

**The New Experimentalism Challenges  
Duhemian Problems:  
*We Have Ways to Make Them [the data] Talk!***

***Deborah Mayo***

***October 15, 2010***

I am pleased this conference invites us to ponder the  
“*Philosophy of Scientific Experimentation:  
A Challenge to Philosophy of Science*”

I think the time is ripe to raise the programmatic  
meta-question:

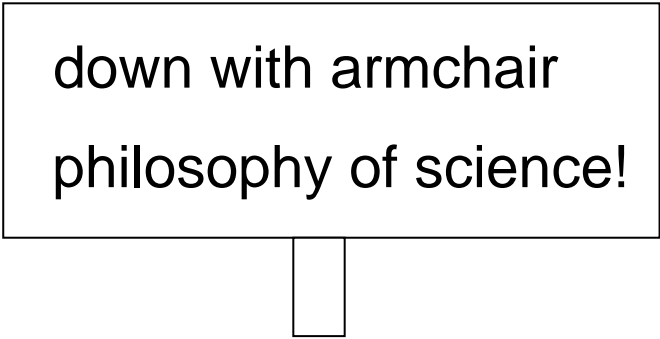
**Has philosophy of science been challenged  
sufficiently by “the new experimental  
movement”?**

I for one do not think it has, and the goal of my talk  
is to try to encourage this group to launch a new  
project aimed at improving things...

Aside: My slides are really just little notes for me, I’m forbidden  
from using power point because I am an artist and can never  
stop making them more (potentially) beautiful...

To those of us who came of age, philosophically, in the early 80's, the “*new experimentalist*” project felt like a new birth of freedom...

We marched with placards,



down with armchair  
philosophy of science!



down with theory domination!

Breaking out from the logical positivist orthodoxy, we firmly resolved to be dynamically engaged with, knowledgeable about, and genuinely *relevant to scientific practice.*

Never mind pronouncements on ideally rational methodology,

*We would examine methodologies of science empirically*

Never mind seeking uniform logics between statements  $E$  and hypotheses  $H$ ,

*We would study how scientists obtain (debate, discard)  $E$  to begin with, and the complex linkages between*

- **data**
- **experiment**
- **theory/hypotheses**

We came to realize:

- accounts that start with neat and tidy evidence statements forfeit being relevant as “a forward-looking” account of how scientists proceed.

Such accounts come up short even as a reconstruction of how scientists learn.

- Most importantly, we came to suspect that understanding the role of experiment is the key to solving long-standing philosophical problems about observation, inference, evidence  
(at least as I saw it)

Take the well-known Duhem problem ...

The challenge posed by “**Duhem's problem**” (roughly) is to provide a principled way to determine which hypothesis should be rejected or disconfirmed when experiment disagrees with  $H$ 's prediction (“the anomaly”).

This task scarcely so daunting once we move away from high-level theory, to the manifold local experimental tasks of distinguishing real effects from artifacts, of checking instruments, and estimating and subtracting out effects of backgrounds.

Many illuminating experimental case studies (by people at this conference) illustrate how scientists successfully deal with Duhemian problems in actual cases.

Some work with cutting-edge modeling tools to discern artifacts, omit confounding factors, (whether machine learning, causal modeling, climate models, experiments to discern types of gravitational radiation, etc., etc.)

But if asked:

*Have we now solved or made progress on  
Duhem's problem, I doubt they will say yes?*

The thinking, perhaps, was that we would come back to the general problems of evidence after being steeped for a while in case studies of practices of various particular sciences.

**If so, the time seems ripe to return to them!**

**The failure to solve these problems has led the vast majority of philosophers of science to give up on them ...**

**Peter Achinstein:**

“scientists do not and should not take . . . philosophical accounts of evidence seriously”  
(2001, p. 9)

because they are based on a priori computations;  
whereas scientists evaluate evidence empirically.

**Alan Chalmers** similarly claims scientists...  
are not in need of advice from philosophers”  
(1999, p. 252),

and the only general things philosophers can say  
are limited to “trivial platitudes” such as “**take  
evidence seriously**” (ibid. p. 171).

