Real Phenomena, Theory and Design of Laboratory Experiments in Economics

Shyam Sunder
Yale University
(Preliminary Draft, June 26, 2001)

What is the relationship between theoretical models and laboratory experiments? What is the relationship between these two and the real phenomena of interest? How should we design laboratory experiments to take these relationships into account? These are some of the most frequently asked questions in experimental economics. This note continues development of a perspective on these questions from Friedman and Sunder (1994, pp. 1-20) and Sunder (1995, p. 491-3).

I assume that the purpose of building theory as well as design and conduct of laboratory experiments is to gain a better understanding of real (i.e., naturally-occurring) phenomena of interest. Real phenomena are complex, and it is rarely possible to understand them completely. Theory identifies one or a few critical variables that may help us gain a satisfactory, though rarely complete, understanding of the phenomena. Instead of thinking of a theory as being right or wrong, it is better to assess a theory by how useful it is for understanding the real phenomena of interest. When there are two or more theories of the same phenomena, we compare them on the basis of their usefulness in understanding the real phenomena.[1]

The essence of theory is simplification and abstraction from details of the real phenomena. While the explanatory power of theory is important, simplicity is at least equally so. Theory simplifies and abstracts away from unnecessary details of reality by making assumptions. Some are key assumptions, while others are merely for convenience. A lack of correspondence between assumptions of convenience and reality is the essence, not a defect, of theory.

With its highly simplified assumptions, a theory is to real phenomena what a stick figure is to a human body. Its correspondence to reality is crude, at best. The theorist develops statements that are tautologically true whenever the assumptions hold. Given tautological nature of theory, and assuming that there are no errors in derivation, what does it mean to empirical "test" the theory in laboratory or field? We consider this question under three conditions.

1. Single Theory Experiments

When there is only one candidate theory for the phenomenon of interest, empirical "test" of a theory consists of assessing its robustness to deviations from its assumptions of convenience. If the data are gathered from an environment that

[1]In response to doubts about the usefulness of experimental environments as a means to compare theories and delimit the extent of their generalization, Plott (???) has argued that laboratory economies are also "real" economies in the sense that they represent observation of real behavior and outcomes under controlled circumstances. After all, few models in economics claim to inapplicable to such controlled environments. The point is well made, and I concur, for the purpose. However, I use "real" in these sense of naturally occurring phenomena whose understanding is the principal driver behind efforts to build theories, and testing them from data gathered in the field or laboratory. For this second purpose, understanding of the naturally occurring economies is the beacon to guide the design of laboratory economies. In this sense, we call the former real, and the latter their model. If we regard both as real for this purpose, laboratory economies risk becoming self-referential.

corresponds exactly to the assumptions of the theory, one should not expect any difference between its predictions and the data. If a difference is observed, it must be attributed either to a gap between the data and model environments, or to a logical error in derivation of the tautologies. Empirical observation is a rather costly way to discover such errors.

It is not easy to find in the field, or create in the laboratory, conditions that correspond exactly to the assumptions of the theory. Even if we did, there is little we could learn from such observations except the presence or absence of derivation errors. Almost always, we attribute deviations of data from theoretical predictions to a lack of perfect correspondence between the assumptions of theory and the environment from which the data are gathered. In a strict sense, the data reject the theory, and the theory rejects the environment from which data are gathered as defective. What useful scientific inferences can be made from such a mutual lack of correspondence?

Scientific value of empirical "test" of a single theory lies in assessment of how robustly the predictions of the theory correspond to data as the environments from which the data are gathered become less similar to the convenience assumptions of the theory. In Figure 1, we represent the distance, on some appropriately defined metric, between the data environment and the assumptions of the theory on x-axis. The y-axis measures the correspondence, predictive or explanatory power of the theory with respect to the data.

The figure shows three empirical relationships, marked A, B and C respectively, schematically. The intersection of the three curves at top left corner on y-axis marks the unlikely event when the two environments correspond one hundred percent, as do the data and the model predictions.

