II Explanation and Testing
Exchanges with Clark Glymour

Deborah G. Mayo

Clark Glymour's contribution explores connections between explanation and testing and some applications of these themes to the discovery of causal explanations through graphical modeling. My remarks touch on a number of subthemes that have emerged from the back-and-forth exchanges Glymour and I have had since ERROR06,[1] especially insofar as they connect with problems and arguments of earlier chapters. As we get close to the end of the volume, I want also to provide the reader with some directions for interconnecting and building on earlier themes.

1. *Experimental Reasoning and Reliability*: How do logical accounts of explanation link with logics of confirmation and testing? When does *H*'s successfully explaining **x** warrant inferring the truth or correctness of *H*?
2. *Objectivity and Rationality*: Do explanatory virtues promote truth or do they conflict with well-testedness? How should probabilistic/statistical accounts enter into scrutinizing methodological desiderata (e.g., promote explanatory virtues) and rules (e.g., avoid irrelevant conjunction, varying evidence)?
3. *Metaphilosophical Themes*: How should probabilistic/statistical accounts enter into solving philosophical problems? What roles can or should philosophers play in methodological problems in practice? (Should we be in the business of improving practice as well as clarifying, reconstructing, or justifying practice?)

[1] Especially beneficial was the opportunity afforded by the joint course taught by Glymour and Spanos at Virginia Tech, Fall 2007.

1 Explanation and Testing

The philosophical literature on explanation is at least as large as the literature on testing, but the two have usually been tackled separately. I have always felt that the latter is more fundamental than the former and that, once an adequate account of inference is in hand, one may satisfactorily consider key explanatory questions. I find Glymour's thoughts on linking explanation and testing sufficiently insightful to enable me to take some first steps to connect the severity account of testing to some of the classic issues of explanation.

Glymour's insight is "that explanations and the explanatory virtues that facilitate comprehension also facilitate testing the claims made in explanation" (this volume, p. 331). Although he does not claim explanation and testability always go hand in hand, he considers examples where "the explanations specify tests whose results, if in accord with the explanations, converge on the truth *conditionally*" (p. 332). It will be interesting to see what happens if we replace his "test" with "severe test" as I define it. Recall our basic principle about evidence:

***Severity Principle (weak)*:** Data **x** (produced by *G*) do *not* provide good evidence for hypothesis *H* if **x** results from a test procedure with a very low probability or capacity of having uncovered the falsity of *H* (even if *H* is incorrect).

A test "uncovers" or signals the falsity of *H* by producing outcomes that are discordant with or that fail to "fit" what is expected were *H* correct. This weak notion of severity suffices for the current discussion.[2]

To Glymour's question, When does *H*'s successfully explaining **x** warrant inferring the truth or correctness of *H*?, our answer would be "only when *H* is severely tested by **x**." (*H* could still be viewed as an explanation of **x** without anyone claiming that *H* is warranted *because H* explains **x**. Here I only consider the position that *H* is warranted *by dint* of *H*'s explaining **x**.) I want to consider how much mileage may accrue from this answer when it comes to a key question raised by Glymour's discussion. For a model or theory *H*:

- How do/can explanatory virtues (possessed by *H*) promote *H*'s well-testedness (or *H*'s capacity for being subject to stringent tests)?

[2] Note that satisfying this weak principle does not mean the test is severe, merely that it is not blatantly insevere.

1.1 A Few Terminological Notes

Some informal uses of language in this exchange warrant a few notes. "*H* is severely tested" is a shorthand for "*H* passes a severe test," and to say that "*H* is (statistically) falsified" or "rejected" is the same as saying "not-*H* passes the test." Although this wording may seem nonstandard, the advantage is that it lets us work with the single criterion of severity for assessing any inference (even with test procedures built on controlling both erroneous acceptances and erroneous rejections). "*H* is false" may be seen to refer to the existence of a particular error that *H* is denying. The correctness of *H* may generally be seen to assert the absence of a specific error or discrepancy. To allude to examples in earlier exchanges, *H* may be viewed as the alternative hypothesis to a null hypothesis that asserts not-*H*. (For example, "*H*: $\mu > 2$ passes severely" may be identified with rejecting a null hypothesis "H_0: $\mu \leq 2$".)

