Some Methodological Issues in Experimental Economics

Deborah Mayo†

The growing acceptance and success of experimental economics has increased the interest of researchers in tackling philosophical and methodological challenges to which their work increasingly gives rise. I sketch some general issues that call for the combined expertise of experimental economists and philosophers of science, of experiment, and of inductive-statistical inference and modeling.

1. Introduction. The goal of the symposium out of which our group of papers emerged is to introduce, within a Philosophy of Science Association forum, some of the philosophical and methodological issues relating to current work in experimental economics, which I abbreviate throughout as *expereconomics*. As our proposal for this symposium remarks: "Experimental economics is an interesting arena because its methodological conventions are still relatively fluid, and thus there seems to be a genuine opportunity for philosophers of science to contribute to the debate and ultimately [hopefully] improve the methodological practice in the field. Conversely, experimental economists seem to have come across methodological problems that call for more attention on philosophers' part. To sum up, we see the exchange between experimenters and philosophers very much as a two-way street." My aim will be to survey some broad philosophical and methodological questions that may help in developing a road map for this two-way street.

2. Mature Expereconomics Is Ripe for Methodological Attention. Growing attention to foundational problems among expereconomists is often traced to the field's success and growing acceptance. "As a young discipline, experimental economics struggled to survive within a generally skeptical and dismissive scientific community. Downplaying difficult methodolog-

†To contact the author, please write to: 235 Major Williams, Virginia Tech, Blacksburg, VA 24061-0126; e-mail: mayod@vt.edu.

Philosophy of Science, 75 (December 2008) pp. 633–645. 0031-8248/2008/7505-0012$10.00

ical problems, therefore, may have been a simple survival strategy" (Guala 2005, 159). This "survival strategy," many now recognize, has led expereconomists to "lag behind their colleagues in other disciplines in discussing important methodological issues. Whereas methodological issues are a standard part of almost any curriculum in psychology, they are virtually missing in the education of [experimental] economics" (Schram 2005, 225). With expereconomics now at the heart of some of the most dynamic new developments in economics, it is increasingly recognized that the time is ripe to openly acknowledge, and attempt to remedy, methodological shortcomings in expereconomics research.

3. Methodological Validity: Who Bears the Burden? A key aim in introducing experimentation into economics is to enable researchers to probe implications of economic theories in a more controlled manner than is possible from collecting economic data and recording economic behavior as they naturally occur. Achieving this aim, however, depends on the relevance of the growing body of specially triggered lab phenomena to the implications of economic theories. This gives rise to what is often called the problem of "external validity": How can we generalize from the apparently artificial environments of the expereconomics lab to "real-world" economic phenomena?

A common "textbook" response, containing more than a twinge of defensiveness, is that "an honest skeptic of external validity bears the burden of guessing what makes the lab environment substantially different than the real world" (Friedman and Cassar 2004, 29). The typical situation where this issue arises is when lab results apparently conflict with behavior predicted by a received economic theory *T*. The defenders of theory *T* may deny that the experimental lab result is sufficiently relevant to the predicted behavior and thereby deny that the results constitute an anomaly for *T*. The textbook strategy is for the expereconomist to throw the ball back to *T*'s defenders and ask them to furnish specific grounds for why theory *T* should not be expected to hold for the experiment in question. If, for example, theory *T* has implications for buying behavior, then it would seem to apply to buying mugs or chocolates in the expereconomist's lab; so conflicts between *T* and buying behavior observed in the lab would seem to count as genuine anomalies for *T*. It is the defenders of *T* who bear the burden of saying what factors prevent the lab behavior from validly testing *T* if any weight is to be given to their denial that a genuine anomaly for *T* is at hand—at least according to the textbook recommendations. But this seems too strong. Granted, it would be unwarranted for the defense of *T* to boil down to protecting *T* from any and all anomalies, but to place the burden of proof entirely on the critic's shoulders goes too far in the opposite direction. It is true that if the critic identifies a factor thought to invalidate the anomaly, the experimenter can use this constructively to improve the experiment, and the expereconomist's willingness to do so is all to the good. But what if the critics do not pinpoint a factor? Should the expereconomist's construal of the lab result (as genuinely anomalous for *T*) thereby stand? To claim to have evidence for an effect simply because no critic has pinpointed why lab results lack external validity would itself be to commit a fallacy of "argument from ignorance." So how to respond to "external validity" challenges is one issue to be addressed in expereconomic methodology.

