *justification* of Einstein's theory. And here Dyson took a very firm stance indeed: "After a careful study of the plates I am prepared to say that *there can be no doubt that they confirm Einstein's predictions*. *A very definite result* has been obtained that light is deflected in accordance with Einstein's law of gravitation" (cited in: Earman and Glymour 1980, p. 77; my emphasis). Now whether this is a problematic statement of course hinges on whether Dyson et al. had good reasons for dismissing data set 2. So did they? It seems not. To see this, let us first assess Kennefick's third claim, namely the claim that the 1979 data re-analysis provided an "after-the-fact-justification" for Dyson et al.'s exclusion of data set 2.

First note that it is not unproblematic to cite a data re-analysis as justification for the neutrality of an original data analysis with respect to a theory that had been just proposed, when the re-analysis was carried out at a time when the theory potentially causing bias has been accepted for decades and when the people carrying out the original analysis have become celebrated heroes. If anything, the bias towards a result compatible with the theory must be exceedingly more likely for the re-analysis than for the original analysis. This is no mere logical possibility. In fact there is a clear indication that the re-analysis was indeed biased towards Einstein's theory. A re-analysis has got to make a decision about whether an actual change of scale had occurred or not. The re-analysis Kennefick is citing, it appears, simply *assumes* (without argument) that a change of scale did not occur. It is therefore not particularly surprising that they produced a result that coincides with Dyson at al.'s "no change of scale"-estimation. But such a re-analysis does nothing to justify precisely what is at stake, namely the assumption *that* no change of scale had occurred.

Let us now consider Kennefick's second point. For the sake of the argument let us assume that the reasoning Kennefick attributes to Dyson et al. is accurate and Dyson et al. felt licensed to exclude data set 2 from their considerations because (i) if the instruments had worked in the intended manner, the data would have supported Einstein's theory, (ii) the instruments at Principe (producing data set 3) in fact worked in the intended manner, and (iii) it was plausible to assume that the instruments producing data set 2 should behave just as the instruments at Principe. But all of these premises are problematic. First, it is at least questionable whether a deflection of 1.52'' would have been real evidence for Einstein's theory.[20] Second, with regard to premise (ii) and (iii), one must refer to Earman and Glymour's (1980, p. 79) observation that the images of data set 3 were in fact *more blurred* than the images of data set 2! Given that a blurring of the images was likely to be caused by a change of focus, which in turn would imply a change of scale rather than star light bending, Kennefick's chain of reasoning can be turned around. Rather than inferring the truth of the "no scale assumption" for data set 2 from data set 3, one might question the accuracy of the "no scale assumption" for data set 3 on the basis of the doubtfulness of this assumption in the analysis of data set 2. And then, one might want to ask whether it is legitimate at all to draw any inferences between the two data sets, as Kennefick supposes. At the very least, it's not clear that one can. In any case, nothing in the original publications suggests that Dyson et al. drew any such inference. On the contrary, they treated data set 2

[20] Recall that Earman and Glymour point out that the value inferred from data set 3 was not in much better agreement with Einstein's theory than with Newton's, and that the value for data set 2 is even lower than the one of data set 3.

and 3 as being of *different* quality. Whereas data set 3 was argued to be reliable, data set 2 received a detrimental assessment:

> The results obtained with a similar instrument at Sobral are considered to be largely vitiated by systematic errors. (Dyson et al. 1920, p. 330)

This is awkward. The possible systematic error Dyson et al. mention is the change of scale. And as we saw above, conceding a change of scale would have implied a result that was inconsistent with Einstein's prediction. Apparently Dyson et al. conceded a change of scale and then simply dismissed the whole data set (by postulating an undetected error source) *because* the reduced data would have implied a result inconsistent with Einstein's theory.