Line *A* marks an empirical relationship in which the predictive power of the model drops off sharply with even small deviations of laboratory environment from the model assumptions. This relationship suggests that while the model is literally true—it is a tautology after all—it is not robust to small deviations. Since the purpose of a model is to help us understand the real phenomena, the sharp drop-off in the predictive power of the model suggests that the model is unlikely to be a useful guide to such phenomena, which occur in circumstances that deviate from the model in varying degrees.

Line *B*, on the other hand, marks a highly robust model. Its predictions correspond reasonably well to the data even in the presence of large deviations from the model environment. Such models, rare as they may be, are useful guides to real phenomena.

Line *C* marks an intermediate case between *A* and *B*. The predictive power of the model remains high for moderate deviations in environment before declining to lower levels.

The shape of curves *A*, *B*, and *C* may be sensitive to how we define the distance between the model and data environments plotted on the x-axis. There may be no unique metric of these differences. Friedman and Sunder (1994) suggest isolating the key assumptions of the model from the assumptions of convenience. For example, while monotonically increasing preferences for dividends may be a key assumption in a model of security markets, the number of states of the world may be only an assumption of convenience. In testing the robustness of a single theory using laboratory experiments, we can conduct multiple experiments, all holding the key assumptions, but progressively relaxing the assumptions of convenience. This could be done, in our example, by using

the number of states of the world as a metric on x-axis, the number of states in the model being used as the zero of the x-axis.

A second approach to testing the robustness of the model is to vary the number of alternative choices available to subjects, and therefore the number of possible outcomes of the experiment. While the model may predict one choice and outcome, the experimental design must permit at least two, in order for the theory to be falsifiable. The greater the number of viable alternatives available to the subjects, and greater the number of viable outcomes of the experiment, more robust is the model if it's prediction is supported by the data.

An Example of Single-Theory Experiment

Vernon Smith's double auction experiment (1961, Figure 2) is, perhaps the best-known example of testing the robustness of a single theory in both the dimensions we have mentioned above. The theory of demand and supply was the only model considered. Its predictions are derived from a large number of assumptions including the auction form (Walrasian), atomistic competition among a large number of traders. Smith conducted an experiment in which the orderly Walrasian auction was replaced by a chaotic multilateral bargaining, and the infinity of traders were replaced by a mere handful. We can mark the two auction forms on x-axis of Figure 1 (Walrasian at origin), or plot the number of traders on the x-axis, using origin for the number theoretically deemed necessary for atomistic competition.

Smith's experiment showed us that the huge difference between the theoretical and experimental environments in both auction form and the number of traders made little difference to the predictive power of the law of supply and demand (see Figure 2). Furthermore, the experimental design allowed the participants a choice from a very large space of possible actions, and therefore allowed a very large number of possible outcomes. From all of these possibilities, the data corresponded to a few points in the close neighborhood of the price, allocation, and efficiency predictions of the demand and supply model.

Thus there were at least two reasons why so many readers found Smith's results so compelling in support of the demand and supply model, even though there was no alternative model available. Since the predictions of the model held for an auction form radically different from the one for which the model had been developed, it suggested that the applicability of the model extended well beyond the auction institutions used in either the model or the experiment. This proved to be true when subsequent experiments were conducted with scores of other auction forms. Before Smith's experiment, the Walrasian auction form may have been seen as a key assumption of the demand and supply equilibrium model.

Similarly, for those who may have regarded atomistic assumption of demand and supply equilibrium as a key assumption of the model, the results of the experiment came as a surprise. Even the most fervent believers of the model hardly expected it to do so well with less than ten traders.

If Smith's experiment had achieved all this by offering the traders only two choices—trade at either the theoretical equilibrium price or at, say, one-half of the equilibrium price—the readers may have felt that the equilibrium result, if observed, had been forced. Instead, in Smith's experiment, the subjects could choose their actions to

bid or ask freely, and choose the prices at which they could bid and ask from a set of over one hundred integers.

