History of Economic Ideas, xiv/2006/1

REVIEW ARTICLES

★

THE LOGIC OF EXPERIMENTAL DISCOVERY

Ivan Moscati
Bocconi University
Department of Economics

A review essay on Francesco Guala, *The Methodology of Experimental Economics*, New York, Cambridge University Press, 2005, pp. xiv+286.

Guala's book is a study in the philosophy of science investigating the role of experiments in scientific discovery. Experimental economics offers the author a series of illustrations and case studies, and allows him to anchor his epistemological discussion to a specific and concrete example of experimental science. Yet, as Guala stresses, the work is not a handbook of experimental economics, nor a textbook of experimental methodology that teaches how to design economic experiments.

Guala's fundamental question revolves around the problem of what counts as a rational scientific inference. He sketches an original theory of inductive inference that attempts to account for the role experiments have in actual scientific practice, and in economics in particular.

The book effectively mixes a descriptive and a normative approach. On the one hand, Guala refers constantly to scientists' actual practices, and his theory of inductive inferences tries to capture the concerns that drive experimenters' research. On the other hand, Guala discusses from a normative viewpoint «whether these practices can be justified epistemically or, in other words, whether they can be expected to generate proper (rationally justified) scientific knowledge» (p. 38).

The work is divided into two parts. Part One (chs 2-6) examines the relationship between scientific hypotheses and experiments, and the way experiments allow researchers to draw reliable inferences about the validity of scientific hypotheses. Part Two (chs 7-11) discusses the relationship between experiments and the real world, and the so-called external validity problem. This concerns the extent to which the results obtained in the experimental setting can be generalized to real-world situations.

The focus of the present review essay is on Guala's epistemological tenets rather than on the examples he draws from experimental economics. Section 1 illustrates the epistemological tenets of Part One of the book, while section 2 is devoted to those contained in Part Two. Section 3 points out two issues left open by Guala's study.

1. Theory and Experiments

In chapter 2 Guala introduces the notion of experimental *replications,* as distinct from that of experimental *repetitions*. Repetitions of an experiment try to preserve the original experimental design in order to test whether the data provided by another experimenter are reliable. Replications involve instead some modification of the original design in order to investigate whether a phenomenon observed in the original experiment is robust to changes in the experimental setup. For instance, an economic experiment can be replicated by using larger or smaller amounts of monetary payoffs, or by giving the subjects more or less information about the experimental situation. If the phenomenon observed in the original experiment is recorded in different setups, it is more difficult to rationalize it as an outcome of a specific experimental environment. The phenomenon appears robust and calls for a scientific explanation.

Whereas repetitions are rarely performed, replications are a central part of experimental work. As Guala illustrates in the following chapters, replications are crucial because they are the basic tool for inductively inferring the causes of phenomena.

In chapter 3 the author lucidly presents the Duhem-Quine problem, its challenge to the basic Hypothetico-Deductive (HD) model of testing, and the role experiments have in a 'Duhem-Quine world'. According to the basic version of the HD model, scientists formulate a hypothesis *H* that logically implies a certain empirical statement *e*. Then two cases can occur. In the Refutation case, *e* is not observed, so that we can conclude by logic alone (and more exactly, by *modus tollens*) that *H* is false. In the Confirmation case, *e* is observed. Yet, from *e* we cannot deduce that *H* is true, but only that *e* confirms, supports, or corroborates *H*, meaning that the truth of *H* appears in some sense more likely in the light of *e*. According to the basic HD model there is then an asymmetry between Refutation and Confirmation: in the former deductive logic alone is sufficient to test a scientific theory, whereas in the latter we also need a logic of inductive inference to assess to what extent *e* confirms *H*.

Upon issues of inductive inference there is no agreement among philosophers. Some hypothetico-deductivists, notably Karl Popper among them, even deny any role for inductive inference in science. According to Popper, *modus tollens* and Refutation alone constitute the essence of scientific reasoning. However, Popper's view has been challenged by the Duhem-Quine thesis.

In different contexts, the French physicist P. Duhem and the American philosopher W. O. Quine pointed out that any scientific hypothesis *H* is never tested in isolation, but always together with a number of auxiliary hypotheses. The auxiliary hypotheses concern, among other things, the initial conditions of the test, the right specifications of free parameters, the correct functioning of the instruments, the absence of disturbing events and the fulfillment of all other *ceteris paribus* clauses. Let us label this set of auxiliary assumptions as $K=K_1$ and K_2... and K_n. When the HD model of testing accounts for the presence of auxiliary hypotheses, it takes the sophisticated form: *H* and *K* imply *e*.