This brings us to the perhaps most important reason why Kennefick's defense of the British eclipse physicists fails: his arguments are simply ineffective against Earman and Glymour's critique. Earman and Glymour are well aware of Dyson's moderate skepticism[21] and they are also well aware of the attempted justification Dyson and Eddington gave when throwing out data set 2. Alas, *none* of this bears on their critique that "it is hard to see decisive grounds for dismissing one set [i.e. data set 2] *but not the other* [i.e. data set 3]", given that data set 3 was the "the worst of all" (p. 74-5; added emphasis). Earman and Glymour's assessment receives independent support from the physicist C.W.F. Everitt (1980), who wrote that data set 2 "was thrown out though the evidence for them [i.e. data set 2] was much better than that for the 1.61±0.30 arc-sec measurement at Principle", i.e., data set 3. Indeed, Earman and Glymour cite Dyson et al.'s contemporary, the American astronomer W. Cambell, who in 1923 wrote: "as the few images on his small number of astrographic plates [at Principe] were not so good as those on the astrographic plates secured in Brazil, and the results from the latter were given almost negligible weight, the logic of the situation does not seem entirely clear" (p. 29). Even Kennefick (2007) himself, in a more extensive discussion of the case, concedes that data set 3 was so "meager" that "another experimenter [than Eddington, who gathered and analyzed the results] would have been tempted to discard [it] altogether". So the real question then must be: why did the British physicists discard data set 2 rather than data set 3, even though data set 3 was no better (perhaps worse) than data set 2? This is the question that Earman and Glymour really

[21] In fact, Earman and Glymour (1980) explicitly mention that Dyson thought that Einstein's theory was "too good to be true" (p. 85). But again, this should not be read as strong, but only as *moderate* scepticism. Dyson was "deeply interested" in Einstein's theory and showed "enthusiasm" for the carrying out of the eclipse expedition (cited in: Kennefick 2009, p. 40).

posed. Kennefick provides no answer to this question. Perhaps even more interesting is the question why the physics *community* took Einstein’s theory to be confirmed by the positive light bending evidence and why they accepted that data set 2 was to be ignored.

So here now finally is how the concept of TDR makes sense of this historical episode. The TDR thesis, recall, says that independently well supported theories may drive scientists to treat as reliable those results that are consistent with the theory and as unreliable those results which are not. And indeed, what the above discussion shows is that the data consistent with Einstein’s theory (data set 1 and 3) were regarded as reliable and the data inconsistent with it (data set 2) were not. In agreement with the TDR thesis, Dyson et al. sought to bring in agreement with the theory those data that contradicted it (data set 2).[22] This they did by questioning the most natural assumption of a change of scale being the cause for the change in focus. Although this would indeed have brought the result *closer* to Einstein’s prediction of 1.75’’, 1.52’’ was still considerably below it. Faced with a choice of interpreting data set 2 as a result *against* Einstein’s theory or as a result at most loosely compatible with it, they chose to entirely dismiss the sample as unreliable. In contrast, there were apparently no efforts to undermine data set 3, even though it was qualitatively no better (or even worse) than data set 2. Both of these observations are thus explained by the TDR thesis. Another observation by Glymour and Earman can be explained. As they point out, the British physicists presupposed a trichotomy of possible results: results consistent with Einstein’s theory, results consistent with Newton’s theory, or no star light deflection. Crucially, the star light deflection inferred from data set 1, Earman and Glymour explain, was interpreted as falling into the first category *even though* the measured value was significantly higher than the prediction by Einstein. Indeed this continued to be so for decades of light bending measurements (von Kluber 1960). The fact that the physics *community* accepted these results as evidential support for Einstein’s theory must be explained by the supposition that the excess magnitude of the measured light bending values might be due to experimental artifacts rather than being expression of the true physical causes. This supposition is implied by the TDR thesis. What remains to be clarified is the question, what the independent empirical support for Einstein’s theory of general relativity was in 1919. The only real candidate is Einstein’s explanation of Mercury’s

[22] In fact, the raw data of data set 2 indicated a light deflection of 0.86’’ rather than 0.93’’. The latter value was gained only after restricting the analysis to a few particularly bright stars.

perihelion, which had remained unexplained for decades in the Newtonian paradigm and which was therefore perceived as major feat (cf. Brush 1989). Indeed Dyson et al. (1919, 1920) explicitly mentioned this in the presentation of their results. Again, the advance of Mercury's perihelion lends support to Einstein's theory only under V-support: it clearly was no temporally novel prediction (thus violating P-support), and it is now known that Einstein did indeed try to accommodate it when constructing his theory (Earman and Glymour 1978), thus violating UN-support.

