Experimental Localism and External Validity

Francesco Guala[†‡]

Experimental "localism" stresses the importance of context-specific knowledge, and the limitations of universal theories in science. I illustrate Latour's radical approach to localism and show that it has some unpalatable consequences, in particular the suggestion that problems of external validity (or how to generalize experimental results to nonlaboratory circumstances) cannot be solved. In the last part of the paper I try to sketch a solution to the problem of external validity by extending Mayo's error-probabilistic approach.

1. Introduction: Latour's "Hasty Generalizations." In *The Pasteurisation of France* (1988) Bruno Latour reconstructs Pasteur's struggle to enroll the scientific community in his research program of experimental microbiology and extensive vaccination. Latour aims at explaining Pasteur's success, a success that—he suggests—went right from the start well beyond what was justified by (common, or reasonable) scientific standards. Latour takes side with the dissenters, like Koch and Peter, who criticized Pasteur's "hasty generalization" from a bunch of vaccinated sheep to "a *general* method, applicable to all infectious diseases" (1988, 29). The skeptics questioned the stock of empirical facts upon which Pasteur based his inferences. According to Latour, "no one can deny that in 1881 this stock was extremely limited" (1988, 30). What Pasteur had done, in other words, did not support the generalized theories of infection and cure that he in fact formulated. The upshot is that Pasteur's "generalizations" were just rhetoric. In reality, Pasteur simply "exported" his findings by shaping reality on

[†] To contact the author, please write: Centre for Philosophy of the Social Sciences, University of Exeter, Amory Building, Rennes Drive, Exeter EX4 4RJ, Great Britain; e-mail: f.guala@ex.ac.uk.

[‡] Financial and logistic support from the Cognitive Science Laboratory of the University of Trento, where this paper was written, is gratefully acknowledged. Matteo Motterlini provided useful comments on an earlier version—of course all the remaining mistakes are mine.

Philosophy of Science, 70 (December 2003) pp. 1195–1205. 0031-8248/2003/7005-0028$10.00

the model of his laboratory of the Rue d'Ulm. Where this was not done properly, the "generalizations" failed.

> It often seemed for instance, that the antianthrax vaccine refused to pass the Franco-Italian border. However much it tried to be "universal," it remained local. Pasteur had to insist that the practices of his laboratory be repeated exactly if the vaccine were to travel. (Latour 1988, 93)

Latour takes the 1881 experiment at Pouilly le Fort as a paradigmatic case: its successful outcome was achieved by turning the farm into a laboratory. Where reality has not been carefully engineered, according to Latour, scientific knowledge does not apply. Experimental results travel from lab to lab, but never really get to grips with the "outside world." Here are a couple of quotes, in typical Latourian aphoristic style:

> When people say that knowledge is "universally true," we must understand that it is like railroads, which are found everywhere in the world but only to a limited extent. To shift to claiming that locomotives can move beyond their narrow and expensive rails is another matter. Yet magicians try to dazzle us with "universal laws" which they claim to be valid even in the gaps between the networks. (1988, 226)

> Whatever is local always stays that way. (1988, 219)

I shall call "*radical localism*" the view that experimental results do not apply to the world out of the laboratory.[1] There is a family resemblance between Latour's radical localism and the views of other philosophers of experiment. Ian Hacking, in an attempt to downplay the importance of high-level theories, points out that in "normal science" experiments are typically concerned with the measurement of a parameter, the replication of a phenomenon, the reduction of "noise," etc. Hacking calls the hypotheses involved in tasks like these, "*topical hypotheses,*" "to connote both the usual senses of "current affairs" or "local," and also to recall the medical sense of a topical ointment as one applied to the surface of the skin, i.e., not deep" (Hacking 1992, 145). Nancy Cartwright (1999), similarly, argues that the success of experimental science (the ability to explain, predict, and control phenomena) vindicate local, context-specific models, rather than general theories. These views seem milder, but in fact point in the same direction as Latour's: if general theories are less important than traditionally assumed, how are experimental results generalized, if at all? According to the standard view of theories, experiments provide evidence confirming general theories, so that these, in turn, can be used to explain other phe-

[1] Cf. also Latour (1987, 247–254). David Gooding (1990, ch. 6) and Andy Pickering (1995) at times seem to defend similar views.

nomena in their unrestricted domain of application. Theories transform the particular into general, because they tell us more about the world than the limited set of data that has been used to justify them. But if we are to take the above accounts of scientific practice seriously, the extension of laboratory knowledge must be much more complicated than that.