After this experiment, two key assumptions, auction form and atomistic competition, were downgraded to the status of an assumption of convenience. Since weaker assumptions make the theory stronger, the experiment generalized and thus considerably strengthened the theory. The correspondence between the data and the model prediction is all the more compelling when the set of possibilities from which the data are chosen by the subject's actions so large.

The value of experimental testing of single theories lies in their power of generalization. Experiments help rescue the models from the knife-edge confines of the assumptions necessary to derive their formal results, and show that their predictions may be useful approximations over a set of real phenomena with positive measure. Returning to the Smith example, environments that meet the assumptions for which the perfect market equilibria are derived is a set of measure of zero. Smith showed that the predictions of the model hold to approximation acceptable to many over a range of real markets.

Models achieve their power from abstraction from and simplification of real phenomena. These tautologies based on precise assumptions have no empirical content. Experiments are a way of fleshing out the "stick figure" models, so we can assess whether the models provide us useful guidance in a nontrivial range centered on the model. This is the sense in which experiments, as well as field data, given empirical content to the otherwise sterile tautologies or models. Of course, without tautologies to guide experimental design, the latter becomes an endless groping in the dark.

This view of the role of experiments implies that even if it were possible, there is little reason to try to create conditions in the lab that correspond precisely to the assumptions of the model. Greater the deviation of the laboratory conditions from the model assumptions (of convenience), greater is the added generalization of the model if the it predicts well. If the predictive power drops below the acceptable level, the lab environment has discovered that it lies beyond the boundaries to which the model can be generalized. Another experiment closer to model assumptions will reveal the acceptable limits of generalization.

The single theory experiments do not reject a theory, even if they show the theory to be unsatisfactory is organizing the data. There is little advantage to rejecting a theory until an alternative is available. Poor performance in a single theory experiment may well encourage us to look for alternatives. When such alternatives become available, we are ready for two or multiple theory experiments.

1. Multi-Theory Experiments

When two or more alternative theories relevant to the same real phenomena are available, experimental design can set up a competition among them. First consider competing theories based on identical assumptions. The purpose of such a competition is, again, to assess the empirical content of the theories by identifying the range of empirical environments over which some may have an advantage over the other(s).

The key to setting up a theory competition is creating a laboratory environment in which two or more theories make distinct predictions. Figure 3, for example, shows the predictions of three models, Theory1, Theory2, and Theory3 respectively, to lie at

various points along a straight line.[2] Suppose the data are found to be clustered around point X between the predictions of Theory1 and Theory3, one may infer that the predictions of Theory1 and Theory2 dominate the predictions of Theory3 under the relevant empirical generalizations. Instead, if the data cluster around Y between Theory2 and Theory3, we could draw a similar inference about the data favoring Theory2 and Theory3 over Theory1. Data that lie at W to the left of Theory1 would favor Theory1 over Theory2 and Theory3 while data lying at Z to the right of Theory3 would favor Theory3 over Theory1 and Theory2.

Of course the strength of all such inferences depends on the statistical properties of data that we need not go into at this point. While the central tendency of the data is often useful for making such inferences, our confidence declines with increase in dispersion of the data.

In designing the experiment, we need to keep in mind that larger the distance between the predictions of the competing theories, greater the chance that the data from a given experiment will help us identify a winner with some confidence. The distance between the theory predictions can be measured in terms of either the expected monetary payoffs associated with the respective outcomes, or the magnitude of dispersion of data.

For example, if outcomes Theory1, Theory2 and Theory3 yield expected payoffs of $1, $2, and $4 respectively, the first measure of distance suggests that the predictions of Theory1 and Theory2 may be twice as hard to distinguish as the predictions of Theory2 and Theory3. Using expected monetary payoffs also makes the distance invariant to conversion ratios between experimental points and real money payoffs.

When the distance between theories is assessed by expected payoffs, there is an additional design trade-off in uncertain environments. Other things being the same, a larger difference between the predictions of two theories is usually is accompanied by a smaller probability of the difference being observed.