2 Is There a Tension between Explanation and Testing?

2.1 A Goal for Science versus a Criterion for Inference

Pondering Glymour's question of the relationship between explanation and testing lets me elucidate and strengthen the responses I have given to my "high-level theory" critics in previous chapters. According to Chalmers, "[a] tension exists between the demand for severity and the desire for generality. The many possible applications of a general theory constitute many possible ways in which that theory could conceivably break down." (Chapter 2, p. 61). This leads him to recommend a weakened notion of severity. But there is a confusion here. The aim of science in the account I favor is not severity but *finding things out*, increasing understanding and learning. Severity is a criterion for determining what has and has not been learned, what inferences are warranted, and where our understanding may be in error. From the start I emphasized the *twin goals* of "severity and informativeness" (Mayo, 1996, p. 41). Trivial but severely passed claims do not advance learning. So we have, first, that the goal is finding things out, and second, that this demands learning from error and controlling error (which is not the same as error freedom). Does this mean that we shortchange "explanatory virtues"? My high-level theory critics say yes.

Musgrave (this volume) alludes to an earlier exchange in which Laudan (1997) urges me to factor "explanatory power back into theory evaluation."

There Laudan claims that "in the appraisal of theories and hypotheses, what does (and what should) principally matter to scientists is not so much whether those hypotheses are true or probable. What matters, rather, is the ability of theories to solve empirical problems – a feature that others might call a theory's explanatory or predictive power" (p. 306). Like Glymour, I deny that what matters is explanatory power *as opposed to* truth – truth does matter. Like Laudan I deny that one is after "highly probable" hypotheses in the sense of epistemic probabilists (e.g., Achinstein's Bayesian, or Musgrave's "justificationist"). What we want are hypotheses that have successfully been highly probed. What I wish to explore – taking advantage of Glymour's aperçu – is whether increasing predictive power is tied to increasing the probativeness of tests.

Laudan's "problem-solving" goal is covered by the twin informativeness/severity demands: we want informative solutions to interesting problems, but we also want to avoid procedures that would, with high probability, erroneously declare a problem solved by a hypothesized solution *H*. Far from showing "bland indifference to the issue of a theory's ability to account for many of the phenomena falling in its domain" (Laudan, 1997, pp. 306–7), when a theory fails to account for phenomena in its domain, that is a strong indication of errors (and potential rivals) not yet ruled out – concerns that are close to the heart of the severe tester. Moreover, as we saw in the case of GTR, such gaps are springboards for creating new hypothesized solutions of increased scope and depth, as well as developing tests that can reliably distinguish them.

2.2 Powerful Explanations and Powerful Tests

Glymour asks: "We know that under various assumptions there is a connection between *testing* – doing something that could reveal the falsity of various claims – and . . . coming to understand. Various forms of testing can be stages in strategies that reliably converge on the truth. . . . But how can *explaining* be any kind of reliable guide to any truth worth knowing?" (p. 331)

One answer seems to drop out directly from the severe testers' goals: to reliably infer "truth worth knowing," we seek improved tests – tests that enable us to corroborate severely claims beyond those already warranted. This leads to identifying interconnected probes for constraining claims and controlling error probabilities, which simultanously results in frameworks and theories that unify. Starting from severe testing goals, we are led to stress an initial hypothesis by subjecting it to probes in ever-wider

domains because that is how we increase the probability of revealing flaws. We thereby are led to a hypothesis or theory with characteristics that render it "explanatory."