4. Getting Beyond an (Alleged) Distinction between Deductive Testing and Inductive Inference. To address the issue, it is useful to understand how this dispute about who bears the burden of proof connects to a common distinction expereconomists tend to make between *testing* theories—which is regarded as hypothetical-deductive—and *inductively arriving* at and appraising empirical regularities and experimental claims. External validity challenges are thought to apply only or mainly to the latter, not to the former. In fact, it is often said that the central achievement of expereconomics is to have shifted attention to testing theories since (deductive) testing frees expereconomists from having to capture aspects of genuine economic phenomena: "Scepticism about the external validity of experiments was countered by a strategy of argument that has come to be called 'Blame the theory', perhaps first diagnosed by Chris Starmer (1999). This is to point out that any unrealism of the laboratory environment mirrors the unrealism of the theory being tested" (Sugden 2008, in this issue). If a deductively derived hypothesis from a received theory *T* disagrees with the experiment, then *T* requires modification with some auxiliary hypothesis that, conjoined with *T*, can be tested next. From this perspective, the onus is on the defender of *T* to supply the auxiliary in terms of an experimental design that more nearly captures the artificiality of the theory.

However, as expereconomists accumulated more and more of what Sugden (2005) calls *exhibits*—replicable experimental designs that reliably produce interesting results—they began to move beyond theory testing to investigations of these exhibits in their own right. These investigations concern the *inductive* tasks of generating and explaining the anomalous effects themselves. With this inductive task, expereconomists are prepared to accept the burden of methodological critiques about the external validity of their hypotheses; hence, we should expect texts to increasingly move away from the still-common defensive standpoint.

In truth, both the testing of theories and investigating hypotheses about experimental exhibits involve inductive or "ampliative" inference in the sense that both require going beyond the data to experimental hypotheses that generalize or explain the experimental effect. These experimental hypotheses generally concern reliably produced experimental effects (exhibits) that are anomalous for preexisting economic theories; thus, purported explanations of these effects are sometimes called "deviation theories" or hypotheses (Sugden 2005). However, if the theory says something about, say, aspects of preferences, then if one does not have evidence for a genuine regularity concerning preferences, then one does not have an anomaly regarding the theory. Or again, if there are legitimate questions about, say, the inductive exhibit involving the "endowment effect," then there are legitimate questions about whether a genuine anomaly had been found to begin with. So issues of inductive validity (of which external validity is only one) arise both in testing received theories and in the more current tasks of explaining anomalous exhibits. (Admittedly, certain hypothetico-deductive conceptions of testing among philosophers may have encouraged the distinction I am now questioning.)

Instead of seeing testing as *deductive* and *induction* as distinct from testing, I propose that we view experimental learning as inductive testing of various claims. What gets distinguished is the nature of the hypothesis or inference. For example, at some stages of an inquiry the concern might be to infer hypotheses that the effect is genuine (not an artifact) and reliably reproducible. In yet others, the interest might be in whether the genuine effect is due to some factor(s), whether it is adequately captured by means of a given model of the data generation procedure, or whether aspects of the parameters of the model may be estimated as having certain values. Thus we can consider methodological issues regarding (inductive) experimental inferences in expereconomics all together. Rather than be straitjacketed by some outmoded conceptions of testing (e.g., as purely deductive), we can organize the different issues in terms of different problems or questions.

