This paper is adapted from
Roth, A.E. "Let's Keep the Con Out of Experimental Econ.: A Methodological Note" **Empirical Economics** (Special Issue on Experimental Economics), 1994, 19, 279-289.
Reprinted in J.D. Hey (editor), **Experimental Economics**, Physica Verlag, Heidelberg, 1994, 99-109.
Reproduced here with permission (15.03.99 Dr. Martina Bihn, Economics Editor, Physica-Verlag).

Lets Keep the Con out of Experimental Econ.: A Methodological Note

BY
Alvin E. Roth

When Edward Leamer (1983) wrote the well known critique of econometric practice whose title I have adapted and adopted, he was concerned that the credibility and utility of econometric research had suffered because of differences between the way econometric research was conducted and the way it was reported 1.). He wrote (p36- 37):

"The econometric art as it is practiced at the computer terminal involves fitting many, perhaps thousands, of statistical models. One or several that the researcher finds pleasing are selected for reporting purposes. This searching for a model is often well intentioned, but there can be no doubt that such a specification search invalidates the traditional theories of inference. The concepts of unbiasedness, consistency, efficiency, maximum-likelihood estimation, in fact, all the concepts of traditional theory, utterly lose their meaning by the time an applied researcher pulls from the bramble of computer output the one thorn of a model he likes best, the one he chooses to portray as a rose. The consuming public is hardly fooled by this chicanery."

Leamer emphasized the contrast between the problems facing an econometrician and those facing an experimenter. He used agricultural experiments as his example of experimental research, and suggested that experimental methods, both in the laboratory and in the field, might prove increasingly useful to economists.

In the intervening years, the rapid growth of laboratory experiments in economics has amply justified this confidence. While we still have much to learn about the uses of experimentation in economics, it is already clear that carefully controlled experiments permit us to draw some kinds of inferences with far more confidence than we could hope to do from any available nonexperimental data.

At the same time, experimental economists, like econometricians, need to be careful not to let the way we report experiments diverge in important dimensions from the way they are sometimes actually conducted. This has become increasingly clear in the last ten years, as there have begun to be enough experimental economists so that it is no longer unusual for experimenters with different theoretical predispositions to be investigating the same questions. Some thought is in order about how to report experiments so as to make these dialogues as efficient and informative as possible. In this regard, there are useful things to be learned from the older experimental traditions in other sciences. This

is so even though, as we gain more experience with economic experiments, we see that they do not look exactly like experiments in agriculture, or chemistry, or medicine, or even in psychology, but rather that the nascent tradition of experimental economics has some distinguishing characteristics. And indeed, some of the issues we need to pay attention to, to make sure that we report economic experiments in a way that is as informative as possible about the way we conduct them, arise from these distinctive features of economic experimentation.

My purpose here is to raise some of these issues, concerning how and in what detail we report experimental procedures, what data are reported, how data are aggregated for reporting purposes, and how they are analyzed. Some of the potential pitfalls facing experimenters are similar to those facing econometricians: when pilot experiments are used to search through alternative experimental procedures and parameters, and to decide which experimental investigations shall proceed to the reporting stage, then, if this is not fully reported, it is easy to misinterpret the significance and robustness of the reported results.

Many of these issues can be highlighted by focusing on a deceptively simple matter, namely the divergent practices among experimental economists about what unit of data is called "an experiment." While this divergence is sometimes merely a matter of terminology, it sometimes masks differences in reporting practices that can lead to ambiguity about what is being reported.

What is "an experiment?"

I will focus on two approaches to reporting laboratory experiments that have emerged in the economics literature. They do not exhaust the range of approaches, indeed many reported experiments fall between them. But I think the difference between these two approaches, and their different strengths and limitations, have not been as widely appreciated as they need to be, and that this may contribute to some ongoing controversies in experimental economics.

The first approach, which I will call the method of **planned experimental design**, is the approach that has received the most formal attention in other experimental sciences. In it, investigators fill in the cells of an experimental design. Each **cell** of the experiment consists of **trials** conducted with some fixed set of experimental conditions, and the experimental design specifies which conditions will be varied, and what settings will be observed. The whole set of observations is what is referred to as the "experiment," and these observations are reported and analyzed together. There is a large literature on different kinds of experimental design, i.e. on different ways of identifying the effects of particular variables.