2. External Validity. Radical localism is surely a sort of skepticism, but quite different from the mix of relativism and antirealism customarily associated with the sociology of science. More precisely, it is an anti-inductivist kind of skepticism. But instead of relying on the *logical* problem of induction, radical localism uses the history of science to argue for the de facto or practical difficulty of generalizing from laboratory to nonexperimental circumstances.

Despite (or perhaps because of) its apparent descriptive accuracy, radical localism is extreme and disturbing. To be sure, *some* science travels from lab to lab without ever being faced with unconstrained reality.[2] But not *all* science works that way, and indeed scientific knowledge would be a poor thing if it were limited to that. Here's the disturbing aspect of radical localism: we would be disappointed, were we to find out that physiologists experiment with drugs on animals, but will never be able to tell whether these drugs can cure us (humans).[3] And similarly, we would be disappointed were we to find out that economists' experiments can teach us nothing about the working of real-world economic systems. Radical localism must face a normative challenge: as worldly decision makers, we *require* experimental knowledge to apply outside of the laboratory.

But this does not constitute a proper answer yet. If we take localist accounts of scientific practice seriously, we have a problem to solve. The problem of generalizing experimental results arises from a tension between, on the one hand, our desire to understand natural and social phenomena and, on the other hand, the fact that such phenomena usually take place in circumstances in which it is difficult to collect useful information about them. Our most reliable knowledge is achieved in the lab, where simpler, more manageable, and indeed rather special circumstances are created "artificially." Experimenters in the human sciences use a special terminology to capture this tension, by distinguishing between the "internal"

[2] Hacking (1992) speaks of a process of "self-vindication": in "mature" experimental science (by which he means basically some areas of physics), problems are set in the laboratory, solved in the laboratory, and the solutions appraised in the laboratory.

[3] Not to mention the fact that animal experimentation surely would be morally outrageous, if it were not able to help in the cure of human beings. See LaFollette and Shanks (1995) for a critique of animal experimentation along this line.

and the "external validity" of experimental results.[4] *Internal* validity is achieved when the structure and behavior of a laboratory system (its main causal factors, the ways they interact, and the phenomena they bring about) have been properly understood by the experimenter. For example: the result of an experiment E is internally valid if the experimenter attributes the production of an effect B to a factor (or set of factors) A, and A really is the (or a) cause of B in E. Furthermore, it is *externally* valid if A causes B not only in E, but also in a set of other circumstances of interest, F, G, H, etc. Problems of internal validity are chronologically and epistemically antecedent to problems of external validity: it does not make much sense to ask whether a result is valid outside the experimental circumstances unless we are confident that it does therein.

Philosophers of science have paid relatively little attention to the internal/external validity distinction. If you have been trained in the empiricist tradition, it is not obvious whether the distinction is philosophically legitimate in the first place. If you believe that science is ultimately devoted to the discovery of nomic generalizations, and that laws are (deterministic or probabilistic) regularities between events, then the two problems of validity lose most of their significance. But consider the following (fictional) example: a psychologist attributes the behavioral response (B) of his subjects to the experimental treatment (A), whereas in fact it is due to another feature of the experimental conditions (C), such as for example an unintended "demand effect." According to the Humean conception embodied in the Standard View, the problem with the A-B relationship is that A is not regularly associated with B, or in other words, that "For all x, if Ax then Bx" is not a genuine (general) law of nature. But exactly the same charge can be raised against the C-B relationship: outside those specific experimental conditions, B is neither associated with A, nor with C. Yet, we are here dealing with two different kinds of failure: one is a failure of causation, the other one of generality. The incapability of distinguishing between these two failures—which are, in contrast, appropriately distinguished by scientists—is a shortcoming of the Humean conception of science.