Another example that can be cited in illustration of the disambiguation of data conflicts through TDRs is the notorious Millikan-Ehrenhaft dispute about the measurement of charge of electrons in the 1910s (Holton 1978). Although Millikan and Ehrenhaft conducted similar experiments, in which oil drops (Millikan) or colloidal metal particles (Ehrenhaft) were suspended in an electric field, they obtained very different results. Whereas Millikan's measurements revealed integer multiples of 1.592×10^{-19} coulomb, Ehrenhaft's results were much messier. They indicated a *continuous* spectrum of charges. Why Ehrenhaft produced the data that he did, however, was never fully established. As late as 1940 Albert Einstein wrote "[c]oncerning his [Ehrenhaft's] results about the elementary charge I do not believe in his numerical results, but I believe that nobody has a clear idea about the causes producing the apparent sub-electronic charges he found in careful investigations" (Holton 2008).[23] Despite the reasons for Ehrenhaft's results being obscure not only by contemporary but also by later standards, Millikan's experiments were accepted from 1913 onwards as having discovered the charge of electrons.

Previous discussions of the Millikan-Ehrenhaft dispute have focused very much on whether Millikan's conduct in his data reduction was proper or not. In particular, discussion has centered on the question of whether Millikan illegitimately discarded data (mentioned in his laboratory notebook but not in his published papers) which did not fit his theoretical prejudices (Franklin 1981; Niaz 2005). These discussions, however, have little to say about the question of why the physics community accepted Millikan's results as establishing the charge of the electron, when Millikan's results were clearly contradicted by Ehrenhaft's results and when the reasons for Ehrenhaft's results were unknown. Should one not have

[23] Other such quotes can be found in Holton (1978) and Hon (1985). Of course there speculations about what went wrong. One of the charges against Ehrenhaft was that he did not use a corrected version of Stokes's law, as Millikan did. However this correction presupposed that charges come in integral multiples, which was exactly what was at stake. This, in fact, Ehrenhaft himself pointed out to his critics (cf.Hon 1985).

expected that the community would have suspended their judgment about whose results were correct as long as they had no grasp on that matter? Again, the most plausible (epistemic) explanation for the physics community siding with Millikan's rather than with Ehrenhaft's results is that the physics community was being led by TDRs in their judgments about which results to lend more credence to. Although there was no fully-fledged theory about electrons at the time, the idea that electrons should come in integral units of charge was an idea that had been made plausible by J. J. Thomson in his explanation of cathode rays roughly 15 years before Millikan's measurements (cf. Falconer 1985). Indeed, if there are discrete electrons then there *must* be integral units of charge (and multiples thereof) and there *cannot* be a continuum of charges. It is plausible that not only Millikan but also the majority of the physics community had this in mind when weighing Millikan's against Ehrenhaft's data. In other words, it is plausible that the physics community applied TDRs when assessing those data, i.e., TDRs that were driven by the hypothesis of the electron as a discrete unit of charge. This hypothesis, in turn, had received its most significant support from its application to the phenomenon of cathode rays by J.J. Thomson in 1897. Again, if that hypothesis had not received any support from this application (as P- and UN-support would have it), then the use of TDRs in order to assess Ehrenhaft's data would have been unwarranted. So, once again, in order to make rational sense of this historical episode, we are forced to accept that the cathode ray experiments lent support to the electron hypothesis (on the basis of the V-support criterion). Indeed, Millikan himself assumed that J.J. Thomson had (in Millikan's own words) achieved an "unambiguous establishment of the electron theory of matter" (cited in Holton 1978, 40).