For example, when we speak of an experiment that uses a 3x2 factorial design, we mean one in which there are two experimental variables, one of which is observed at 3 settings and one of which is observed at 2 settings, and in which all 6 combinations of these settings are observed. (It is often impractical to employ factorial designs, which look at all combinations. Many experiments examine only some subset of the possible cells.) Each trial of the experiment consists of an observation at a particular pair of settings (and may be a complicated event involving many transactions by many subjects), and each cell of the experiment consists of independent trials at the same

choice of settings The data of the experiment consist of all the trials in all the cells examined.

Under this approach, once an experiment has been designed and conducted, few if any questions of judgement arise about what data to report. Although the deletion of occasional data points from the analysis may be noted with the explanation that they were outliers, or that some breakdown of the experimental procedures took place in the trial in which they were collected, the presumption is that all the data collected are reported.

This does not mean that there are not ways in which judgement may have played a role that needs to be communicated if the experiment is to be properly understood. While it is obviously a matter of judgement what experiments to conduct, i.e. which conditions to vary and what parameters to set, what is less obvious because it is often unreported is how these decisions might have been influenced by preliminary "pilot" experiments that may have been run. This is where the question of unreported search may arise. I will return to this subject, and why I think there is room for improvement in the way we report these matters, after introducing the second approach to reporting experiments that has become common in the literature.

In what I will call the method of **independent trials**, each trial is itself regarded as an experiment. The inclination to do this in economics experiments arises from the fact that each trial may be a complex event consisting of multiple decisions and observations 2. But when each trial is regarded as a separate experiment, the potential for problems associated with unreported matters of judgement is magnified, because even the question of what trials to report, or to report together in the same paper, may be taken to be a matter of judgement 3. This kind of problem is further magnified in an active laboratory which may have related investigations proceeding simultaneously, since then the problem of sorting which trials to present as evidence bearing on a particular question may be decided only when the time comes to write a formal report. And the distinction between which trials are pilot experiments and which trials are "actual" experiments may become entirely arbitrary.

The potential for trouble in treating each trial as an independent experiment depends on the kind of conclusions the experimenter seeks to draw from the data. One relatively **un**troublesome use of the method of independent trials in experimental economics is in investigations intended to show that some theory is not a good predictor for every situation to which it might be thought to apply, i.e. investigations intended to find counterexamples.

A good example is the famous "paradox" of Allais (1953). To establish that there are some risky choices for which expected utility theory is not a good predictor, Allais reported two pairs of lotteries for which a substantial percentage of subjects, when faced with the task of choosing one lottery from each pair, made choices inconsistent with expected utility theory. This is a reliable result: you can present a pair of lotteries like Allais' to your class in the confident expectation that many of your students will make the choices that Allais observed. Subsequent experimenters have considerably expanded the kinds of lotteries about which this can be said, and they customarily report these as independent trials. That is, it is customary to report tasks for which subjects consistently

violate expected utility theory without reporting the entire set of tasks that the experimenter may have examined.

In this case it is not clear that much information is lost by this practice. Suppose the task reported by some investigator, which shows subjects violating expected utility theory, were the result of a search in which he presented ten different tasks to subjects, and that the responses to the other nine tasks were largely consistent with utility theory. We cannot conclude from this anything about the percentage of cases in which expected utility theory will be unreliable: another experimenter, with less insight, might have had to search through a hundred choice tasks before finding a good counterexample, and yet another investigator might have found the same example on his first try.

A more problematic use of the method of independent trials involves the search for examples that illustrate a theory. Such examples are sometimes thought of as constructive proofs that the theory is not behaviorally vacuous, i.e. as demonstrations that the theory applies at least in some observable situations. Suppose an investigator is interested in finding a game in which a certain kind of equilibrium outcome can be observed, and that, using the method of independent trials he examines ten games, identifies one in which the equilibrium is regularly observed, and reports only that one. (Since he regards each trial as an independent experiment, he regards those trials involving the other nine games as unrelated, failed experiments 4.) Even if the investigator has no intention of implying that this kind of equilibrium will be reached in all games, some information is lost when the nine games in which non-equilibrium outcomes were observed go unreported. In particular, there might be some common process that accounts for the outcomes of all ten games, and which coincides with the equilibrium outcome of the one reported game merely by accident. Including a brief account of the other nine games could give clues to other investigators, while a failure to report them could easily lead astray even investigators seeking to follow up on the reported experiment with new experiments, if they stick too closely to games with the structure of the single reported game. My point is that, while there can be very good reasons to carefully select experimental tasks and conditions through search or other means, the manner in which this selection is carried out is a reportable part of the experiment.