3 When the "Symmetry" Thesis Holds

According to Carl Hempel, the link between explanation and testing is provided by an equivalence known to philosophers as the symmetry thesis: premises explain a phenomenon if and only if the phenomenon could be predicted from the premises. (Glymour, this volume p. 332)

Although the symmetry thesis is strictly false, Glymour, in our exchanges, tries to trace out when a symmetry does seem to hold. This led to my own way of linking the two (which may differ from what Glymour intended or holds). To begin with, the entanglement between successful explanations and probative tests is likely to remain hidden if we insist on starting philosophical analysis with a given theory and ignore the processes by which explanatory theories grow (or are discovered) from knowledge gaps.

3.1 The Growth of Explanatory Theories: The Case of Kuru

To move away from GTR, I consider some aspects of learning about the disorder known as Kuru found mainly among the Fore people of New Guinea in the 1960s. Kuru, and (what we now know to be) related diseases (e.g., Mad Cow, Crutzfield-Jacobs disease [CJD]) are known as "spongiform" diseases because the brains of their victims develop small holes, giving them a spongy appearance. (See Prusiner, 2003.)

For H to successfully explain **x** – as I am using that term – there must be good evidence for H.[3] For example, we would deny that witchcraft – one of the proposed explanations for Kuru – successfully explains Kuru among the Fore. Still, **x** may be excellent evidence for H (H may pass a severe test with **x**) even though one would not normally say that H explains **x**: hence the asymmetry. As with the familiar asymmetric examples (e.g., barometer-weather), the evidence that a patient is afflicted with Kuru warrants inferring H:

H: the patient has a hole-riddled cortex,

while the cortex holes do not explain her having Kuru.

[3] The data that is explained, **x**, need not constitute this evidence for H.

Perhaps this example may be scotched by our (usual) stipulation that the hypothesis *H* refers to an aspect of the procedure generating **x**. But other examples may be found where we would concur that *H* is well tested by **x**, and yet *H* does not seem to explain **x**. For example, from considerable data in the 1950s, we had evidence **x**: instances of Kuru cluster within Fore families, in particular among women and their children, or elderly parents. Evidence **x** warrants, with severity,

H: there is an association of Kuru within families

but to say that *H* explains **x** sounds too much like explaining by "dormative properties."

These are the kinds of cases behind the well-known asymmetry between explanation and tests. Focus now on the interesting activity that is triggered by the recognition that *H* does not explain **x**: Prusiner, one of the early Kuru researchers, was well aware of this explanatory gap even without a clue as to what a satisfactory theory of Kuru might be. He asked: What causes Kuru? Is it transmitted through genetics? Infection? Or something else? Are its causes similar to those of other amyloid diseases (e.g., Alzheimer's)? How can it be controlled or irradicated? This leads to conjectures to fill these knowledge gaps and how such conjectures can be in error – which, in turn, provides an incentive to create and test more comprehensive theories.

That Kuru was a genetic disorder seemed to fit the pattern of observed cases, but here we see why the philosopher's vague notions of "fit" will not do. By 1979 it was recognized that a genetic explanation actually did not fit the pattern of data at all – Kuru was too common and too fatal among the Fore to be explained as a genetic disorder (it would have died out of the gene pool). Rather it was determined with severity that the correct explanation of the transmission in the Fore peoples was through mortuary cannibalism by the maternal kin (this was a main source of meat permitted women):

H: Kuru is transmitted through eating infected brains in funeral rites.

Ending these cannibalistic practices all but eradicated the disease, which had been of epidemic proportions. The fifty or so years it has taken to arrive at current theories – of which I am giving only the most sketchy glimpse – is a typical illustration of the back-and-forth movement between

A. the goal of attaining a more comprehensive understanding of phenomena (e.g., the dynamics of Kuru and related diseases, Mad Cow, CJD) and
B. the exploitation of multiple linkages to cross-check, and subtract out, errors.