Although I shall not defend this point here, a main assumption of this paper is that in this case (as in many others) we should side with the scientists. Thus, to begin with, we need a non-Humean account of causation, able to capture the conceptual distinction between mere associations and genuinely causal relations. Whereas internal validity is fundamentally a problem of identifying causal relations, external validity involves an inference to the *robustness* of a causal relation outside the narrow circumstances

[4] This terminology is popular in experimental psychology and in some branches of the social sciences. Cook and Campbell (1979) provide a classic discussion.

in which it was observed and established in the first instance. Thus, secondly, we must make sure that generality requirements are not smuggled in as part of the definition of causation. Knowledge of general causal relations is obviously useful and to be sought for pragmatic reasons, but local causal knowledge is causal knowledge nonetheless (cf. also Hausman [1998, 186–191, 224–227]). By following this strategy, both problems of validity can be rehabilitated.

3. Mayo's Error Probabilities. What is so special with laboratory experimentation? What is it that makes scientific investigation and inference relatively easier inside, and more problematic outside the lab? Roughly, a good experimental design is one that maximizes control. This idea is very familiar to scientists—here's the definition of "control" from a popular textbook on research methods in the social sciences:

> Control: a procedure designed to eliminate alternative sources of variation that may distort the research results. (Frankfort-Nachmias and Nachmias 1996, 587)

But this is not rigorous enough. Deborah Mayo (1996) tries to capture practitioners' intuition, as in the above quote, by means of the notion of "severe test." A severe test is meant to eliminate error (the "alternative sources of variation that may distort the research results"):

> An experiment E is a severe test of hypothesis H if and only if H implies observation O, and there is a very low probability of observing O in the event that H is false.

H is a "topical" hypothesis, Hacking's style, about the occurrence of a certain kind of error. For example, H = "the measured value X* of variable X is an artefact of the measurement instrument." A good experimenter should try to minimize the chance of making the sort of error specified by H, for example, by using different independent instruments in order to determine X*. Such a procedure, in case of approximately convergent results, triggers an argument from coincidence that is very familiar to realist philosophers of science: given that the measurement instruments rely upon independent assumptions and independent mechanisms, it would be an extremely unlikely coincidence to observe again and again identical (or very similar) values for X*, if X* was an artefact of the measurement procedure.[5]

What do we learn from experiments, then? According to Mayo, a passing test for H has limited ("local") positive implications. We may learn, in

[5] To be sure, Mayo grounds the normative appeal of no-miracle arguments on error-probabilistic reasoning. We shall bracket such issues here.

the example above, that X* is not the artefact of a specific measurement apparatus, but is robust to different procedures of estimation. This is not such a great achievement, one might say, because there may be *other* errors affecting our result. But in order to check for those mistakes, we need different (error-specific) controls. Normally, it will be impossible to implement all such controls at once—we learn little by little, so to speak. Mayo puts it this way:

> What does it mean to learn that H is indicated by the data? It means that the data provide good grounds for the correctness of H—good grounds that H correctly describes some aspect of an experimental process. What aspect, of course, depends on the particular hypothesis H in question. One can, if one likes, construe the correctness of H in terms of H being reliable, provided one is careful in the latter's interpretation. Learning that hypothesis H is reliable . . . means learning that what H says about certain experimental results will be often close to the results that would actually be produced—that H will or would often succeed *in specified experimental applications*. What further substantive claims are warranted will depend on the case at hands. (Mayo 1996, 410; my italics)

Notice that this approach does not allow inferences to what will happen *outside* the relevant class of experimental circumstances. This is of course very much in the spirit of localism. You cannot extend to human beings the results of experiments on mice, for example, unless you have good (experimental) grounds to believe that certain differences between the anatomy of mice and human beings do not matter (i.e., they are not error-generating differences). But remember that external validity inferences are inferences to circumstances that we *know* to be different from the experimental situation in some respects. In order to make such inferences reliably, we must ask (and check) whether the differences between the experimental and the target system are error-generating or not. What kind of error is an external validity error? And what kind of device would minimize its occurrence?