5. A Framework of Experimental Problems and Hypothesized Solutions.

To have a robust framework in which to house the variety of methodological issues in expereconomics, it suffices to erect a very general organizing scheme:

1. Experimental inquiry starts with a *main problem or question* and sets out to develop and probe various hypothesized solutions or answers. Examples of possible questions are (*a*) Is this experiment a genuine application of theory *T*? (If not, why not?) (*b*) Is this a genuine effect? If so, is it a genuine anomaly for theory *T*? (*c*) How can we discriminate rival hypothesized explanations for the anomaly?

2. Different threats of error arise in relation to each problem, and corresponding methodologies for checking and correcting claims may be developed. Collecting a reservoir of "real effects" or "exhibits" may be of great interest in its own right, as we increasingly see in expereconomics. Without anything more elaborate, one may set sail discussing the various problems and issues that would arise in developing an adequate foundation for the methodology of expereconomics.

5.1. Methodological Critique and the Burden of Proof. Consider the question raised earlier about "burden of proof." We can group together what is required for a critic to warrantedly block an inference *H* from an experiment, whether in the course of defending theory *T* against an alleged anomaly or otherwise. A critic cannot legitimately block *H* by declaring, in effect, that no matter how rigorous the tests that *H* passes, everything is just as if *H* were true or adequate when in fact *H* is false. That would be a highly unreliable procedure (that may take the form of what I sometimes call "gellerization" or "rigging"). However, a legitimate critic need not bear the burden of identifying specific factors that invalidate the experimental inference to *H* in order to block it—however useful such suggestions might be in improving tests. If all the critic is doing is denying that the experiment supplies evidence for some *H*, it suffices to raise certain types of methodological questions. In particular, it suffices to identify what might be called "canonical" flaws that in general create obstacles to the inferential moves on which the given inference depends (at any of the stages from data collection, modeling, and interpretation). The onus is then on the experimenter to at least address these challenges, show how one gets around the obstacle in the particular experiment, or show how one might do so in an improved experiment.

I think it is preferable to speak of general methodological flaws or potential sources of invalidity or bias rather than keep to the (often ambiguous) external/internal validity terminology; and anyway, discussions of "external validity" do not seem to cover some of the main methodological worries that we wish to raise. Before giving examples in the realm of expereconomics, I can illustrate my point by reference to an epidemiological inference to some causal hypothesis about a risk factor, for example, hormone replacement therapy. Because it is known that certain biases and confounding effects can invalidate the inference at hand, the epidemiological (e.g., case control) researcher would normally address these concerns, even if the critic had not borne the burden of showing that any such bias invalidated the particular inference.

5.2. Expereconomics-Laden Effects. I shall identify in brief snippets some of the legitimate methodological concerns that might arise. Several of them connect to what might be called "expereconomic-laden effects." Expereconomists inductively arrive at hypothesis H, where H is designed to account for the observed "exhibits," which are reliably triggered by experimental designs; and then H passes tests by accounting for the data exhibited (perhaps by suitably adjusting a parameter in H). A legitimate question would be how the researcher is getting around known threats of self-sealing or circular inferences whenever this kind of insular loop seems to be involved. Until the question is laid to rest, the inference is in abeyance, which is not to say that there is evidence against H, only that more is required to have evidence for H. For example, having gotten good at generating evidence $\mathbf{x}$ of "ultimatum effects," a deviation theory H such as "inequity aversion" or "fairness" will pass tests by dint of being in accordance with further evidence $\mathbf{x}'$ of ultimatum effects. But is this really to probe the purported explanation? Is this insularity a problem? The concern is that even if H is false as an explanation of the anomalous exhibit $\mathbf{x}$, hypothesis H, having been designed to account for this exhibit, would be in accordance with any newly triggered $\mathbf{x}'$. To put this probabilistically, the concern is that $P(H \text{ fits } \mathbf{x}';\ H \text{ is false})$ is high.