Despite the inaccuracies of each link on its own, they may be put together to avoid errors. A unique aspect of this episode was the identification of a new entity – a prion – the first infectious agent containing no nucleic acid. (Many even considered this a Kuhnian revolution, involving, as it does, a changed metaphysics.) This revolutionary shift was driven by local experimental probes, most notably the fact that prions (e.g., from scrapie-infected brains) remain infectious even when subjected to radiation and other treatments known (through prior severe tests) to eradicate nucleic acids (whereas they are inactivated by treatments that destroy proteins). Only in the past ten years do we have theories of infectious prion proteins, and know something of how they replicate despite having no nucleic acid, by converting normal proteins into pathological prions.

Prion theories earned their badges for (being warranted with) high severity by affording ever stronger arguments from coincidence through interrelated checks (e.g., transmitting Kuru to chimpanzees). Moreover, these deeper prion theories were more severely corroborated than were the early, and more local, hypotheses such as *H*: Kuru is transmitted among the Fore through eating infected brains. (It does not even seem correct to consider these local hypotheses as "parts of" or entailed by the more comprehensive theories, but nothing turns on this.) However, and this is my main point, the same features that rendered these theories better tested, simultaneously earned them merit badges for deepening our understanding – for explaining the similarities and differences between Kuru, CJD, and Mad Cow – and for setting the stage to learn more about those aspects of prions that continue to puzzle microbiologists and epidemiologists. (Prusiner received the Nobel prize in 1997; see Prusiner, 2003.)

The position that emerges, then, is not that explanatory virtues are responsible for well-testedness, nor even that they are reliable signs of well-testedness *in and of themselves*; it is rather that these characteristics will (or tend to) be possessed by theories that result from fruitful scientific inquiries. By fruitful scientific inquiries I mean those inquiries that are driven by the goal of *finding things out* by reliable probes of errors.

4 Explanation, Unification, and Testing

Hempel's logical account of explanation inherits the problems with the corresponding hypothetico-deductive (HD) account of testing: *H* may entail, predict, or otherwise accord with **x**, but so good a fit may be highly probable (or even guaranteed) when *H* is false. The problem with such an HD account of confirmation is not only that it is based on the invalid affirming the consequent – that, after all, is only problematic when the hypothesis fails to be

falsified. Even in the case of a deductively valid falsification, warranting the truth of the first premise "if *H* then **x**" is generally problematic. The usual assumption that there is some conjunction of "background conditions" *B* so that *H* together with *B* entails **x** has scant relationship to the ways hypotheses are linked to actual data in practice. Furthermore, the truth of the conditional "if *H* (and *B*) then **x**" does not vouchsafe what even HD theorists generally regard as a minimal requirement for a good test – that something has been done that could have found *H* false. This requires not mere logic but evidence that anomalies or "falsifying hypotheses" (as Popper called them) are identifiable and not too easily evadable. This, at any rate, would be required for a failure to falsify to count as any kind of evidence for *H*. This leads to my suggested reading of Glymour's point in his discussion of Copernicus. Ensuring that anomalies for *H* are recognizable goes hand in hand with *H* possessing explanatory attributes, such as the ability to transform a contingent empirical regularity into a necessary consequence. Says Glymour, "If the empirical regularity is false, then the entire Copernican framework is wrong; nothing is salvageable except by *ad hoc* moves (as in, *oh well, Mars isn't really a 'planet'*)." By contrast, "The same regularity is accounted for in Ptolemaic theory by adjusting parameters" (this volume, p. 336). I take the point to be that the former and not the latter warrants taking the agreement between *H* and the observed empirical regularity as *H* having passed a genuine test (even if weak). In the former case, the hypothesis or theory sticks its neck out, as it were, and says such-and-such would be observed, *by necessity*, so we get a stronger test.

