PERGAMON

Studies in History and Philosophy of
Modern Physics 34 (2003) 131–134

Studies in History
and Philosophy
of Modern Physics

www.elsevier.com/locate/shpsb

Discussion

# Novel work on problems of novelty? Comments on Hudson

Deborah G. Mayo

*Department of Philosophy, Virginia Tech, Blacksburg, VA 24006, USA*

Hudson's paper, "Novelty and the 1919 Eclipse Experiments," focuses on two main questions, both of which have been the subject of debate among philosophers of science: The first is (1) when, and why, should novel data be preferred or required in evaluating hypotheses and theories? The second is (2) were there good grounds for discounting one of the eclipse results in the 1919 testing of Einstein's General Theory of Relativity? Hudson sets out to provide answers to (1) and (2) that differ from, or improve upon, those that I discuss (in Mayo, 1991, 1996), but I do not see that he has made his case. He has not shown that his answer to (1) goes beyond a rewording of my condition; nor has he adduced historical information that speaks to an altered answer to question (2) (i.e., a "no" rather than a "yes"). Nevertheless, further discussion of these issues is needed, and the main goal of my comments will be to highlight two broad problem areas or queries to which new experimentalists might turn their attention—if they are seeking to make real and interesting progress beyond what has already been done. The first is (i) when do violations of use-novelty, and more generally, "double counting" of data prevent hypotheses from passing severe or reliable tests? The second is (ii) what are some general strategies for informal assessments of reliability or severity? Can we identify a handful of canonical models of error and informal arguments from error? Tackling these, it seems to me, will require careful study of historical episodes including an analysis—or reanalysis—of the relevant data and statistical techniques.

## 1. Severity as the rationale for the requirement of use-novelty

Hudson claims he will "reformulate use-novelty in a way that accommodates Mayo's insight" (in Mayo, 1996) that hypotheses can be use-constructed by means of highly reliable use-construction rules. Some examples I give there of reliable

*E-mail address:* mayod@vt.edu (D.G. Mayo).

1355-2198/02/$ - see front matter
PII: S 1 3 5 5 - 2 1 9 8 ( 0 2 ) 0 0 0 8 3 - 7

use-construction rules come from formal statistics, e.g., 95% confidence interval rules (272); others are entirely informal, e.g., using the "three feet of mangled, soot-encrusted steel" (251) from the 1993 bombing of the World Trade Center to both construct and support hypotheses about the person who rented the van carrying the explosive device. In fact, the example that first convinced me that there was something amiss in the use-novelty requirement was of the informal variety—constructing and testing a hypothesis about who dented my 1976 Camaro (276).

What is Hudson's reformulation? "The reformulation I call *prima facie* use-novelty…: where an experimental procedure generates evidence that is used in the construction of a hypothesis, and where there isn't any assurance that the experimental procedure reliably indicates the truth of the hypothesis constructed, we should not suppose that the use-constructed hypothesis is confirmed by this evidence." (7) But in my view this would be true for *any* hypothesis and is not special for the case of use-constructed hypotheses. So, it is not clear how this is any kind of a reformulation of my position. He has simply substituted my "so long as H passes a severe test with data x" with "so long as x reliably indicates the truth of H"; but one has reliable evidence for H to the extent that H passes a severe test—that is, to the extent that the ways H could be in error have been investigated and eliminated.

Oddly, Hudson has not explained why there is some special problem of reliability in the case of use-constructions in the first place. The key dispute, after all, has been over the epistemological rational behind the supposition that violating novelty threatens reliability. I have argued that violating use-novelty *can* cause special problems, and I have identified problematic cases where it is easy to use-construct a hypothesis even though that hypothesis is false. I don't see that Hudson has done any (much needed!) additional work in identifying *types* of problematic cases and distinguishing them from unproblematic ones. I will elaborate a little.

It is well known, for example, that if one is allowed to search through several factors and report just those that show (apparently) impressive correlations, there is a high probability of erroneously inferring a real correlation. Because the probability of finding some such correlation *or other* is high, even if all are due to chance, the assertion of a genuine correlation does *not* pass a severe test. However, it is equally clear that using the same data both to identify and test the cause, say, of an explosion, may allow us to infer the cause with very high severity. While severity gives us a platform for judging when to allow double-counting, it turns out to be much more difficult than might be expected to determine just when such data dependencies create obstacles to severity. Clear intuitions about extreme cases are not enough; we need general ways to evaluate the impact on severity or reliability for a host of different data-dependent and "use-construction" methods. If one is looking to cover novel ground, why not try to make some progress here?