The Duhem-Quine problem consists in the fact that the presence of auxiliary assumptions makes deductive logic alone no more sufficient to refute a scientific theory. In fact, what happens if *not-e* is observed? *Modus tollens* tells us that at least one assumption among *H*, K_1, K_2..., *or* K_n is false, but is silent about which one/ones is/are responsible for *not-e*. In other words, as Guala notices, deductive logic alone is not sufficient to refute *H*. Consequently, an inductive reasoning is needed to decide whether to reject *H* rather than some auxiliary assumption. This also means that the difference between Refutation and Confirmation in hypothesis testing is not so acute as the basic HD model suggests: both hang in the end on some form of inductive reasoning, as delicate and fallible as it may be.

Guala calls attention on the fact that the Duhem-Quine problem also explains the importance of experiments in hypothesis testing. With respect to other forms of empirical tests, laboratory experiments allow the scientist to control better the auxiliary assumptions, so that the Duhem-Quine problem becomes more manageable in the lab.

In chapter 4 Guala discusses the notion of causation, and the way individuals attempt to infer underlying causal relations from observable data. Guala's basic claim is that ordinary people as well as scientists try inductively to discover the causes of phenomena by controlled variation and comparison. Let us imagine that our hypothesis *H* is that X causes Y. Given a certain state of affairs, we try to modify only one element (the putative cause X), keeping all other background conditions *K* fixed or, at least, under control. Then we compare the initial and the final states of affairs, and tend to interpret the variation Y occurred in the state of affairs as caused by X. As Guala points out, for inferring causal relations experimental techniques are particularly useful since they make it possible to modify 'surgically' only one element at a time, and to control accurately the background conditions.

In chapter 5 Guala critically examines the idea that predictive success is the essential condition for supporting a scientific hypothesis *H*, as Popper, Imre Lakatos, and Milton Friedman, among others, maintained. The discussion of the predictive success criterion belongs to the more general issue concerning what counts as a justified scientific inference when the Duhem-Quine problem is taken into account. As observed above, in a Duhem-Quine world the refutation or the confirmation of a scientific hypothesis cannot be decided by deductive logic alone. Rather, some form of inductive reasoning is needed to evaluate how much support/disconfirmation a certain empirical evidence *e* gives to an hypothesis *H*. As Guala clearly puts it:

> Inductive support is a three-place relation among *H*, *e*, and *K* [the auxiliary assumptions] rather than a two-place relation between *H* and *e* [...]. It would be nice if one could provide a set of rules such that by inspecting the purely formal features of *H* and *e*, it were possible to establish whether *e* supports *H*, and to what extent. Once the *K* enter the picture, however, the issue of inductive support becomes contextualized: one cannot answer it by merely looking at the features of *e* and *H*. An empirical investigation is

necessary in order to establish whether the context is 'right' for *e* to be truly confirming evidence for *H*.

(p. 109)

According to the criterion of predictive success, the fact that a hypothesis *H* implies (and in this sense, explains) some already known evidence *e* is not sufficient to give *H* inductive support. In addition, *H* must also predict some evidence *e* that will be observed in a subsequent moment. Guala notices that the rationale behind this criterion seems to be the following. Since it is 'easy' to adjust *post hoc* an hypothesis in order to make it consistent with the already known evidence, even a false hypothesis could easily fit the existing evidence. On the contrary, it seems quite unlikely – 'a miracle' – that a false hypothesis generates successful predictions. But then, Guala observes, «predictive success is just an epiphenomenon of [a] more fundamental principle of inductive support» (p. 107), according to which *H* is supported by *e* if the probability of observing *e* when *H* is true is high, whereas the probability of observing *e* when *H* is false is low.

If this is the case, the temporal relation between *H* and *e* occupying the central stage in the predictive success criterion (the fact that *e* must follow *H*) is not actually relevant. Moreover, the predictive criterion focuses on the relation between *H* and *e* only, and fails to mention the background conditions *K* that are instead decisive for assessing the link between *e* and *H*. According to Guala, a theory of inductive support that fails to account for the importance of background conditions also fails to account for the importance of experiments in science:

Unless we model the *K* in the relation of inductive support, our theory of induction will be unable to explain the advantages of controlled experimentations over nonexperimental methods of investigation. In experimental science, the *K* can be systematically checked and thus possible source of errors eliminated.