Although it is by no means a requirement for the soundness of the TDR thesis that scientists are aware of their use of TDRs, they sometimes are, as the following case shows. In the 1970s several experiments were performed to test the prediction of parity violation in weak interactions as entailed by the Salam-Weinberg-Glashow (GWS) model. In 1976 atomic physics experiments (with bismuth) at the Universities of Oxford and Washington were unable to confirm parity violation. In early 1978 another experiment on parity violation in the higher energy ranges was carried out at the Stanford Linear Accelerator Center (SLAC). E122, as it was called, was the first (and last) high energy physics experiment on parity violation. Contrary to the atomic physics experiments at Washington and Oxford, it confirmed the GWS prediction. Pickering (1984) has pointed out that with E122 the assessment of the Washington and Oxford experiments changed dramatically, even though "there was no *intrinsic* change in the status of the Washington-Oxford experiments", since "[n]o data were withdrawn and no fatal flaws in the experimental practice of either group had been proposed" (301).[24] But E122 had no direct implications for the atomic parity experiments. As Franklin mentions, the atomic parity violation experiments "test the [GWS model] in a very different energy range and test different electron-quark couplings from the high energy physics experiments" (Franklin 1990a, 189). However later in 1978 also positive evidence for *atomic* parity violation emerged in experiments carried out by Soviet physicists from Novosibirsk. Another positive result was published in early 1979 by a group at Berkeley. Although one might have thought, with Franklin (1990b, 1990a), that the Washington-Oxford experiments and the results from Novosibirsk and Berkeley "neutralized" each other, with a suspension of judgment about atomic parity violation being correct or not, this is not what happened. As Pickering (1990) points out, rather awkwardly, several research reviews favored those results that confirmed the GWS theory, without there being clear experimental grounds on which those results could have been preferred. On this rather shaky basis, Dydak (1979), for instance, concluded that the results from Novosibirsk and Berkeley are "in good agreement with the standard model" (13).

In an attempt to explain the dismissal of the Washington-Oxford experiments, despite the failure to detect any errors in them, Franklin (1990b, 1990a) provides the following arguments. First, Franklin stresses the presence of systematic uncertainties in the Washington-Oxford experiments. Second, although conceding that it was never established why the Washington-Oxford experiments gave negative results, Franklin *speculates* that the early atomic parity violation experiments contained *unknown* systematic uncertainties and claims that the later atomic physics parity violation experiments that were carried out in the early 1980s (including re-runs by the Washington-Oxford groups with a different experimental apparatus) which did indeed confirm the GWS's prediction, were "more accurate" than the early ones. Third, Franklin notes that the results published by the Washington and Oxford groups in a joint preliminary report and in more detailed separate

[24] Pickering furthermore points out that E122 was a rather peculiar experiment (at least at the time) in that it was performed once and was never replicated. If replicability is a good guide to the reliability of experimental results (as Franklin stresses himself in other contexts), then the results of E122 were not immune to being subject to doubt.

publications were "mutually inconsistent" and were therefore looked upon with suspicion. None of these arguments, however, is convincing.

First, systematic uncertainties are nothing special. E122 had to deal with them too.[25] And, rather curiously, Franklin discusses the ways in which systematic uncertainties were controlled in the latter, but not in the former experiments (Pickering 1990). Second, Franklin's speculation that the early Washington-Oxford experiments contained unknown systematic uncertainties is really not more than that: a speculation. For the claim that the latter atomic parity experiments were more accurate than the early Washington-Oxford experiments, Franklin cites merely their alleged inconsistency. This brings us to Franklin's third argument. Here two points can be made. First, the fact that the preliminary report by the Washington-Oxford group was significantly different from their later more detailed reported experiments ought not to give one too much of a headache. After all, their first report was preliminary. It always takes time to fine-tune an experimental apparatus. Franklin claims that the physics community was troubled by this, but he provides no evidence for this claim apart from a private conversation with a single physicist who "recalls such discussion" (Franklin 1990b, 168). Second, if the consistency of results was really as epistemically meaningful as Franklin has it, physicists should have dismissed the positive results from Novosibirsk (1.07±0.14) and Berkeley ($2.3^{-1.4}_{+3.1}$) rather than the results from Washington (0.0-0.2) and Oxford (0.0-0.1). Indeed one must be doubtful about the reliability of the results from Novosibirsk and Berkeley: the latter was only two standard deviations from zero, and the Novosibirsk group not only had a dubious track record (Pickering 1990, 462), but would ultimately report a result of about twice the value (-20.2±2.7) obtained by the later experiments by Oxford and Washington (9.3±1.5 and 10.4±1.7, respectively). Incidentally, these final results on atomic parity violation were, contrary to Franklin's assertion, everything but mutually consistent.