The supposition seems to be *if it's empirical it is best left to the scientists*, but **this is a mistake....**

*Scientific inference is too important to be left to the scientists!*

Fresh methodological problems arise in practice surrounding a host of methods and models used to learn from incomplete and uncontrolled, data—often, ironically, in the very fields in which philosophers of science immerse themselves (e.g., psychology, epidemiology, ecology, biology, economics, astrophysics)

Scientists who look to contemporary philosophy of science still find discussions divorced from issues of current practice.

[*Ironically*, they go back to philosophers we were to replace, notably Popper (recently, statisticians, e.g., Cox, Senn, Gelman, Shalizi)]

***What happened? “Was it all a dream?”***

(Worse, was it just my dream?...)

Appreciating the diversity of methods in collecting, modeling, analyzing and interpreting data, and the plethora of criteria in evaluating evidence gives many philosophers of science cold feet when it comes to generalizing...

***But I hope to convince you that we can offer important “systematic” insights and strategies for solving problems...***

[And, by the way, when I speak of solving Duhem, *I do not mean we always pinpoint blame but deny we never can, and deny the particular under-determination allegations thought to stymie learning*]

Solving problems in a manner relevant to practice  
(as I see it) requires

- mounting critiques of actual practice (where warranted)

as well as

- showing how to get more of the kind of knowledge we actually get in science

(science is “successful” but how, and how can we be more successful?)

Going back to the “trivial platitude”

At least *the experimentalist can answer the question of what it means to ‘take evidence seriously’?*

***Clearly, evidence is not being taken seriously if in appraising hypothesis  $H$ , it is predetermined that a way would be found to “save  $H$  from anomaly” (even if  $H$  is false) .***

- Velikovsky’s “scotoma dodge” (Worrall)  
(Any culture showing no records of the cataclysmic events of his theory are suffering collective amnesia.)
- Drug companies refuse to construe any data as evidence of risk, “junk science”.
- Rather than “follow where the evidence leads, they take the data where they want it to go!  
(e.g., in politicized science)

## Minimal (Weak) Requirement for Evidence

*Data  $x_0$  provides poor evidence for  $H$  if it results from a method or procedure that has little or no ability of finding flaws in  $H$ , even if  $H$  is false.*

[While weak, it is stronger than a mere falsificationist requirement: it may be possible to falsify  $H$ , while a procedure makes it virtually impossible to obtain falsifying evidence.]

We (of the trivial platitude) would block attempted “saves” if a procedure had no ability to find or admit flaws in  $H$ .

The basic stipulation can be put in a million ways (Cox’s weak principle of repeated sampling)

If an account of evidence can't do this much, it's hard to see what it has to do with obtaining evidence or grounds for inference, learning or the like...

- One can get considerable mileage even stopping with this negative conception...
- *But note we get the positive converse in the favorite experimentalist argument...*

### **“Experimental Argument From Coincidence”**

for distinguishing “real effect from experimental artifacts”

(Hacking's argument from coincidence to dense bodies)

*A more home-grown variation* (though obviously unrealistic)

If no change in weight registers on any of a series of well-calibrated and stable scales, both before leaving and upon my return from London, even though, say, they easily detect a difference when I lift a .1-pound potato, then we argue from coincidence that the data warrant inferring that my weight gain is negligible within the limits of the sensitivity of the scales.

*H*: my weight gain is no greater than  $\delta$ , where  $\delta > 0$  is an amount easily detected by these scales.

*H*, we would say, has passed a *severe test*: were I to have gained  $\delta$  pounds or more (i.e., were *H* false), then this method would almost certainly have detected this.

Perhaps underdeterminationists would say I could insist all the scales are wrong—they work fine with weighing vegetables, etc.  
(Hacking's Cartesian demon of scales)

## **Rigged alternative $H'$ :**

$H$  is false but all data will be as if it is true.

*All experiments systematically mask the falsity of  $H$*

Were we to deny  $\mathbf{x}_0$  is evidence for  $H$  on account of the *possibility of rigging*, we would be prevented from correctly finding out about weight or whatever...

## **No Rigging! (No Gellerization)**

If the scales work reliably on test objects with known weight, what sort of *extraordinary circumstance* could cause them all to go astray just when we do not know the weight of the test object (can the scales read my mind? (C.S. Peirce)

**It is the learning goal that precludes such rigging, conspiracies, gellerization** — *highly unreliable strategy.*

Depending on the example, this argument from *coincidence* (or argument from no rigging) may only indicate there is a real effect (not an artifact of the experiment)

*We still need to distinguish that inference from various theories to explain the effect*

Much less does evidence for a “real effect” warrant realism (entity realism, or other)

Experimental learning is piecemeal  
(my own favorite way of distinguishing the pieces is by considering what error of inference is of concern).