Questions about expereconomics-laden interpretations of various effects give rise to some of the most serious, and also the most philosophically interesting, generic methodological challenges, although they are rarely put in just these terms. What counts as an adequate response will depend on what precisely is being inferred; developing responses to such questions would promote important distinctions about just which inferences are defensible.

6. Selection Effects and Data-Dependent Hypotheses in Expereconomics. The issue of how data-dependent hypotheses (also sometimes called "use-constructed" hypotheses) and data-dependent searches influence the validity of inferences is highly tricky and has long been the object of controversy. If one is allowed to search through several factors and report just those that show (apparently) impressive correlations or other effects, there is a high probability of erroneously inferring a real effect. In that case, the hypothesis of a genuine effect does not pass a severe test. However, it is equally clear that using the same data both to identify and to test the cause of, say, a plane crash may allow us to infer the cause (e.g., a particular type of bomb) with very high reliability. I discuss the general issue both within and outside of statistical contexts at length elsewhere (e.g., Mayo 1996, 2008; Mayo and Cox 2006); here I am just identifying it as a member of potential flaws about which one might raise methodological questions in expereconomics. Expereconomics also gives rise to some novel variations on the whole "double-counting" issue.

i) For example, data-dependent searches to find an experimental design that will reliably display an effect of interest do not count against the recipe ultimately found. (Whether the resulting "exhibit" is useful for one or another purpose is a distinct question.) "Trying and trying again" to learn enough to develop a reliable design for triggering an effect is no more problematic than, say, Jean Perrin's searching through 50 substances until hitting on gamboge on which to base Brownian motion experiments, or searching through numerous experimental animals until finally finding one—the New Zealand rabbit—that would show the known teratagenic effects of thalidomide. The hypothesis here, *H*, is a claim about the experimental design for reliably triggering the exhibit of interest. Data-dependent searches for the recipe in *H* can be seen to promote rather than discount the reliability of the resulting inference.

ii) Similarly unproblematic are data-dependent "stress tests," as Vernon Smith (2002) calls them, with the goal of deliberately trying to get an effect *not* to show up so as to better understand where both theories and experimental designs fail and explore the mechanisms at work. Data-dependent stress tests often serve to set reliable bounds on how far off experimental conditions can vary and still have a theory hold or an effect show up. Much work in expereconomics in fact seems to be directed at learning to trigger "exhibits," whether to better understand a theory, answer challenges to inferences based on less well understood experiments, or explore ways to distinguish the causes of anomalies.

7. Expereconomic-Laden Constructs. One of the things that introduces difficulties into methodological appraisals of expereconomic results is that at times an experimental effect is being used merely as an exhibit or reliably produced effect, whereas at other times it may be regarded as serving as part of an explanation for that effect. The assessment of how well warranted inferences are in these two cases will differ, particularly when data dependences and double uses of data enter. These difficulties are heightened by the extent to which these effects are often themselves complex experimental constructs. While these constructs are often developed to serve as quantitative consequences of substantive economic theories, questions about the relationship between the testable and theoretical notions often arise.

Consider the so-called *endowment effect*. This effect describes a well-known anomaly for "received" economic theories of consumer behavior. The anomaly is that people seem to demand more if they are selling an object with which they are endowed than if they are asked for the minimum amount of money they would prefer to have rather than receive the object. An example would be if I regard anything over 50 cents as preferable to receiving the "Virginia Is for Lovers" mug; but then once given the mug and asked for the minimum amount I would agree to sell it for, I require more than 50 cents, say I require $1.00. Sometimes 50 cents is called my "choice-equivalent" value and $1.00 my "willingness to accept" valuation. In probing this effect, experimental subjects can be buyers, sellers, or "choosers," and their valuations are used to quantify

> *Willingness to accept* (WTA): Minimum amount sellers demand to give up their good.
>
> *Choice equivalent* (CE): Minimum amount that "choosers" prefer to receiving the good.
>
> *Willingness to pay* (WTP): Highest amount buyers are willing to pay for the good.