4. From the Laboratory to the Outside World. External validity errors can be of various sorts. One may observe phenomenon Y in the lab, and incorrectly infer that the same phenomenon takes place (or can take place) also in a given field setting. Or one may establish that X causes Y in experiment E, and erroneously infer that X causes Y also in nonlaboratory circumstances F, G, etc. First, consider that in order to specify the error, one must specify a target. If you worry that Y may not occur out of the lab *generically,* there is little you can do to figure it out. Instead, scientists usually worry about the extension of an experimental result to a specific target: to a population of patients who are ill from a given disease, for example, or to a certain kind of economy with specific characteristics. The obvious thing to do, then, is to go out and have a look: if you observe the right sequence of Xs and Ys in the target, for instance, you will be encouraged to believe that, since X causes Y in the lab, it does the same in the target. The temptation to frame this approach in terms of "analogical inference" is strong. In biomedical research, for instance, the inference would take this form (adapted from Thagard 1999, 140):

(1) Humans have symptoms (disease) Y.
(2) Laboratory animals have symptoms (disease) Y.
(3) In animals, the symptoms (disease) are caused by factor (virus, bacteria, toxin, deficiency) X.
(4) The human disease is therefore also caused by X.

This is an *analogical inference*.[6] Analogical reasoning is often deemed "risky" and, therefore, merely "heuristic" (able to suggest, rather than establish, hypotheses). In part, this is just another way of saying that analogical reasoning is fallible. But all inductive inferences are fallible (otherwise, they would be deductive rather than inductive in the first place), and external validity inferences surely involve an inductive step. The point is rather: how do we distinguish reliable from unreliable inferences? Drawing positive analogies is basically a mapping procedure, where elements of a set or properties of an object are put in correspondence with elements or properties from another set or object. But every object is similar to every other object from an infinite number of respects. There is potentially an infinite number of mappings to be drawn, many of which will be uninteresting or even misleading from a scientific viewpoint.

To put it another way, consider the role played by statistical regularities in *internal* validity inferences. In the right circumstances, for example in the context of a well-designed experiment, statistical associations may constitute evidence for causation. But the very same correlations observed in uncontrolled circumstances do not bear equal weight. One thing is to observe that factors X and Y are regularly associated in the field, where many (uncontrolled) factors could have been responsible for their instantiation. Quite another is to "trigger" X and observe Y in the context of an accurately designed experiment, where the "other factors" have been appropriately shielded or controlled for. A correlation between X and Y provides evidence for the hypothesis that X causes Y only if we have made sure (by means of appropriate experimental design and data-analysis) that

[6] See also LaFollette and Shanks (1995, 147) for a similar analysis.

the correlation could emerge from *that* data-generating process only. The circumstances matter: the same piece of data can bear different weight depending on whether the background circumstances are "right" or not. The same moral applies to external validity inferences: analogical correspondences are not enough. We need "good" analogies—but what makes a "good" analogy in the first place?

The idea, roughly, is that we need to create (or select) circumstances in which it is really unlikely to observe certain data, unless the external validity hypothesis is true. In this case, the data is the correspondence between observed features of the target and observed features of the experimental system; the external validity hypothesis is that the relata belong to similar causal mechanisms. Now, the probability of observing such a correspondence (were the hypothesis false) is low if we have eliminated alternative reasons why such a correspondence might occur, other than the causal similarity between the two systems. Notice how some typical localist themes emerge once again: if you want to generalize from A to B, you better make sure that A and B are as similar as possible. This is in fact the logic underlying the best-known external validity control—representative sampling. If you want to generalize to population B, you better make sure that you have in the lab good representatives of the individuals in B (you need students if you want to generalize to students, housewives for housewives, mammals for mammals, etc.). But unfortunately in most cases this will not be the end of the story: experimental conditions include not only a pool of subjects, but also a number of environmental factors, treatments, and boundary conditions in general. In many cases, it is the environment and the treatment that worry us the most. (Think at the stylized, abstract tasks of experimental cognitive psychology, for example.) Representative sampling will not solve these problems, but other devices based on the same logic will be of some help. Let me show how by means of an example.