The TDR thesis, contrary to Franklin's and Pickering's accounts,[26] provides a plausible explanation for the history of experiments on parity violation. First, the TDR thesis explains why physicists performed further experiments that would confirm the GWS atomic parity

[25] One way of dealing with uncertainties is the replication of the results with slightly different experiments. But as mentioned in the previous footnote, the results from E122 were never replicated.

[26] Pickering (1984, 301-2) tries to explain this episode in terms of his concept of 'social symbiosis', according to which theorists and experimentalists drew on each other's work in order to justify their respective practices. This concept, however, does not explain why the physics community went along with the GWS model rather than with its possible alternatives. See also (Schindler under review) forthcoming.

prediction, despite their initial negative results. Second, the TDR thesis explains why, although E122 had no direct bearing on the atomic parity violation data, it fuelled the TDRs derived from the GWS model that demanded parity violation also in the atomic realm. This explains why the physics community took a skeptical attitude towards the Washington-Oxford experiments particularly after E122. Lastly, the TDR thesis explains why Washington-Oxford experiments were deemed unreliable despite no errors being found. That is, they were deemed unreliable because the inductive support (indeed P-support, in this case) of the GWS model gave physicists good reasons for such doubts. Fittingly, Steven Weinberg has commented on this episode that

> [the GWS model's] naturalness was being used to help physicists weigh conflicting experimental data. (Weinberg 1994, 100)[27]

Remarks like Weinberg's, which very much suggest at least some awareness of TDRs in the scientific community, can even be found in the original publications. When Sudarshan and Marshak (1957) and Feynman and Gell-Mann (1958) long before the GWS model simultaneously proposed their V-A theory of weak interactions, not less than four experiments contradicted it. Sudarshan and Marshak (1957) demanded that "all of these experiments should be redone" and Feynman and Gell-Mann (1958), after emphasizing the theoretical elegance of their model[28], stated that

> These theoretical arguments seem to the authors to be strong enough to suggest that the disagreement with the Helium 6 recoil experiment and with some other less accurate experiments indicates that these experiments are wrong. (ibid. p. 198)

And indeed, driven by the virtuous properties of the V-A theory of weak interactions, physicists re-did the relevant experiments, re-checked various potential confounding sources, and eventually produced the desired results (Franklin 1990b).

4 Conclusion

In this paper I invigorated the time-honoured concept of theory-ladenness with the introduction of the notion of theory-driven-data reliability judgments (TDR). According to

[27] In the context of theory-choice and the Copernican system McMullin (1993) has argued against T. S. Kuhn that the "naturalness" of a theory is to be equated with the theory's coherence. I believe that is what Weinberg too has in mind here.

[28] Feynman and Gell-Mann name the following internal properties: universality, symmetry, parity violation, conservation of leptons, preservation of invariance under CP and T, and the simplicity of the model. This case was brought to my attention by Allan Franklin. He discusses the case in detail in Franklin (1990).

the TDR thesis the assessment of data is guided by the very same theories these data are supposed to test. Although not epistemologically problematic *per se*, TDRs do imply theory-ladenness of the problematic kind when their application leads to the postulation of undetected error sources in those experiments which the theory driving TDRs is inconsistent with. I argued that the threat of theory-ladenness can be averted if our notion of independent support allows for a theory to be supported by evidence that it neither anticipated nor explained in a use-novel way. According to the notion of independent support proposed here, a theory may receive strong independent support from evidence that it explains in a virtuous way.

Acknowledgments
This paper has benefited enormously from the input from several individuals. In particular I would like to thank Allan Franklin, Matteo Morganti, Wolfgang Freitag, Marcel Weber, Steven French, the members of reading groups at the Universities of Konstanz and Aarhus, and altogether ten anonymous referees for reading previous versions of this paper and for providing very detailed feedback. I furthermore am grateful to the audiences at the Department of Philosophy at the University of Konstanz, at the *Jour Fixe* of the Zukunftskolleg at the University of Konstanz, and at the first *Philosophy of Scientific Experimentation* (*PSX*) workshop at the Center for Philosophy of Science at the University of Pittsburgh for their comments on my presentations of remote ancestors of this paper. I feel deep gratitude towards the Zukunftskolleg at the University of Konstanz for providing me with a fellowship that allowed me to work on the ideas developed in this paper.