A hypothesis about the endowment effect can be expressed in terms of quantitative parameter θ:

> H_1: $\theta > 1$,

where θ = WTA/CE (Bateman et al. 2005).

The complexities introduced can and do quickly grow in the expereconomic literature. Classifying types of methodological questions may make it easier for them to be systematically raised. To begin with, this example exemplifies a case in which there is an interest in discriminating rival explanations for an already accepted anomaly. That is, hypothesis H_1 is accepted as having been shown, and contrasting explanations are offered. In particular, one explanation for the anomaly expressed in H_1 is stated by means of hypothesis H_2:

> H_2: CE/WTP = 1, "no loss aversion in buying" (NLIB).

Hypothesis H_2 implies that buyers do not evaluate as a loss the money that is given up to purchase a good in a normal transaction, at least if it is within their budget. If in fact people do experience loss aversion for the money they spend, then we would expect

> H_2': CE = α_mWTP, where positive α_m is the "loss aversion" coefficient for money.

(Hypothesis H_2 was developed on the basis of results in which experimenters found α_m to be 1 or close to 1.) Hypotheses H_2 and H_2' correspond to rival hypotheses to account for the anomaly of interest, and an interesting "adversarial collaboration" was conducted to decide between them objectively (Bateman et al. 2005).

While I am leaving things at a sketchy level, the example illustrates the sense in which theories dealing with characteristics such as equity, fairness, reciprocity, and so on all have special meanings for the expereconomist, and there are a cluster of ingenious ways to render them concrete and measurable. Once immersed in the expereconomist paradigm, one begins "to speak like a native" (as Kuhn [1962] might say), and a philosophical critique would need to combine a degree of "immersion" with a meta-level scrutiny of the relationships between the various concepts, constructs, and experimental analogues. Things get further complicated in using experimental data to test rival hypotheses. The parameters of interest, such as θ, are embedded within statistical hypotheses, and the experiments designed to estimate values such as α attempt to capture the assumptions of the statistical models involved.

8. Statistical Issues in Interpreting Expereconomics Data. Whether the predictions concern individual choice or game-theoretic experiments, the testable predictions tend to be framed, analyzed, and interpreted statistically (via simple statistical significance tests, reporting *p*-values and/or confidence intervals). Thus, critiques of statistical methodology and concerns with statistical fallacies—topics of special interest to me—introduce a cluster of questions for expereconomic analyses in their own right.

Although expereconomists express the hope that their interpretations are sufficiently strong to sidestep subtle statistical debates (sometimes called the "interocular eye test"), the fact is that their discussions—for example, the adjudication of the fascinating "adversarial collaboration" above—rely heavily on discriminations afforded by statistical experiments. The researchers from the rival camps generally agreed that the statistics spoke in favor of H_2' rather than H_2, and various explanations were given. (For example, holders of H_2—the prediction associated with NLIB—suggested that the items purchased, mugs or chocolates, may have been perceived as outside the subjects' discretionary budget.) A distinct study, however, discussed in Novemsky and Kahneman (2005) is claimed to shore up support for H_2, in apparent conflict with the results of the adversarial collaboration experiments. But do the results really conflict? The evidence for H_2 is that it is *not rejected* at the .05 level. The predicted ratio (of 1) is not rejected by the data because 1 is included in the estimated 95% confidence interval (which has to be bootstrapped by computer). Even if one grants all the assumptions, interpretive questions arise which the rival group (in Bateman et al. 2005) could well have noted on their behalf. Most notably, the failure to find evidence against the null prediction (1) is not evidence *for* it, but at most allows setting bounds ruling out specifiable discrepancies from the null. In particular, the data provide evidence that the ratio is less than the upper confidence interval bound—provided that the statistical assumptions are at least approximately satisfied—at confidence level .95. Since the goal of the experiments and their statistical analyses was to adjudicate between holders of rival deviation theories H_2 and H_2', such metastatistical considerations would clearly be relevant. In my view, even minor improvements (in the statistical corner) would go a long way toward addressing day-to-day issues as to whether expereconomic exhibits are artifacts, how to improve designs, and how to use the available data more effectively in distinguishing rival interpretations of effects.