The most troublesome use of the method of independent trials is when the investigator interprets his data as supporting the general predictions of some theory, such as a theory which says that all games will result in a certain kind of equilibrium outcome. If trials in which the theory fails are not reported together with those in which it succeeds, an entirely erroneous impression about the success of the theory can be given 5..

What can we learn from the experimental traditions in other sciences?

The first thing that becomes apparent from the experimental traditions of other sciences is that there seems to be no foolproof way to set guidelines for what constitutes "all the relevant data" from an experimental investigation. When the National Academy of Science's Committee on the Conduct of Science attempted to describe how data should be treated, they noted this difficulty with the following cautiously worded story (1989, pp2-3):

"One well-known example of this difficulty involves the physicist Robert Millikan, who won the Nobel Prize in 1923 for his work on the charge of the electron. In the 1910s, just as most

physicists were coming to accept the existence of the electron, Millikan carried on a protracted and sometimes heated dispute with the Viennese physicist Felix Ehrenhaft over the magnitude of the smallest electrical charge found in nature... Ehrenhaft used all the observations he made, without much discrimination and eventually concluded that there was no lower limit to the size of an electrical charge that could exist in nature. Millikan used only what he regarded as his 'best' data sets to establish the magnitude of the charge and argue against the existence of Ehrenhaft's 'subelectrons.' In other words, Millikan applied methods of data selection to his observations that enabled him to demonstrate the unitary charge of the electron.

"Millikan has been criticized for not disclosing which data he omitted or why he omitted those data. But an examination of his notebooks reveals that Millikan felt he knew just how far he could trust his raw data. He often jotted down in his notebooks what he thought were good reasons for excluding data. However, he glossed over these exclusions in some of his published papers, and by present standards this is not acceptable."

It is of course difficult to judge particular cases, especially in hindsight once the phenomena in question seem well understood. Elsewhere in the same document (p14) the Committee on the Conduct of Science characterize as fraud the deliberate practice of "selecting only those data that support a hypothesis and concealing the rest ('cooking' data)." 6. What is clear is that the larger the role that unreported "methods of data selection" play in determining what data are presented, the harder it is for readers to reliably interpret the data.

Another theme that stands out in the experimental literature is that experimenters must constantly guard against self deception: particularly when an investigator has clear intuitions about what should happen, it is easy to read these into the data 7. Many of the specialized experimental methods commonly employed in some sciences (e.g. double blind trials) are addressed directly at such problems. And much of the general aim of experimental methodology is to reduce the scope for the subjective expectations of the investigators to play a critical role.

In summary, one method of data selection that is clearly beyond the pale is to select only data that conform to the predictions of the hypothesis being (nominally) tested. But this is not a simple thing to guard against when procedures are used in which experimenters' subjective expectations may play a large role. Many experimental methods, including the use of planned experimental designs, are intended to guard against inadvertently selecting data in this way.

The potential for mis-communication due to different theoretical predispositions

A subsequent committee of the National Academy of Sciences, the Panel on Scientific Responsibility and the Conduct of Research, quotes the historian Jan Sapp as follows

"What 'liberties' scientists are allowed in selecting positive data and omitting conflicting or 'messy' data from their reports is not defined by any timeless method. It is a matter of negotiation. It is learned, acquired socially; scientists make judgments about what fellow

scientists might expect in order to be convincing. What counts as good evidence may be more or less well-defined after a new discipline or specialty is formed; however, at revolutionary stages in science, when new theories and techniques are being put forward, when standards have yet to be negotiated, scientists are less certain as to what others may require of them to be deemed competent and convincing."(Sapp, 1990, p113 as quoted in NAS, 1992, p39.)

Experimental economics is certainly a relatively new technique in economic research (despite tracing its origins back over sixty years, see Roth, 1993 or Roth, 1995), and the rapidly growing acceptance of experimental evidence undoubtedly constitutes something of a revolution, so in light of the above quotation it should not surprise us to find that the views I have outlined here do not meet with complete agreement (recall footnote 5).