I might note here that it makes no difference to the task at hand whether an experimentalist is a realist or not: the same Duhemian problems are raised by philosophers on both sides of this divide, (and the same work is required to reliably pinpoint blame).

*Pinpointing Blame:* Suppose then our experimentalist is armed with a minimal principle of evidence and an argument from coincidence

To Duhem's claim that since a scientist "can never submit an isolated hypothesis to the control of experiment," an experimental disagreement with a prediction does not show us the particular hypothesis that is at fault.

The experimentalist says (or should say) that we do not need to actually control all factors in order to pinpoint blame correctly.

**Enough "experimental knowledge" will do!**

It suffices to appeal to a variety of strategies, appropriate for different cases, to:

- **distinguish the pattern of effects** of different factors
- learn enough about observed effects (often by simulations) in order to "**subtract them out**"
- vary methods sufficiently so to rule out artifacts and attain **robustness**
- distinguish between conflicting hypotheses of the cause of an anomaly by **distinguishing how well tested each is.**

It is useful to delineate two distinct types of strategies:

- (a) "**Blocker**" strategies: strategies to block or criticize attempts to explain away anomalies and save a threatened claim or theory.
- (b) **Error-detection strategies** to positively identify the factor responsible.

To Duhem's lament that the experiment teaches us "there is at least one error; but where this error lies is just what it does not tell us" (Duhem 1954, p. 185)

We respond: *that may be true when experiment is approached with "white gloves": we need to become shrewd inquisitors of errors, interact with them, simulate & amplify them:*

"we have to learn to make them talk!"

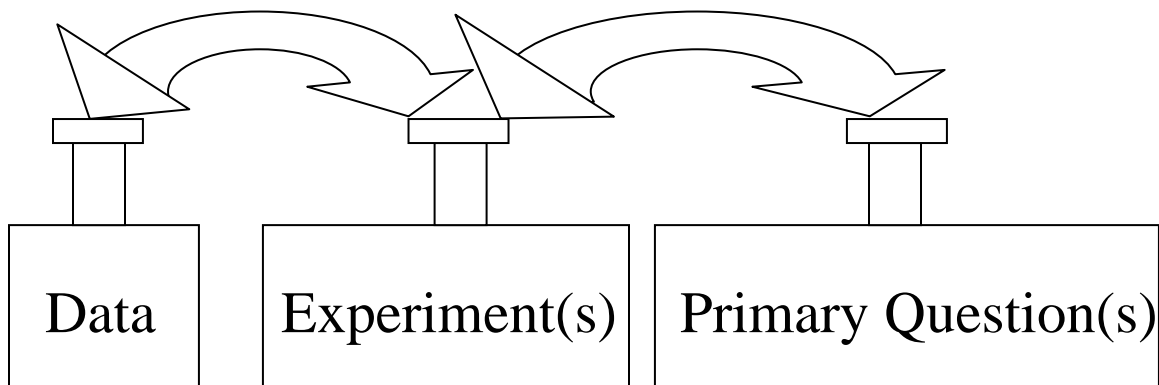
It is easy to arrive at a multitude of hypotheses "to account for" data  $\mathbf{x}_0$ , when it suffices to merely posit that "factor F is responsible for  $\mathbf{x}_0$ ", but things are very different if you must take account of details of the raw data and experimental conditions.

*What kind of framework can we use to reconsider the Duhemian (and other) problems of evidence?*

As Hacking famously put it, “**experiment lives a life of its own**” apart from high-level theory.

*But it should be a real home, not a life in the street: it has its own models, parameters, and theories (whether they exist in nature, on paper, computers, or in our thoughts)*

These homes are often represented as unadorned boxes with arrows coming out the chimneys:



In reality we need to recognize all the work of planning, generating modeling data, linking to questions we can experimentally pose/answer...and how answers are relevant to various primary questions of interest — *would need more houses and chimneys* (but Giere assures me 3 at a time is max).

*However you like to arrange things, the new experimentalist approach lets us split off a cluster of queries.*

In the context of the Duhemian problem, we may group them in two:

- **Is there a real effect?** (*how big?*)
- **Pinpointing its cause?**

If experimentalists were to pursue this, it might require revisiting some well-known episodes.