Perhaps Sugden's recent call for expereconomists to begin giving greater attention to small effects will awaken interest in putting statistical principles and critiques to work. A "deviation theory which isolates one causal mechanism will account for deviations that, on average, are smaller than those that are revealed in the exhibit" it tries to explain. These are "likely to be perceived as unexciting by referees, editors and readers" but may well be central to understanding well-known exhibits that may be induced by combinations of weak causal mechanisms. Detecting small effects requires that "experimentalists use samples that are large enough to detect weak effects" and/or (I am adding) more effective statistical analyses (Sugden 2005, 300).

9. How to Evaluate a Normative Rule of Inference. The development of an adequate methodology for expereconomics might contribute some new perspectives to ongoing interest in reconsidering the interpretations of some of the well-known classic experiments at the interface of psychology and economics. Experiments such as those performed by Kahneman, Slovic, and Tversky (1982), we know, reveal substantial deviations from the Bayesian model of "rational" inference even in simple cases in which the prior probabilities are given, and even with statistically sophisticated subjects. In dealing with such anomalies, the focus has often been on altering the experimental design or educating the subjects to get their experimental judgments to conform with the answers expected from probability theory.

However, the allegedly "correct" answers assume that the inferences are to be viewed as Bayesian assignments (using "uniform" prior probability assignments), whereas frequentists would not appraise evidence in the Bayesian manner. Gerd Gigerenzer shows how to make violations of Bayesian probabilistic calculations (e.g., conjunction fallacy) *disappear* by presenting the information, not in terms of probabilities, but in terms of "natural frequencies," which he maintains is "in the same form as in the environment in which our ancestors evolved" (2000, 76). Cosmides and Tooby (1996, 1) remark that "this result adds to the growing body of literature showing that frequentist representations cause various cognitive biases to disappear, including overconfidence, the conjunction fallacy, and base-rate neglect" requiring a reexamination of the common conclusions in the literature on judgment under uncertainty "that our inductive reasoning mechanisms do not embody a calculus of probability."

But one might argue instead that both the traditional and Gigerenzer's "frequentist" experimental designs are flawed: they have no chance of detecting different models of rational inductive reasoning that involve neither Bayesian updating nor assigning probabilities to hypotheses. I consider just one well-known chestnut: the so-called conjunction fallacy.

Conjunction fallacy.—The fact that certain experiments show that subjects take the various premises given as evidence as better grounds for a conjunction *A* & *B* than for a conjunct *A* is taken to show that people commit a "conjunction fallacy." Since the probability of a conjunct exceeds that of a conjunction, this appears anomalous for received accounts of "rationality" in which evidential warrant is assumed to be a posterior probability assignment. In the familiar example the experimental subjects are given information about Linda:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

When subjects are asked "which is more likely?"

1. Linda is a bank teller
2. Linda is a bank teller and is active in the feminist movement

a substantial proportion choose answer 2. One way to avoid this apparent anomaly is to rigorously train subjects, as Gigerenzer has, in the ways of frequencies. But perhaps the subjects understand "likely" (or whatever word the experimenter uses) along the lines of the statistical notion of likely or its informal counterpart in terms of the better explanation. Then there is no anomaly in the first place. (The likelihood of *H* given data $\mathbf{x}$ is $P(\mathbf{x}; H)$, in contrast to the probability of *H* given $\mathbf{x}$, which is $P(H; \mathbf{x})$.)

To illustrate, suppose that the "data" given to the subject are anomalous for the first choice in 1, whereas the answer in 2 presents a plausible "auxiliary" that, conjoined with 1, explains or accounts for the data. Suppose, for example, that one is presented with the data $\mathbf{x}$: very low HIV viral loads are detected in recently infected patients.