The essence of the position I have taken is that negative results, i.e. results contrary to the experimenter's expectations or to received theory, are potentially very informative, and ought to be reported even if the experimenter decides to pursue results more in accord with his expectations or with standard theory. But what one investigator may view as informative results contrary to the predictions of received theory may be seen by another investigator as a failed experiment. For example, I have seen it argued that an experiment to detect a Nash equilibrium is like an experiment to detect a subatomic particle, and that once it has been reliably detected in one experiment, experiments which fail to detect it can be regarded as failures 8.

Now I certainly don't want to discourage scientists who believe that their experimental apparatus is defective from tinkering with it until it works better. (In an economic experiment, this might involve changing the instructions to the subjects, the mechanics of the experiment, some parameters, or even the game or market being investigated.) My point is that there is room for substantial miscommunication if the search for conditions that yield the desired result is not reported in the resulting papers, which instead present as evidence only those trials that "succeeded," and do not mention the nature of the "failures" and the search for conditions which avoided them.

Concluding remarks

For a variety of reasons, experimental economists have often found that the phenomena they wish to explore, and the hypotheses they wish to test, require experiments in which each trial is a complex event, which may involve many separate transactions. And because of the technology of economics experiments, trials tend to be conducted sequentially rather than (as in agriculture experiments) simultaneously. Perhaps for these reasons, there has been a tendency among some experimental economists to regard each trial as a separate experiment. This practice greatly increases the potential for data selection to play a primary role in the analysis and interpretation of experimental results, and consequently for investigators' subjective expectations to influence the data they report. And when the procedures and parameters for the experimental trials ultimately selected for reporting are influenced by unreported pilot experiments, the process may come to bear more than a passing resemblance to the kind of search about which Leamer (1983) alerted econometricians.

This latter comment can also apply to experiments reported according to a planned experimental design. And under either method, the questions of what experiments to conduct, and which experiments to report together and which separately (and which not at all) are questions of art, as opposed to matters of clearly defined practice. The difference between the two methods, then, has to do with the fact that the method of independent trials, by considering each trial to be an experiment, makes **all** the decisions about data presentation questions of art, while the method of planned design restricts this artistic freedom by requiring the experimenter to first report the design (i.e. all the observations to be made in the experiment), and then to report all of these observations, and to report them together. Insofar as all experimenters need to take care not to let their prior expectations play too great a role in determining which parts of the data they take seriously, experimenters who regard each trial as a separate experiment may want to reconsider whether this is the best approach for the questions they wish to address. And to avoid unnecessary confusion, experimenters who are reporting all the data in their design should probably avoid calling each trial an experiment 9.

There is room for us all to do a better job reporting what kind of pilot experiments we have conducted, and how they may have influenced the design choices made in the experiments from which the reported data were gathered. It may not always be possible or desirable to conduct pre-planned experimental designs, as the results in an early cell of an experiment may call for a change in plans. But if the process by which the data are collected is reported, the potential for miscommunication can be reduced.

As in econometric research, we may draw different inferences about the robustness of results that result from a search 10. In this connection, we should also, as editors and referees, give some thought to making sure that we do not create incentives that encourage poor reporting practices. This is not a simple thing, for it means both showing a tolerance for ambiguity of results in well designed experiments, and being prepared to publish replications of various sorts, particularly those aimed at investigating the robustness of conclusions 11. 12 But the experience of other fields suggests that there is a limit to what can be accomplished by exhortation of authors or editors. As a profession, our best defense against erroneous conclusions resulting from unreported or incompletely reported search is to encourage experimenters to follow up on one another's work, with experiments which change elements of the design that might influence the outcome in ways not accounted for by the theory or theories proposed to account for the results.

Indeed, one of the reasons that experimental economics has been as productive as it has is that reported results have by and large been straightforwardly replicable: when a carefully conducted experiment is repeated, the likelihood that the data will be similar seems to be high. But precise replication gives little information about robustness. What ultimately gives us our best indication of the robustness of experimental results is replication with some variation of experimental parameters and conditions. And, particularly in the last ten years as experimental economists have become numerous enough so that investigators with different theoretical predispositions have started to examine the same questions, we have seen that investigators with different intuitions about a question can sometimes design experiments that lead them to different conclusions. In this respect, some of the controversies in the experimental literature focus on questions of robustness, and will be more productively carried out when

methodological issues concerning how much search has been conducted and how much data selection has been employed are clearly reported.