The history of attempted improvements on HD accounts has been to add some additional requirement to the condition that *H* "accord with" evidence **x** in order to avoid too-easy confirmations. We have seen this with requiring or preferring theories that make novel predictions (e.g., Chapter 4). The same role, I suggest, is behind advocating certain explanatory virtues. However, because such attributes are neither necessary nor sufficient for meeting the "weak severity" requirement for a genuine test, it is more effective to make the severity requirement explicit. To be clear, I am not saying explanatory virtues are desirable only to vouchsafe genuine tests – they are desirable to achieve understanding and other goals, both epistemic and pragmatic. Here, my aim has been limited to exploring *how explanatory virtues may simultaneously promote grounds for inferring or believing the explanation*. It would be of interest to go further in the direction in which Glymour is valuably pointing us: toward analyzing the connection between explanatory power on the one hand and powerful tests on the other.

5 Irrelevant Conjunction

The issue of irrelevant conjunction is one on which Glymour and I have had numerous exchanges. On Hempel's logical account of explanation, Glymour notes, anything "lawlike," true or false, can be tacked on to a Hempelian explanation (or statistical relevance explanation) and generate another explanation (this volume, p. 332). That is,

If *H* explains **x**, then (*H* and *J*) explain **x**, where *J* is any "irrelevant" conjunct tacked on (the Pope's infallibility).

How would such a method fare on the severity account? From our necessary condition, we have that (*H* and *J*)'s explaining **x** cannot warrant taking **x** as evidence for the truth of (*H* and *J*) if **x** counts as a highly insevere test of (*H* and *J*). (See also Chapter 3, pp. 110, 123.)

A scrutiny of well-testedness may proceed by denying either condition for severity: (1) the fit condition, or the claim that (2) it is highly improbable to obtain so good a fit even if *H* is false. Here we are only requiring *weak* severity – as long as there is some reasonable chance (e.g., .5 or more) that the test would yield a worse fit, when *H* is false, then weak severity holds. Presumably, we are to grant that (*H* and *J*) fit **x** because *H* alone entails **x** (never mind how unrealistic such entailments usually are). Nevertheless condition 2 is violated. Say we start with data **x**, and that *H* explains **x**, and then irrelevant hypothesis *J* is tacked on. The fact that (*H* and *J*) fits **x** does not constitute having done anything to detect the falsity of *J*. Whether **x** or not-**x** occurs, the falsity of *J* would not be detected, and the conjunction would pass the test. Because this permits inferring hypotheses that have not been well tested in the least, the HD account is highly unreliable. (In a statistical setting, if the distribution of random variable *X* does not depend on hypothesis *J*, then observing *X* is uninformative about *J*, and in this informal context we have something similar.)

To go further, we should ask: when would we come across an assertion that a conjunction of hypotheses (*H* and *J*) explains **x**. Most commonly, saying *H* and *J* explain(s) **x** would be understood to mean either:

1. together they explain **x**, although neither does by itself, or
2. each explains **x** by itself.

A common example of case 1 would be tacking onto *H* an explanation of an *H*-anomaly (e.g., the Einstein deflection of light together with a mirror distortion explains the eclipse results at Sobral, (see Chapter 4)); an example of case 2 arises when *H* and *J* are rival ways of explaining **x**.

But the problem case describes a situation where H explains $\mathbf{x}$ and hypothesis J is "irrelevant." Although this is not defined, pretty clearly we can dream up the kind of example that the critics worry about. Consider

H: GTR and *J*: Kuru is transmitted through funerary cannibalism.

Let data $\mathbf{x}$ be a value of the observed deflection in accordance with GTR. The two hypotheses do not make reference to the same data models or experimental outcomes, so it is not clear that one can even satisfy the "fit" condition for a severe test.

That hypothesis $K = (H \text{ and } J)$ fits $\mathbf{x}$ requires, minimally, that $P(\mathbf{x}; K) > P(\mathbf{x}; \text{not-}K)$, and this would not seem to be satisfied (at least for an error statistician). Perhaps sufficient philosophical rigging can define something like a "Kuru or gravity experiment." However, the main force for rejecting the well-testedness of the conjunction of H and J is clearly that J has not been probed in the least by the deflection experiment.