(p. 110)

One theory of inductive support that takes into account the three-place relation among the hypothesis *H*, the empirical evidence *e*, and the auxiliary assumptions *K* is the Bayesian view of inference, which Guala analyzes and criticizes in Chapter 6. According to the Bayesian view, the degree of support provided by *e* to *H* and *K* is measured by a probability value, and, more exactly, by the posterior probability $P(H \,\&\, K|e)$ calculated via Bayes' rule:

$$P(H \,\&\, K|e) = \frac{P(e|H \,\&\, K)P(H \,\&\, K)}{P(e)}$$

where $P(H \,\&\, K)$ is the scientist's prior belief that *H* and a certain set of auxiliary assumptions *K* are true. Even if the Bayesian approach accounts for the importance of auxiliary assumptions, according to Guala it leaves too much room for subjectivity: «Bayesians [...] leave too much freedom for scientists to choose their favorite prior beliefs [...]. As a consequence, it is always possible for a dogmatic scientist to stick 'rationally' to a certain belief in spite of the evidence» (p. 136).

Guala supports a more objective approach to inductive inference, whose fundamental principle is the one introduced in Chapter 5: *H* is supported by *e* if the probability of observing *e* when *H* is true is high, whereas the probability of observing *e* when *H* is false is low. Since these probabilities are influenced by the auxiliary conditions *K*, evidence *e* is said to *support objectively H* if the *K* allow a strong inference from *e* to *H*, that is, if the *K* are so well under control that it is highly unlikely that the presence or the absence of *e* is due to the *K*.

Guala does not attempt to define the principle of objective inductive support in a more formal way. Rather, he contends that various circumstances show that the pursuit of objective support actually drives experimenters' research. First, most experimental work can be seen as an endeavor to guarantee objective support to a certain hypothesis *H* as an explanation for a given evidence *e* by checking for and eliminating possible disturbing factors *K* that could have originated *e*. Guala calls such procedure of check and elimination *eliminative induction*. Second, the principle of objective support seems to account for what is considered a bad/good experimental paper: «A bad paper is usually rejected [...] if the experiment it describes (with *that* particular design, with *those* background conditions) is such that it does not allow a strong inference from the collected evidence to the hypothesis at stake» (p. 129). Third, the search for objective support by eliminative induction rationalizes why replications are so important in experimental science: by replicating an experiment under different conditions, scientists check whether a certain phenomenon *e* is due to *H* or to some auxiliary factor *K*. Finally, the search for objective support explains why experimental knowledge grows progressively and slowly: eliminative induction takes time.

2. Experiments and the World

Chapter 7 opens the second part of the book and introduces the external validity problem. As stated above, the external validity of an experiment is the extent to which its results and insights can be extended to real-world situations of interest. In biochemistry it takes the form of 'the *in vitro-in vivo* problem', in psychology it is sometimes called 'ecological validity problem'. In economics it is often called 'parallelism problem' and essentially concerns the question whether real subjects in actual economic situations behave as experimental subjects do in the lab. Guala's thesis is that there is no universal answer to the external validity problem, but only local and empirical answers to be searched by inductive reasoning.

The basic idea behind this thesis is that the same causes generate the same effects, but only under the same conditions. Therefore, in order to assess whether a certain experimental result can be applied to a specific real-world situation, we have to check empirically and case-by-case whether the experimental conditions and the real-world conditions are comparable. The challenge, Gaula acknowledges, «is [then] to define more precisely the sort of methods that can be used in order to draw reliable external validity inferences» (p. 159).

Chapter 8 describes a successful case of external validity inference, in which the US Federal Communications Commission (FCC) sold a group of telecommunications licenses at auction in 1993-1994. Before the real sale, the auction mechanism elaborated by the FCC game theorists was tested through a series of experiments, which turned out to be extremely useful to correct and fine-tune the mechanism.

In chapter 9 Guala returns to the question of specifying when external inferences are valid. This happens when the experiment at issue and the real-world situation we are interested in are *analogous*, that is, when the causal mechanism generating the phenomenon in the laboratory also generates it in the field, and the differences between lab and field conditions are not relevant. In other words, external inferences are valid when the experiment and the real-world situation are analogous. But, again, when does such analogy exist?

This is an empirical question, to be addressed locally (we are interested in the analogy between a specific experiment and a specific real-world situation), and whose answer is fallible. Put differently, this is again a problem of inductive logic, and Guala coherently applies to it the principle of objective inductive support proposed in the first part of the book. Now the hypothesis H at stake is that there exists an analogy between the experiment and the real-world target. The evidence e is the correspondence between the *observable* features of the experiment and the *observable* features of the target. The analogy hypothesis H is objectively supported by e if there is a high probability of observing e when H is true, whereas there is a low probability of observing e when H is false. Moreover, the differences between the experimental conditions K_E and the real-world conditions K_W must be so well under control that it is highly unlikely that the correspondence between the observable features of the experiment and the observable features of the target is due to the K_E and K_W conditions.