In conclusion, one of the principal roles of experimental methodology is to help investigators avoid the danger of too easily accepting their prior hypotheses. The experience of other sciences suggests that this is a problem that will be always with us: evidence counter to the preferred hypothesis may look like experimental error ("the subjects didn't understand"), there is a natural tendency to keep trying different parameters when things aren't working out, but to stick with parameters which do work out, etc. As in econometric research, experimental results which are the end product of a search for conditions and parameters that will yield certain kinds of results are difficult to interpret without some understanding of the search process. And all of these problems are exacerbated when investigators regard each trial as an independent experiment.

But overall, the experimental enterprise contains many elements that promote the identification of robust results. Chief among these is that experimenters don't have to rely on one anothers' data, or even their choices of parameters and procedures, but can generate their own data from experimental environments well suited to testing their hypotheses precisely. And so **series** of experiments allow the experimental community to build upon and critique one anothers' work in ways that are not as readily available to economists using non-experimental methods. There is every reason for optimism about the contribution that series of experiments, particularly when conducted by experimenters with different points of view who take care to address each others' positions, can make to identifying robust empirical regularities in economics 13. The purpose of this note is to help to enhance the efficiency of this kind of dialogue, by raising some of the methodological issues that both producers and consumers of experimental research must consider in interpreting experimental results.

REFERENCES

Allais, Maurice [1953] "Le Comportement de L'homme Rationnel Devant le Risque: Critique des Postulats et Axiomes de L'ecole Americane," Econometrica, 21, 503-546.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson [1986], "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project," American Economic Review, 76, 587-603.

Feige, Edgar L. [1975], "The Consequences of Journal Editorial Policies and a Suggestion for Revision," Journal of Political Economy, 83, 1291-1295.

Kagel, John H. and Alvin E. Roth (editors) Handbook of Experimental Economics, Princeton University Press, 1995.

Leamer, Edward E. [1983], "Let's Take the Con out of Econometrics," American Economic Review, 73, 31-43.

National Academy of Sciences [1989] On Being a Scientist, Committee on the Conduct of Science, Washington, National Academy Press.

National Academy of Sciences [1992], Responsible Science: Ensuring the Integrity of the Research Process, Volume I, Panel on Scientific Responsibility and the Conduct of Research, Washington, National Academy Press.

Prasnikar, Vesna and Alvin E. Roth [1992], "Considerations of Fairness and Strategy: Experimental Data From Sequential Games," Quarterly Journal of Economics, August, 865-888.

Rosenthal, Robert and Kermit L. Fode [1963], "The Effect of Experimenter Bias on the Performance of the Albino Rat," Behavioral Science, 8, 183-189.

Roth, Alvin E. [1993]

- "On the Early History of Experimental Economics," Journal of the History of Economic Thought, 15, Fall, 1993, 184-209.

Roth, Alvin E. [1995], "Introduction to Experimental Economics," Chapter 1 of Handbook of Experimental Economics, John Kagel and Alvin E. Roth, editors, Princeton University Press, 3-109.

Roth, Alvin E. and Ido Erev [1995] "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," (presented at the Nobel Symposium on Game Theory, June 18-20, 1993, Bjorkborn, Sweden) **Games and Economic Behavior**, Special Issue: Nobel Symposium, vol. 8, January 1995, 164-212..

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir [1991], "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," American Economic Review, vol. 81, December, 1068-1095.

Roth, Alvin E. and Xiaolin Xing "Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions," American Economic Review, 84, September, 1994, 992-1044.

Sapp, Jan [1990], Where the Truth Lies: Franz Moewus and the Origins of Molecular Biology, Cambridge, Cambridge University Press.

ENDNOTES

[Click on note numbers to return to the text at the note.]

1.See also the earlier related critique by Feige [1975], who focused on the incentives for questionable reporting practices inadvertently encouraged by the editorial policies of professional journals, a point to which I shall return in the conclusion. Both Feige and Leamer refrained from citing by name any particularly egregious examples in the

literature of the practices they criticized. I will follow their example, in order not to transform into personal criticism what I intend to be remarks of potential relevance to all experimenters.

2.For example, allowing subjects to participate in many repetitions of a given market may be desirable, to allow them to gain experience of market parameters and of the behavior of others.

3.Some additional confusion is caused by the fact that even some experimenters who conduct and report planned experimental designs have started to follow the practice of calling each trial an experiment. But this confusion can be fixed just by a change in terminology, since these experimenters are simply reporting all the data in their design.