Everyone's favorite example, Einstein's theory of general relativity (GTR), and "tests" using 1919 eclipse results—even though decent tests had to await 30 plus years.

I'm going to be quite sketchy to get to the business at hand (for details of my take on the last 90 years of testing GTR and rival gravity theories, I can point you elsewhere...)

[The anomaly concerned the deflection of light passing near the sun as was discerned during the 1919 eclipse expeditions undertaken by Eddington and others. The deflection effect, while predicted by Einstein's law of gravitation, was an anomaly for Newton's law of gravitation.]

While two of the three results were deemed clearly anomalous for Newton's gravitational theory, a third pointed, not to Einstein's prediction, but, "with all too good agreement to the 'half-deflection,' that is to say, the Newtonian value . . ." (Eddington 1935, 117).

Yet it was discounted as due to systematic error:

A': The data  $\mathbf{x}_0$  (from Sobral astrographic plates) were due to systematic distortion by the sun's heat, not to the deflection of light

— *no real anomaly for GTR*

A' is the denial of A, the telescope was ok, no serious change of focus in 6 months

The result has sparked some debate: Are there grounds to suspect the experimenters "threw out a

good part of the data and ignored the discrepancies”?  
(Earman and Glymour 1980, 85; S. Weinberg).

Should we conclude that scientists went along with this exception incorporation to promote harmony between German and British scientists?

**No!**

*But this could not be discerned without the details of the data analysis and methodology...*

The data-analytic methods, well-known even in 1919 from stellar parallax experiments, quite independent of GTR

Experimenters knew what unequal expansion of the mirror looks like, etc.

[Results were usable only with a sufficiently precise knowledge of the change of focus plates taken during the eclipse and those taken months later]

***That’s just what was absent here—no usable estimate of error.***

Some philosophical accounts would view this “save” on par with Velikovsky’s save, at least with respect to just the data  $\mathbf{x}_0$  (without further tests)

So both would be guilty of rigging, and violating (“novelty”) requirements for evidence (e.g., Worrall there’s at most “conditional support”)

*I think it is important to distinguish these “saves”...*

To make my point, suppose (unlike what happened) that Dyson and Eddington always said “blame the mirror” — even if by chance this time it really was the mirror — we would say their procedure did not warrant the inference to discounting the anomaly!

(broken watch is right twice a day?)

*So can’t just look at a data set, need to know something about the procedure generating it (it’s overall reliability, or probative capacity).*

Admittedly, what's being inferred in this case is quite weak: *These plates, on which the purported GTR anomaly rested, were ruined!*

That's all that's needed for this piece.

**To clarify:**

I am not saying experimental inference is limited to claims like this — far from it — even probing theories, at any level, is done piecemeal.

My point now is just that we need to be very clear as to what is warranted...

“So far as this one set of data are concerned, no anomaly for GTR.”

(But obviously it is no evidence for it either!)

[Aside: Any account of evidence that is prepared to say that the ruined plates give even a tiny bit of evidence for GTR has a problem.]

In the experimentalist treatment I am recommending, the tasks of pinpointing blame are split off and addressed piecemeal

Real effect? Approximate effect size? Causes?

In the eclipse experiments, the first part was to show the effect was real and not an artifact, as well as estimating the deflection effect

A<sub>1</sub>: the chemical in the development not responsible,  
A<sub>2</sub> : the telescopes did not suffer change of focus in  
6 months (from eclipse plate to check plate),

The second part was to ascertain if the effect is attributable to the sun's gravitational field (or if some other factor consistent with a Newtonian account could explain the observed deflection).

N<sub>1</sub>: cooling effect of the moon's shadow controlled or subtracted out

N<sub>2</sub>: corona lens controlled or subtracted out

N<sub>3</sub>: no confounding by mechanical properties of the ether

**Each involved a series of local experimental tests.**

**This way of viewing the problem challenges the ways it is depicted by philosophers of science...**

Larry Laudan claims that the price of my “experimentalist fix” for Duhem’s problem is to have “balkanized” testing, and also to have changed the problem (Laudan, 1997).

The real problem, claims Laudan, is which of existing large-scale theories to accept or prefer.

But it was Laudan (along with Lakatosians and many others) that changed the problem.