## 2. The mirror distortion

The episode of testing Einstein's GTR by means of the eclipse results in 1919 is fascinating. Hudson questions whether researchers knew enough in 1919–21 to discount the Sobral astrographic results as supplying a reliable estimate of the deflection effect. This discounting is controversial because the Sobral data appeared to yield a result in favor of Newton rather than Einstein! In this Hudson agrees with Earman and Glymour (1980) in their important paper on this episode, and disagrees with me. Says Hudson, "Mayo is right that scientists at this time lacked sufficiently precise knowledge of the change in focus with the Sobral plates. Nevertheless, contrary to Mayo, Dyson et al. (1920) *did* use least squares in evaluating the data from the Sobral astrographic plates..and they arrived at a value in accordance with the Newtonian prediction." (Hudson 22). I do not see this to be contrary to me. It's obvious that they *did* run several least-squares estimates. After all, it's only by doing so that they arrived at the potentially pro-Newtonian value, at least as one of the estimates they consider. Perhaps Hudson feels they would not carry out these calculations if they thought some of the assumptions for the method were lacking, but that is not so. The researchers went on to criticize the resulting estimate, and in so doing they used the data that went into the least-squares estimate—hence violating use-novelty. My point was really just to illustrate, once again, the validity of such a double use of data: there is no other (or no better) way to show that the necessary assumptions of the least squares estimate are lacking. The statistical analysis of assumptions of this type of analysis was well-known at the time—and this sufficed to discount the Sobral result as posing an anomaly for Einstein (i.e., Positive information that the assumptions are met would have been required to warrant *not* discounting it!). So long as the assumptions needed to sustain the anomalous estimate are lacking—and Hudson seems to agree they were—the researchers lacked grounds to infer that these results are truly anomalous for Einstein and in accord with Newton. So, something more than Hudson's observations of this historical episode would be needed to question the researchers' stated data analysis.

## 3. Severity assessments and arguing from error: formal and informal

In the final section of his paper, Hudson does attempt to bring out some differences between us, despite the apparent similarities, claiming that for him, "whether a result is a good 'error probe', whether it severely tests a hypothesis, depends largely on whether the experimental process that generates it is a reliable one, where reliability has nothing whatsoever to do with statistics or Mayoian severity." (31) Really? While Hudson has not defined his notion of reliable testing, it seems clear from all he has said in the rest of this paper that our notions do not differ. One wishes he had given some grounds for declaring, all of a sudden, to be holding a concept of reliability having "nothing whatsoever to do with" reliability in "my" sense, which is the usual statistical one. I take it that Hudson just means to emphasize that determining reliability for him (in contrast to me??) turns on lots of empirical information and background knowledge of the manifold sources of error, theories of the instruments, etc. But, of course, I am fully aware of this. Evaluating evidence, I couldn't be more clear, is not a matter of logical or probabilistic relationships, but turns on *empirical* information about how the data were generated and about the overall experimental testing context. The point of my requiring a framework with a multi-level series of models was precisely to call attention to the complex interrelationships between the various theories and models that are needed even to arrive at reliable evidence; more still in order to link data to substantive primary questions. There can be no severity assessment, and hence no warranted inductive inference, without all of these considerations.

Perhaps Hudson just objects to my calling the program "error statistical," thinking it is too narrow. But I make it clear that I construe statistics broadly to include "the conglomeration of systematic tools," formal and informal, for generating, modeling, criticizing, and learning from data (451). If I often discuss methods and principles gleaned from error statistical methodology it is because they offer a treasure chest of ideas, still largely untapped by (yet highly relevant for) philosophers of experiment; and these may be used as formal analogues to the informal tasks and subject-specific details of experimental inquiry. I certainly never meant to limit the error statistical account to problems admitting of formal statistical models and methods. The strongest severity arguments are informal "arguments from error," e.g., inferring the source of the Camaro dent, Hacking's well-known "argument from coincidence" for dense bodies. At the opposite end from such day-to-day arguments from error are cases that admit of literal experimental control of factors; here, too, no formal statistical model is needed to capture experimental arguments. Statistical strategies arise to teach us about the some of the *hardest cases*: those which neither admit of obviously strong arguments from error nor permit experimental manipulation or control. In so doing they offer powerful insights for experimental reasoning in general: the same patterns of arguing from error are instantiated whether quantitatively or qualitatively determined. One might view standard statistical methods as capable of serving a role analogous to the one formal logic has long played in philosophy; but unlike the latter, it combines (just the right blend of) both logical and empirical considerations.

Note, for example, the value of statistical methodology in getting to the bottom of the issue at hand: the conflicting intuitions surrounding the novelty requirement. The common intuition that if the data are not novel then they fail to test or provide reliable evidence for a hypothesis, I came to see, stemmed from an ambiguity surrounding the common charge that "a use-constructed hypothesis will pass the test no matter what"—something that becomes easy to recognize using the "sampling distribution" concept in statistics.

Nevertheless, I heartily agree that there is enormous work that needs to be done in the arena of informal severity assessments and the identification of "canonical models of error." This will require the careful attention to the nitty-gritty details from historical episodes of the sort that I would think experimental philosophers like Hudson are well-positioned to carry out. Such contributions would provide genuine advancements and novel insights to the new experimentalist program.

## References

Earman, J., & Glymour, C. (1980). Relativity and eclipses: The British Eclipse Expeditions of 1919 and their predecessors. *Historical Studies in the Physical Sciences*, *11*, 49–85.

Mayo, D. (1991). Novel evidence and severe tests. *Philosophy of Science*, *58*, 523–552.

Mayo, D. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.