4.But note that there are circumstances in which even an investigator intending to report a planned experimental design might come to regard the other games as pilot experiments: suppose each game is initially examined with just a few subjects, and only in five of the games are some equilibrium outcomes observed. These five games are then examined with a few more subjects, and only in one of them does the equilibrium outcome occur with really high frequency. Finally that one game is examined with more subjects and reported, with some parametric variations, as part of a small experimental design focused on the one game. The investigator regards the other games as failed pilot experiments, and doesn't report them.

5.In this connection I once had the opportunity to hear one experimental economist chide another for having reported that a certain kind of market did not always yield equilibrium behavior. He felt that perhaps a premature negative result had been reported. He went on to say that, in his own research, when he found in an experiment that some economic institution "didn't work," he first tried rewriting the instructions to make sure that they hadn't contributed to the negative result, and if that didn't fix the problem he would try changing the mechanics of the experiment. Often, he said, that fixed the problem. Left unstated was that this search for conditions that would yield the desired result was not reported in the papers that resulted from this activity, which simply presented, as if they were independent experiments, trials that had "worked."

6.In medical research, where each trial may involve a separate laboratory animal, for example, I have heard the practice of treating each trial as an independent experiment and reporting only those that support the hypothesis referred to as "forgetting about the mouse that died."

7.There have even been experiments aimed at elucidating this effect. For example, Rosenthal and Fode [1963] show in a controlled experiment that experimenters who were told that a given rat was experienced or inexperienced at running a maze reported results that reflected their expectations.

8.For the record, I do not agree with the hypothesis that if one game is observed to reach equilibrium it must therefore follow that other games, even similar games, do also, nor do I agree that disequilibrium observations imply an error in the experiment or the analysis. In fact I have reported experiments designed to explore why some games move quickly to their subgame perfect equilibrium while others do not (see Prasnikar and Roth, 1992, Roth, Prasnikar, Okuno-Fujiwara, and Zamir, 1991, and see Roth and Erev,

1995 for a theoretical treatment.) However it is not my intention to discuss the merits of these hypotheses here.

9.It is a simple matter to speak of trials, or experimental "sessions", rather than "experiments" in such a case.

10.To further the analogy with econometrics, the standard error of a (single) regression which is selected from a search understates the ambiguity of the data, and may be a biased estimator. In just such a way, experimental evidence that results from a search through conditions understates the ambiguity of the evidence, and may misrepresent the data, if it omits mention of the search. A related matter is that, because trials are often costly, and because they may involve multiple decisions by subjects who interact with one another in complicated ways, even planned experimental designs for economic experiments often have relatively few cells, few trials per cell, and few truly independent observations per trial. There is room for improved econometric techniques for the analysis of such data. And, here too, replication of results plays a vital role. (Replication in experimental economics means a very different thing than it does in econometrics, where it sometimes refers to the ability of other investigators simply to reproduce the same analyses from the same data--see Dewald, Thursby, and Anderson, 1986. Experimenters seek to replicate a result by generating new data under the same or comparable conditions.)

11.Writing about the applied econometrics literature, Feige (1975), in the paper which caused the **Journal of Political Economy** to start accepting papers in the category "Confirmations and Contradictions," wrote:

"...current journal editorial policies have undoubtedly contributed to (1) an incentive to pursue search procedures for statistically significant results which are spurious as often reported, insofar as they take no account of pretest bias; (2) an incentive for less than candid reporting of intermediate results which could highlight the lack of robustness of statistical tests to alternative model specifications and applications of alternative econometric techniques; (3) an underrepresentation of "negative" results which could otherwise signal empirically anomalous results leading to the rejection of currently maintained hypotheses; and (4) an unnoticed proliferation of published Type 1 errors." (pp1292-93).

12.Counterproductive incentives about what to report may also exist in areas of economics involving other than econometric or experimental data. I have only an anecdote to offer about economic history, having to do with a paper of mine (Roth and Xing, 1994) which contains historical descriptions of the evolution of the timing of transactions in several dozen markets, along with some theoretical models of the observed phenomena. One reviewer suggested that the historical data exhibited an excess variety of behavior compared to that which could be accounted for by the formal theoretical models. Obviously one way to avoid such a criticism would be to report only those historical observations that can be explained by the available theory.

13.For many fine examples of the success of this approach, see the **Handbook of Experimental Economics** (Kagel and Roth, 1995).