5 References

Agafonova, N., A. Aleksandrov, O. Altinok, et al. 2010. Observation of a first candidate event in the OPERA experiment in the CNGS beam. *Physics Letters B* 691 (3):138-145.

Bogen, J. 2010. Theory and observation in science. In *Stanford encyclopedia of philosophy (Spring 2010 Edition)* edited by E. N. Zalta.

Bogen, J., and J. Woodward. 1988. Saving the phenomena. *The Philosophical Review* 97 (3):303-352.

Bovens, L., and S. Hartmann. 2003. Solving the riddle of coherence. *Mind* 112 (448):601-633.

Brush, S.G. 1989. Prediction and theory evaluation: The case of light bending. *Science* 246 (4934):1124.

Culp, S. 1995. Objectivity in experimental inquiry: Breaking data-technique circles. *Philosophy of science*:438-458.

Dydak, F. 1979. Neutral Currents. *Proceedings of the International Conference on High Energy Physics, Geneva, 27 June-4 July, 1979*:1-25 (pages refer to online document).

Dyson, F.W., A.S. Eddington, and C. Davidson. 1920. A Determination of the Deflection of Light by the Sun's Gravitational Field, from Observations made at the Total Eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London. Series A* 220 (571-581):291-333.

Earman, J., and C. Glymour. 1978. Einstein and Hilbert: Two months in the history of general relativity. *Archive for History of Exact Sciences* 19 (3):291-308.

— — —. 1980. Relativity and eclipses: The British eclipse expeditions of 1919 and their predecessors. *Historical Studies in the Physical Sciences* 11 (1):49-85.

Everitt, CWF. 1980. Experimental Tests of General Relativity: Past, Present and Future. *Physics and Contemporary Needs* 4:529-555.

Falconer, I. 1985. Theory and experiment in J.J. Thomson's work on gaseous discharge, University of Bath, Bath.

Feynman, R.P., and M. Gell-Mann. 1958. Theory of the Fermi interaction. *Physical Review* 109 (1):193.

Fodor, J. 1984. Observation reconsidered. *Philosophy of science*:23-43.

Forster, M., and E. Sober. 1994. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science* 45 (1):1-35.

Franklin, A. 1981. Millikan's published and unpublished data on oil drops. *Historical Studies in the Physical Sciences* 11 (2):185-201.

— — —. 1990a. Do Mutants Have to Be Slain, or Do They Die of Natural Causes?: The Case of Atomic Parity Violation Experiments. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, edited by L. Wessels, A. Fine and M. Forbes: University of Chicago Press.

— — —. 1990b. *Experiment, right or wrong*: Cambridge University Press.

— — —. 2002. *Selectivity and discord: two problems of experiment*: University of Pittsburgh Press.

Gardner, M.R. 1982. Predicting novel facts. *British Journal for the Philosophy of Science*:1-15.

Giles, J. 2006. The trouble with replication. *Nature* 442 (7101):344.

Hacking, I. 1983. *Representing and intervening*: Cambridge Univiersity Press.

Holton, G. 1978. Subelectrons, presuppositions, and the Millikan-Ehrenhaft dispute. *Historical Studies in the Physical Sciences* 9:161-224.

Holzman, D. 2009. Tamoxifen, antidepressants, and CYP2D6: the conundrum continues. *Journal of the National Cancer Institute* 101 (20):1370-1371.

Hon, G. 1985. On the concept of experimental error. Ph.D., University of London, London.

— — —. 1987. H. Hertz:'The electrostatic and electromagnetic properties of the cathode rays are either nil or very feeble.'(1883) a case-study of an experimental error. *Studies in History and Philosophy of Science Part A* 18 (3):367-382.

— — —. 1989. Franck and Hertz versus Townsend: a study of two types of experimental error. *Historical studies in the physical and biological sciences* 20 (1):79-106.

— — —. 2003. From Propagation to Structure: The Experimental Technique of Bombardment as a Contributing Factor to the Emerging Quantum Physics. *Physics in Perspective (PIP)* 5 (2):150-173.

Janssen, M. 2002. Reconsidering a scientific revolution: The case of Einstein versus Lorentz. *Physics in Perspective* 4 (4):421-446.

Judson, H.F. 1996. *The eighth day of creation: makers of the revolution in biology*. New York: Cold Spring Harbor Laboratory Press.