5. An Example from Economics. I suppose most people do not realize that economists do laboratory experiments. The reason to choose a case from such an odd discipline is precisely that experimental economics has been criticized ever since its birth by means of arguments of this sort: "Okay, this is what happens in your lab economy, but surely you don't think that *real* economies work that way?!" Of course there is no way of answering such a question, unless it is made more precise. What is the target here? To what sort of settings is a given experimental result supposed (not) to be applicable? And what kind of mistake are we worrying about? So let us take a concrete case and a concrete challenge. "Preference Reversals" (PR) are one of the most debated anomalous phenomena discovered (or "produced," if you like) in the economic laboratory. Without going into the details,[7] PR are a form of apparently intransitive behavior: many subjects choose a lottery X over another lottery Y (X and Y have some peculiar monetary payoffs, with peculiar probabilities), but then assign a higher price to Y than to X. If both pricing and choosing are ways of manifesting one's preferences, then preferences must be intransitive, contrary to standard economic theory.

Research on PR has gone through two stages: initially, experimenters were concerned with establishing the reliability of the methods used to elicit preferences via choosing and pricing; then, they turned to the robustness of the PR phenomenon in different settings. I will not comment on experiments of the first kind, because techniques for the elicitation of preferences are complicated and an adequate discussion would take too much space.[8] My concern here is with the second kind of research. The idea is that PR might be (or, for some economists, *ought* to be) a purely laboratory phenomenon. In real economies, surely intransitivities cannot happen. In order to test such a hypothesis, some experimental economists investigated the robustness of PR to repetition, trading, and arbitrage. Repetition is used to observe how long it takes before subjects realize their "mistake" and revise their preference ordering. Trading is used to investigate whether certain market institutions, like the English auction for example, manage to reduce inconsistencies in the valuation of lotteries. And arbitrage is used to check whether "money-pumping" is an effective way of reducing PR. It turns out that no one of these factors alone is able to eliminate PR completely. Repetition and arbitrage together, however, pretty much do the job (cf. Chu and Chu 1990). But what sort of lesson is this supposed to teach us?

The argument is this: "real markets" involve repeated choice, under specific institutional rules, and are populated by arbitrageurs. Therefore, by making the classic PR experiment more similar to "real markets," we test the external validity of the PR phenomenon. I have put "real markets" between scare-quotes because not all *real* markets are of this kind. The real estate market, for example, is populated by many traders who will not engage in that sort of transaction more than once or twice in their whole life. Mostly, the price is determined by a first-price sealed-bid mechanism. And most traders do not have a chance to learn that their preferences are inconsistent by being repeatedly money-pumped (thank God, one might say!). Thus, the external validity of the experiments used to test the robustness of PR will not stretch to these circumstances. Their results will

[7] Cf. Thaler and Tversky (1990) for a more detailed discussion and review of the literature.

[8] But see Hausman (1992, ch. 13) and Guala (2000) for a methodological analysis of these experiments.

be applicable only to real-world economies with repetition, arbitrage, and the English-auction mechanism.