Hence, not only the internal inferences from experiment to theory, but also the external inferences from experiment to real-world become reliable by eliminative induction, that is, by checking for and eliminating the auxiliary factors that may produce an apparent analogy between the experiment and the real-world situation. In Guala's words:

> External validity inferences are inferences to circumstances that we *know* to be different in some respects from the experimental situation. In order to make such inferences reliably, we must ask (and check) whether the differences between the experimental and the target system can confound the external validity inference or not […]. The logic of inductive inference is eliminative, both *within* and *from* the experiment.
>
> (pp. 197-198)

In chapter 10 Guala elaborates on the idea that experiments somehow mediate between theory and the real world, and discusses the relationships among experiments and other theory-world mediators like models and simulations. Finally, in chapter 11 Guala criticizes the conventional view according to which monetary incentives are a necessary requirement for running reliable economic experiments. Guala notices that the cognitive/psychological mechanisms triggered by money are not completely clear, and that the impact of alternative incentives on behavior have not been adequately investigated, so that «it is hard to establish that monetary incentives are good *in general*» (p. 245). Whether money is the appropriate incentive is again an empirical question, whose answer depends on the specific hypothesis at issue, the specific experimental design, and the particular real-world target of the experiment.

3. Two open questions

Guala's is a stimulating and useful work, and the way the author assesses the role of experiments within his theory of inductive inferences is largely convincing. Still, in my opinion there are two issues that the book does not address satisfactorily and that could be seeds for further research.

First, Guala's principle of objective inductive support demands a more analytical presentation and a more detailed discussion. How can we state this principle in terms of conditional probabilities? Can we model more precisely the way experiments allow us to disentangle the degree of support provided to hypothesis H by evidence e, from the support provided to H by e and the auxiliary conditions K? A criterion of inductive support that is widely used in science and also in experimental economics is the Neyman-Pearson statistical test. The Neyman-Pearson criterion includes subjective elements in the choice of the error probabilities (the probability of rejecting H when it is true, and that of accepting H when it is false), but it is far less subjective than the Bayesian criterion. What is the relationship between Guala's principle of objective support and the quasi-objective Neyman-Pearson test? Why should experimenters prefer Guala's vague criterion to the precise one offered by Neyman and Pearson?

These questions are not addressed in the book. Guala affirms that, since his work is not «a book on induction», he keeps «the discussion at a rather general level» and does not «make any commitment on a number of specific issues of inductive inference» (p. 118). Yet, since the book proposes a specific theory of inductive inference, a more detailed presentation of its central principle would have been opportune.

Second, Guala's theory of scientific inference does not account for the fact that, at least in economics, it is not uncommon that a hypothesis/theory is not abandoned even if it is amply disconfirmed by robust experimental evidence. Let us consider the case of Expected Utility Theory (EUT). Since the 1950s a large amount of experimental evidence has been collected showing that individuals in many relevant circumstances do not behave as expected-utility maximizers. Indeed, in the last thirty years decision theorists have proposed various alternative models of choice under uncertainty, but it is difficult to find them used in other branches of economic theory or in applied economics (for instance, in game theory strategic uncertainty is modeled almost exclusively through EUT). Why have economists not abandoned EUT and embraced some alternative theory of choice under uncertainty?

The case of EUT is in effect no exception. It can be argued that, apart from market experiments like those performed by Vernon Smith, most experimental evidence collected in the last fifty years is at odds with standard economic theory. Yet, it is difficult to indicate a part of economic theory that has been dismissed because it was in contrast with experimental evidence. My point is that Guala's view of experiments and scientific inference does not explain such scant influence of the experimental evidence in economics.

I deem this is because Guala neglects the circumstance that a scientific theory/hypothesis H_0 is evaluated not only with respect to the empirical evidence *e* collected under various conditions *K*, but also and always by comparing it to an alternative theory H_1. In other words, scientific support seems to be a four-place relation among H_0, H_1, *e* and *K*, rather than a three-place relation among *H*, *e*, and *K*. In particular, it could be that H_0 possesses certain epistemic qualities that economists consider highly desirable – *e.g.* simplicity, tractability, determinateness of implications, or unifying power – that instead lack in H_1. In such cases economists typically stick to H_0 even if experimental evidence disconfirms H_0 and supports H_1. Such methodological habit seems in fact to explain why EUT and other parts of economic theory survived experimental disconfirmation.

To be fair, this is a issue of theory choice rather than one of evidential support. Still, I think that a theory of scientific inference and experimentation, and especially one focused on economics like Guala's, should investigate it.