Laudan, recall, developed his account (20 years prior) to deliberately be **“immune from criticism of a Duhemian type”** (1977, p. 42)

“A way out of the Duhemian conundrum may emerge if, far from *localizing* blame or credit in one place, we simply *spread it evenly among the members of the complex*” (1977, 43).

Comparing rivals T and T' (e.g., Newton, Einstein) becomes giving criteria for preferring a complex .

(T & H & A), (T & H & A'), (T & H' & A),  
 (T & H' & A'), (T' & H' & A), (T' & H' & A'),  
 (T' & H & A), (T' & H & A')

**Here's his argument:**

There are always several different combinations of theories and auxiliary hypotheses that equally well accommodate the anomaly.

But scientists do adjudicate between them.

So, in adjudicating between them, scientists must appeal to criteria (that these philosophers regard as) beyond the empirical appraisal of the **well-testedness** of the hypotheses, e.g., as problem-solving ability, explanatory power, and scope.

*Whether this kind of “comparativist” appraisal is a good way to determine which large scale theory or paradigm to accept (I argue it is not) is a separate issue.*

*I deny that resolving Duhemian ambiguities calls for accepting or appraising large-scale theories (confronted with anomalies).*

Thus, where Laudan wants to “steer me back” and Musgrave (2010) says “I want to entice her back” to the project of comparative appraisal of high-level theories (p. 105)

*I, in turn, wish to steer philosophers of science back to the initial Duhemian challenge.*

I have the chutzpah to suggest: Maybe it was given up on too easily, and maybe we know a bit more about experimental testing...

Clearly, we do not always have a warranted way to attribute blame, nor need we always have enough information to properly scrutinize attempts.

*But even then, we are led to strategies that direct progress with Duhemian problems and explain how scientists actually grapple with them.*

They do not spread the blame around evenly, or sum up each large scale theory’s # of “solved problems” (at least not in finding cause of anomalies)

***Doesn’t it matter what is actually responsible, e.g., for building a better theory, better experiments? for learning what is really the case?***

Here's another thing scientists don't do:

Try to write a great big conditional with a huge conjunction in the antecedent and predicted experimental data  $\mathbf{x}_0$  as the consequent...

If  $T \ \& \ N_1 \ \& \ N_2 \ \& \ \dots \ \& \ N_k \ \& \ A_1 \ \& \ A_2$   
 $\& \dots \ \& \ A_r$ , then  $\mathbf{x}_0$

The initial skepticism on Duhemian problems had a lot to do with this traditional hypothetical-deductive logic of testing...

It appears that a failed prediction leads to a disjunction: either  $T$  is to blame, or  $N_1 \dots \dots$  or  $A_1$  or is to blame, or ... etc.

*Actual experimental learning* shows the H-D model is not only divorced from scientific practice

***It shows it is bad scientific practice.***

Theories do not entail observed data, and even if we could write down an entailment, to enumerate and eliminate one by one would be a poor way to learn from data.

Take our GTR example: even in 1919 before the “zoo of rival” to GTR were developed, GTR and Newton’s theories of gravity could at most be seen to allow something like:

If GTR &

$N_1$  &  $N_2$  & & .....&  $N_k$  (a variety of effects not interfering)

$A_1$  & &  $A_2$  & ..... &  $A_r$  & (experiment well run)

*then* the mean deflection is

$$\lambda = 1.75''$$

(.87'' if Newton is right, and light has mass)

*Nor did pinpointing the mirror distortion (denying  $A_2$ ) require slogging through these others factors*

It did require astute use of the data plus aspects of the experiment that could never had been known in advance (including how many samples would be usable).

*Experimental accounts (I hope) — some of them—  
will elucidate the nitty gritty detail.*

They didn't observe a deflection near the sun:

- they record photos of positions of stars, then estimate where they are on the date taken (May 1919),
- then use photos 6 months after when there's no sun to infer, statistically, where the same stars would have been were the sun's effect absent (a control), from which we may estimate a value for what the mean deflection  $\lambda$  would be (were the stars actually right at the sun, which they are not.)

*But isn't this distance from the actual data what makes philosophers of science throw up their hands and declare Duhemian problems hopeless?*

## **Key to reliable experimental evidence\*:**

*The estimated deflection is much more accurate than any of the separate determinations!*

In contrast to the common supposition that the inferred claim is no more reliable than the pieces involved, in good experimental analyses we go from highly shaky data to far more accurate modeled data.