We should emphasize a point regarding the weak severity requirement: it is not merely that the test needs to have some reasonable probability of detecting the falsity of H, if it is false. The indication of falsity (or discrepancy) has to be *because* of H's falsity. For example, we would not consider that a GTR hypothesis H had been well probed by a "test" that rejected H whenever a coin landed heads (of course the fit condition would also fail). In a good test, moreover, the *more false H* is, the higher should be the probability that the test detects it (by producing a worse fit, or a failing result). With the irrelevant conjunct, however, the falsity of J does not increase the detection ability; the test is *not registering J*'s falsity at all.

Someone may ask, but what if one is given a bundle like conjunction K at the start, rather than creating K by tacking J onto H? Which part is well tested? With this question we are back to where we began in our discussions of theory testing (Chapter 1) and in responding to Chalmers and Musgrave. What is well tested is what has passed severely, and a good part of scientific inquiry involves figuring this out. For example, we saw that it was determined that warranting the equivalence principle did not count as severely testing all of GTR, but only metric versus nonmetric theories. The evidence for the equivalence principle did not have the ability to discriminate between the class of metric theories.

6 Metaphilosophical Notes: The Philosophical Role of Probabilistic Inference

"The idea of putting probabilities over hypotheses delivered to philosophers a godsend, an entire package of superficiality" (Glymour, this volume,

p. 334). I share Glymour's indictment of what often goes under the heading of "Bayesian epistemology" (to be distinguished from Bayesian statistics) – at least if the aim is solving rather than merely reconstructing. In Mayo (1996) I described the shortcomings of claims to "solve" problems about evidence, such as Duhem's problem, by means of a probabilistic reconstruction: "Solving Duhem comes down to a homework assignment of how various assumptions and priors allow the scientific inference reached to be in accord with that reached via Bayes's theorem" (p. 457). They do not tell us either how the assignments are arrived at or, more important, how to determine where the error really lies. The same problem arises, Worrall notes, in treating the issue of "use-novelty" among Bayesian philosophers. "The fact that every conceivable position in the prediction vs. accommodation debate has been defended on the basis of some Bayesian position is a perfect illustration of the fact that 'the' Bayesian position can explain everything and so really explains nothing" (Worrall, 2006, pp. 205–6).

Most ironic about this practice is that rather than use statistical ideas to answer questions about methodology, the Bayesian epistemologist starts out assuming the intuition or principle to be justified, the task then being the "homework problem" of finding assignments and/or selecting from one of the various ratio or difference probability measures to capture the assumed intuition. Take the "tacking problem" discussed in Section 5:

> The Bayesian epistemology literature is filled with shadows and illusions: for example, Bayesian philosophers solve the difficulty that logical relations constituting an explanation can be conjoined with irrelevancies – the tacking on problems – by just *saying*, ad hoc for each case, that tacking on reduces degree of belief. (Glymour, this volume, p. 335)

At best, they are able to say that the conjunction gets less support than the conjunct, when what we want to say, it seems to me, is that there is no evidence for the irrelevant conjunct, and the supposed "test by which the irrelevant conjunct is inferred" is a terrible (zero-severity) test. Any account that cannot express this forfeits its ability to be relevant to criticizing even egregious violations of evidence requirements.

The same problem, Glymour observes, occurs in subjective Bayesian attempts to show why explanatory unification supplies greater degrees of belief. Here, too, we are to *start out* assuming some methodological principle (about unification and belief). The task is to carry out the Bayesian computation of hammering out probabilities to accord with the assumed principle. In fact, however, H's unifying power and the warrant for believing or inferring H need not go hand in hand. Surely a theory that incorrectly "unifies" phenomena ought not to earn higher belief: the similarities between Kuru

and other amyloid diseases such as Alzheimer's, for instance, should not give extra credence to a hypothesis that unified them rather than one that posited distinct mechanisms for Kuru and for Alzheimer's. The less-unified theory, in this case, is better tested (i.e., passed more severely).

I call on the philosopher of science with a penchant for probabilistic analysis to join in moving away from analytic reconstructions to link up to statistical inference (both for the problems and promise it affords).