Which is more likely?

1. *H*: HIV is active from the onset of infection.
2. *H* and *A*: Recently infected patients produce sufficient *T* cells so that a very low or 0 viral load is detectable.

H and *A* together (as we now know) do a good job accounting for $\mathbf{x}$, and understanding the experimental question along these lines may well explain the results. In statistical terms, the data **x** are more probable calculated under the assumption of the conjunction *H* and *A* than under *H* alone. This is simply to illustrate another type of question that belongs in an adequate methodology of experiment, under the heading, perhaps, of how to critique experiments that probe norms of "rationality."

This brings to mind a final "self-referential" issue that might arise. In the statistical appraisal of evidence and in interpreting results, expereconomists use (non-Bayesian) techniques of significance tests and confidence intervals. Since these would be "incoherent" by a strict Bayesian conception of reasoning, it seems odd to stipulate the Bayesian model as the norm of rationality in testing subjects.

10. Concluding Comment. Results from the labs of expereconomists are increasingly at the center of some of the most interesting attempts to learn about facets of human economic behavior in controlled settings. At the same time, researchers in expereconomics face skeptical challenges about the nature and justification of their work in relation to more traditional economics research. These challenges invoke a variety of philosophical issues about evidence, theory testing, statistical inference, and modeling in social science. In contrast to other areas of social science, however, philosophers of science have rarely engaged with expereconomists in studying the philosophical and methodological issues to which their work gives rise. We hope to remedy this situation.

REFERENCES

Bateman, Ian, Daniel Kahneman, Alistair Munro, Chris Starmer, and Robert Sugden (2005), "Testing Competing Models of Loss Aversion: An Adversarial Collaboration", *Journal of Public Economics* 89: 1561–1580.

Cosmides, Leda, and John Tooby (1996), "Are Humans Good Intuitive Statisticians after All? Rethinking Some Conclusions of the Literature on Judgment under Uncertainty", *Cognition* 58: 1–73.

Friedman, Daniel, and Alessandra Cassar (2004), *Economics Lab: An Intensive Course in Experimental Economics*. London: Routledge.

Gigerenzer, Gerd (2000), *Adaptive Thinking: Rationality in the Real World*. Oxford: Oxford University Press.

Guala, Francesco (2005), *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

Kahneman, Daniel, Paul Slovic, and Amos Tversky (1982), *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kuhn, Thomas (1962), *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

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

——— (2008), "How to Discount Double-Counting When It Counts: Some Clarifications", *British Journal of Philosophy of Science*, forthcoming.

Mayo, Deborah G., and David R. Cox (2006),"Frequentist Statistics as a Theory of Inductive Inference", in Javier Rojo (ed.), *Optimality: The Second Erich L. Lehmann Symposium,* Lecture Notes Monograph Series, vol. 49. Beachwood, OH: Institute of Mathematical Statistics, 77–97.

Novemsky, Nathan, and Daniel Kahneman (2005), "The Boundaries of Loss Aversion?", *Journal of Marketing Research* 42: 119–128.

Schram, Arthur (2005), "Artificiality: The Tension between Internal and External Validity in Economic Experiments", *Journal of Economic Methodology* 12 (2): 225–237.

Smith, Vernon (2002), "Method in Experiment: Rhetoric and Reality", *Experimental Economics* 5 (2): 91–110.

Starmer, Chris (1999), "Experiments in Economics: Should We Trust the Dismal Scientists in White Coats?", *Journal of Economic Methodology* 6: 1–30.

Sugden, Robert (2005), "Experiments as Exhibits and Experiments as Tests", *Journal of Economic Methodology* 12 (2): 291–302.

——— (2008), "The Changing Relationship between Theory and Experiment in Economics", *Philosophy of Science* 75 (5), in this issue.