The target in the example above is defined implicitly by the features of the experiments. In other cases, we might start the experimental investigation with a concrete, specific target in mind—say, the market for oil leases, or the market for mobile phone licences. In 1986 two economists, John Kagel and Dan Levin, showed that an alleged market failure in the auctions for oil leases in the Gulf of Mexico could be replicated in a laboratory experiment. Their result was convincing because the setting was very similar to the one used by the Outer Continental Shelf to sell real leases. The auction mechanism was the same, uncertainty about the value of real leases was simulated effectively, professional executives participated as subjects. But the similarity between lab and target setting can also be obtained by working in the opposite direction. Many cases of economic engineering are of this latter sort: the auctions for "third generation" mobile phones were designed and accurately tested in the economic laboratory at Cal Tech, before being "exported" in the real world. They did not exist as a "naturally evolved" entity before the experiments took place.[9]

"Exporting the lab" is the radical localist solution to the external validity problem. The point is that it is just one route to external validity—the safest one perhaps, but not the *only* one. Its viability depends on how much we are allowed to intervene and shape reality to fit the experimental prototypes. The standard sequence of trials to test drugs in experimental medicine is a good example of a compromise: experimenters start with animals,[10] move on to human beings in "ideal" experimental settings, and conclude with so-called efficacy trials with patients in more realistic conditions. Which does not mean that real-world conditions themselves cannot, sometimes, be modified so as to make the drug more efficacious or avoid unpleasant side-effects (that's what hospitals are for, among other things). Radical localism, à la Latour, exploits a correct insight. But radical localism captures just a special case of a more general methodology, that can be applied (with varying degrees of reliability) also when the laboratory cannot be exported, but must be adapted to the target at hands. The localist approach affords general methodological prescriptions, after all.

[9] The creation of "artificial" phenomena in the laboratory is highlighted in Hacking (1983). On the construction of the mobile phone auctions, see Guala (2001).

[10] I am simplifying drastically here: to find out which animals are "right" for which kind of investigation is not a trivial matter. On animal models in biomedical science, see La-Follette and Shanks (1995) and Ankeny (2001).

REFERENCES

Ankeny, Rachel (2001), "Model Organisms as Models: Understanding the 'Lingua Franca' of the Human Genome Project", *Philosophy of Science* 68 (Proceedings): S251–S261.

Cartwright, Nancy (1999), *The Dappled World*. Cambridge: Cambridge University Press.

Chu, Y. P., and R. L. Chu (1990), "The Subsidence of Preference Reversals in Simplified and Marketlike Experimental Settings: A Note", *American Economic Review* 80: 902–911.

Cook, Thomas, and Donald Campbell (1979), *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Gooding, David (1990), *Experiment and the Making of Meaning*. Dordrecht: Kluwer.

Guala, Francesco (2000), "Artefacts in Experimental Economics: Preference Reversals and the Becker-DeGroot-Marschak Mechanism", *Economics and Philosophy* 16: 47–75.

——— (2001), "Building Economic Machines: the FCC Auctions", *Studies in History and Philosophy of Science* 32: 453–477.

Hacking, Ian (1983), *Representing and Intervening*. Cambridge: Cambridge University Press.

——— (1992), "The Self-Vindication of the Laboratory Sciences", in A. Pickering (ed.), *Science as Practice and Culture*. Chicago: University of Chicago Press.

Hausman, Daniel (1992), *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

——— (1998), *Causal Asymmetries*. Cambridge: Cambridge University Press.

Kagel, John, and Daniel Levin (1986), "The Winner's Curse Phenomenon and Public Information in Common Value Auctions", *American Economic Review* 76: 894–920.

LaFollette, Hugh, and Niall Shanks (1995), "Two Models of Models in Biomedical Research", *Philosophical Quarterly* 45: 141–160.

Latour, Bruno (1987), *Science in Action*. Cambridge: Harvard University Press.

——— (1988), *The Pasteurisation of France.* Cambridge: Harvard University Press.

Frankfort-Nachmias, Chava, and David Nachmias (1996), *Research Methods in the Social Sciences*. London: Arnold.

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

Pickering, Andrew (1995), *The Mangle of Practice.* Chicago: University of Chicago Press.

Thagard, Paul (1999), *How Scientists Explain Disease*. Princeton: Princeton University Press.

Thaler, Richard, and Amos Tversky (1990), "Preference Reversals", *Journal of Economic Perspectives* 4: 201–211.