7 Do Tests of Assumptions Involve Us in a Regress of Testing?

> Because total evidence is always finite, the entire testing procedure must, so the informal argument goes, be unfounded and depend on assumptions for which there is no test. (Glymour, this volume, p. 337)

The issue of assumptions is fundamental and has already poked its head into several of the contributions in the form of experimental and model assumptions, and assumptions needed to link experimental and statistical hypotheses to substantive hypotheses and theories. For the error statistician, what matters – indeed, what makes an inferential situation "experimental" – is the ability to sustain reliability or severity assessments. Philosophers often point up assumptions that would be *sufficient* to warrant an inference without adequate consideration of whether they are *necessary* to warrant the inference. If a large-scale theory or paradigm is assumed, then "use-constructed" hypotheses are warranted; but it is a mistake, or so I have argued, to suppose the hypothesis cannot be warranted by other means. Experimental relativists could have arrived at inferences about the deflection effect by assuming GTR and estimating parameters within it, or they could have instead warranted deflection inferences without assuming any one metric theory – as they did! Or, to go back to my homely example in Chapter 1, I could have arrived at my inference about George's weight gain by assuming the first scale used was reliable, or I could have done what I did do: use several different weighing machines, calibrate them by reference to known standard weights, and reach a reliable inference about George's weight that did not require assuming the reliability of any particular scale used. In these remarks, in fact, I have been arguing that the impetus to evaluate severity without depending on unknown assumptions simultaneously leads to hypotheses and theories with good explanatory characteristics.

Glymour's work on causal modeling exemplifies the "can-do" attitude of the error statistician: "The appropriate methodological enterprise is not to shout 'So there!' citing Duhem and Quine and such, but to investigate what can be tested and reliably discovered under varying assumptions"

(p. 338). If any single theme should be attached to what I have referred to as the "new experimentalism," it is the idea that the secret to avoiding the skeptical upshots of classic problems of underdetermination is to make shrewd use of experimental strategies. Experimental strategies for me need not involve literal control or manipulation but are strategies for controlling and evaluating severity, at least qualitatively. How far this may be achieved in the observational contexts in which Glymour works is an open question, but I see no reason why a combination of literal experiments (including randomized trials), simulations, and nonobservational inquiries could not be used to address the kind of assumptions he worries about in causal modeling. But looking deliberately at "nonexperimental" or observational data may actually be the best way to understand why certain kinds of experimental controls enable reliable probes of causal connections. Glymour suggests that the conditions that vouchsafe no unobserved common causes "are exactly [those] that experimental randomization is intended to provide" (p. 341). This may offer an intriguing path to explaining how (and when) randomization works, and it also suggests ways to mimic the results when literal randomization is not possible. The advances achieved by Zhang and many others – their utilization in a vast array of social sciences, computer science, and technology – speak to the valuable aperçu of Glymour, as does the work on testing statistical assumptions by Aris Spanos. This takes us to the next exchange by Spanos.

References

Laudan, L. (1997), "How about Bust? Factoring Explanatory Power Back into Theory Evaluation," *Philosophy of Science*, 64: 306–16.

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

Mayo, D.G. (1997a), "Duhem's Problem, the Bayesian Way, and Error Statistics, or 'What's Belief Got to Do with It?'" *Philosophy of Science*, 64: 222–44.

Prusiner, S.B. (2003), *Prion Biology and Diseases*, 2nd ed., Cold Spring Harbor, Laboratory Press, Woodbury, NY.

Related Exchanges

Mayo, D.G. and Miller J., (2008), "The Error Statistical Philosopher As Normative Naturalist," *Synthese (Error and Methodology in Practice: Selected Papers from ERROR 2006)*, 163(3): 305–14.

Parker, W.S. (2008), "Computer Simulation Through An Error-Statistical Lens," *Synthese (Error and Methodology in Practice: Selected Papers from ERROR 2006)*, 163(3): 371–84.