Kennefick, D. 2007. Not Only Because of Theory: Dyson, Eddington and the Competing Myths of the 1919 Eclipse Expedition. *Arxiv preprint ArXiv:0709.0685*.

— — —. 2009. Testing relativity from the 1919 eclipse—A question of bias. *Physics Today* 62 (3):37-42.

Kitcher, P. 1981. Explanatory unification. *Philosophy of science*:507-531.

Kuhn, T.S. 1977. Objetivity, Value Judgment, and Theory Choice. In *The Essential Tension*. Chicago: University of Chicago Press.

— — —. 1996. *The structure of scientific revolutions*: University of Chicago press.

Lyons, J. 2011. Circularity, Reliability, and the Cognitive Penetrability of Perception. *Philosophical Issues* 21 (1):289-311.

Mayo, D.G. 1994. The new experimentalism, topical hypotheses, and learning from error. Paper read at PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association.

— — —. 1996. *Error and the growth of experimental knowledge*: University of Chicago Press.

McMullin, E. 1976. The fertility of theory and the unit for appraisal in science. *Boston studies in the philosophy of science* 39:395-432.

———. 1993. Rationality and paradigm change in science. *World changes: Thomas Kuhn and the nature of science*:55-78.

Niaz, M. 2005. An appraisal of the controversial nature of the oil drop experiment: Is closure possible? *The British Journal for the Philosophy of Science* 56 (4):681-702.

Nolan, D. 1999. Is fertility virtuous in its own right? *The British Journal for the Philosophy of Science* 50 (2):265-282.

Okasha, S. 2011. Theory choice and social choice: Kuhn versus arrow. *Mind* 120 (477):83.

Pickering, A. 1984. *Constructing Quarks: A Sociological History of Particle Physics* Chicago: University of Chicago Press.

———. 1990. Reason Enough? More on Parity-Violation Experiments and Electroweak Gauge Theory. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* edited by L. Wessels, A. Fine and M. Forbes.

Quine, W.V. 1951. Main trends in recent philosophy: Two dogmas of empiricism. *The Philosophical Review*:20-43.

Raftopoulos, A. 2009. *Cognition and perception: how do psychology and neural science inform philosophy?* Cambridge (Mass.): MIT Press.

Rasmussen, N. 1993. Facts, artifacts, and mesosomes: Practicing epistemology with the electron microscope. *Studies in History and Philosophy of Science Part A* 24 (2):227-265.

Schindler, S. 2008. Model, Theory, and Evidence in the Discovery of the DNA Structure. *The British Journal for the Philosophy of Science* 59 (4):619-658.

———. forthcoming. Novelty, Coherence, and Mendeleev's periodic table.

———. under review. A matter of Kuhnian theory choice. The GWS model and the neutral current.

Segall, R. 2008. Fertility and scientific realism. *The British Journal for the Philosophy of Science* 59 (2):237-246.

Steinle, F. 1997. Entering new fields: Exploratory uses of experimentation. *Philosophy of science*:65-74.

Sudarshan, E, and R Marshak. 1957. The Nature of the Four-Fermion Interaction. Paper read at Proceedings of Padua-Venice Conference on Mesons and Newly Discovered Particles, September, 1957.

Suppe, F. 1977. *The structure of scientific theories*: University of Illinois Press.

Thomson, J. 1919. Joint eclipse meeting of the Royal Society and the Royal Astronomical Society, 6 November 1919. *The Observatory, London* 42 (545):389-398.

van Fraassen, B.C. 1980. *The scientific image*. Oxford: Oxford University Press.

von Kluber, H. 1960. The Determination of Einstein's Light-deflection in the Gravitational Field of the Sun. *Vistas in astronomy* 3:47-77.

Weinberg, S. 1994. *Dreams of a final theory*. New York: Vintage Books.

Worrall, J. 1989. Fresnel, Poisson and the ‘White Spot’: The Role of Successful Prediction in Theory-acceptance. The Uses of Experiment, Cambridge: Cambridge University Press.

———. 2002. New evidence for old. In *In the Scope of Logic, Methodology and Philosophy of Science* edited by P. Gardenfors. Dordrecht: Kluwer