## **Experimental lift off!**

\*I propose to call a context “experimental” whenever the minimal requirement for evidence can be determined  
(even better if we can mount a reliable argument from coincidence or severe test)

We are clearly not limited to cases of literal experiments.

**Experimental Approach to Duhem's Problem.**

**We might allow some entailments:**

**(go down) From theory  $T$  to  $H$**

From theory  $T$  to experimental phenomenon

From contemplated experiment to an  
experimental prediction  $H$

**(go up) From data  $x_0$  to  $D$**

From actual raw data to inferred estimates of  
star positions on May 19

To estimates of deflection

To estimates of what the deflection would be  
near the sun  $D$

***Then, within an experimental context  $E$ , we can illuminate how  $D$  can be used to probe  $H$ .***

If we do this fully, we quickly run out of letters, and it doesn't matter how you break it down ... different cases call for different analyses ...

**Even testing rival theoretical explanations of the deflection effect, it is to local experimental arguments that scientists turn:**

In the other 2 experiments, Newtonian defenders did not deny the (deflection effect is real)

- The problem turned to testing purported “saves” of Newton,
- They were not discounted as ad hoc as with rigging (despite double-counting data)

Newton defenders alleged:

N' : The observed deflection is due to factor N, other than gravitational effects of the sun where N is a factor that at the same time saved the Newtonian law from refutation.

- Each was criticized by paying close attention to detailed experimental data (over 3 years) (not treated as “illegitimate rigging”):
- (i) the effect of the conjectured N-factor is too small to account for the eclipse effect; and
  - (ii) if the N-factor were large enough to account for the eclipse effect, it would have other false or contradictory implications.

Each hypothesized N-factor is shown to be false:

*Alternatively, to uphold N' as the way to accommodate the anomaly would be to commit a classic case of what is disallowed in saving a threatened theory.*

It would make it easy for a hypothesis of form N' to pass, even if it is false and auxiliary hypothesis N is true (i.e., high error probabilities).

- violates minimal condition for evidence

## Conclusion

*If new experimentalists are going to challenge philosophy of science, we have to revise some of the standard ways of characterizing problems of evidence and inference!*

- Neither comparative appraisal of large scale theories (spreading the blame equally) nor *H-D* logical deductions will do,
- the links between data and a primary claim under test are forged, not by conjunctions of background hypotheses, but by a variety of intermediate models, experimental and data models and statistical inferences,

We should *put our philosophy of experiment at the level of experiment*—and see how far that can take us!

*What would take center stage would be the variety of strategies and robustness arguments by which we infer something about the data-generating mechanism, quite apart from theories under test.*

- The failed prediction, “the anomaly” may be a highly modeled entity far away from the raw data and background factors in a full report of the nitty-gritty details of an experiment.
- These offer a crucial source of evidence for discriminating auxiliaries
- More than striving to check if auxiliaries and assumptions hold, we may identify tools to distinguish the effects of given factors.
- These tools often work by estimating backgrounds, and "subtract out" influences of factors other than some intended one.

A. Block a purported "save"

1. guilty of rigging (Velikovsky)
2. use detailed data  $\mathbf{x}_0$  to argue factor  $F$  could not account for (does not "fit") the results

B. Permit a "save" of  $H$  (blame the purported anomaly on something other than  $H$ )

(e.g., the mirror distortion in one set of plates in testing GTR)

- 1. Show experiment not well run (e.g., did not give a reliable estimate of the deflection effect, if any)
- 2. pinpoint the source of the alleged anomaly (e.g., mirror distortion looks different from deflection effect)

Two overall experimental points:

- (i) Even a hypothesis  $H$  that fits the data perfectly may be poorly tested: *Fit is not enough,*
- (ii) *But is so easy to get a good fit?(need an adequate account of “fit”)*

Without peering into the actual data  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  we forfeit an important way out of “there’s always an alternative that fits” criticism

By the same token, within the anomalous data one may find indicators or “signatures” of the likely cause. ...

and thereby pass severely, a hypothesis that “factor  $F$  is responsible for *this* anomaly”, and do so independently of the background theories that are thought to be problematic.

Having gotten this far, experimental philosophers can turn to even more general problems of “underdetermination” which I have no time to discuss today.