Second, suppose that the predictions of the three theories are prices of 100, 200 and 250 respectively, and the standard deviation of prices is 25. Then the distance between Theory1 and Theory2 is two standard deviations while the distance between Theory2 and Theory3 is four standard deviations, making it more difficult to differentiate the former two on the basis of data. Thus our ability of differentiate among theories in multi-theory experiments depends both on the distance between the predictions of the theories as well as the dispersion of the data.

An Example of Multi-Theory Experiment

Plott and Sunder (1982, Figure 4) is an example of a two-theory experiment. The authors set up a competition between two relevant theories, prior information (PI) and rational expectations (RE) equilibrium to explain the outcomes of the same asset market with uncertainty and information asymmetry. Each period of the market was labeled either X or Y depending on the outcome of a random variable. They compared the experimental data to the predictions of the two theories with respect to prices, asset allocations and allocative efficiency.

[2] While I assume that the predictions of these theories lie on a ratio scale in this example, analogous arguments can be made for interval, ordinal or nominal data. I shall return to this topic later. I also assume that the differences among predictions of a theory are confined to a single dimension. This, too, can be generalized as discussed later.

The price predictions of the two theories are identical (400) under state X but differ by 45 points under state Y (220 for PI and 175 for RE). After some experience, X transaction prices cluster close to 400. Almost all Y transaction prices lie between 175 and 220. However, after some experience they tend to cluster close to the RE prediction of 175. RE wins the race as measured by predictability of price.

An inference in favor of RE is made possible by the small dispersion of the price data in later periods of the experimental session. If the dispersion of prices remained large, it would not be possible to place much confidence in such an inference.

The distance between the prices predicted by the two models is 45 points, which in itself does not have much meaning. Converted into dollars at the rate of $0.003 per point, this distance amounts to $0.135. Since the subjects were willing to trade to take advantage of price differences of a penny or two, the 13.5-cent difference between the two equilibrium prices was large enough to use the data to distinguish between the two models.

In design of this experiment, the probabilities of X and Y were set to 0.4 and 0.6 respectively and this design yielded a difference of 45 points in the prices predicted by the two models in state Y. This meant that only 60 percent of the observations, on average, can be useful in setting a competition between the two theories with respect to transaction prices (recall that both models predict the same price under state X). We could increase the probability of Y to, say, 0.65 so more of the price observations could help us distinguish between the two theories. However, the increase in the number of observations can be achieved only by changing the price difference between the predictions of two models from 45 to 30 points. Similarly, the price difference can be increased from 45 to 60 points but we must sacrifice the proportion of useful information by 0.05 to 0.55. The experimenter must balance these two considerations in choosing the parameters. For example, the expected value of absolute difference between the price predictions of the two theories in state Y is 27 under the parameters used. By changing the probability of Y from 0.6 to 0.375, this expected value could be raised to 42.175. This will yield fewer observations but a greater difference between the predictions of the two models. Whether this expected value is the appropriate design parameter to be optimized to serve the goal of picking out the better theory remains to be determined.

Figure 5 shows the results of the model competition with respect to asset allocations. The square markers plot the number of securities that were held in “wrong hands” at the end of periods (X for solid, Y for hollow), relative to the predictions of RE model. Similarly, circular markers plot the number of securities that were held in “wrong hands” at the end of periods (X for solid, Y for hollow), relative to the predictions of RE model. If the data correspond precisely to a model, this graph should show zero. Again, a comparison of hollow squares (RE-Y) and hollow circles (PI-Y) shows that the RE model is a clear winner in this race too.

Since the loss of potential surplus from trade is proportional to the number of securities in wrong hands, Figure 5 also indicates that the RE model is much better at predicting the efficiency of these markets than the PI model is.

Generalizability

As mentioned earlier, the value of empirical evidence for generalizing the usefulness of models beyond the environments where their assumptions hold arises from

the relaxation of assumptions of convenience in the laboratory. In the Plott and Sunder (1982) example, we can see this process at work.

The price and allocation predictions of the RE and PI models are based on the standard demand and supply mechanism backed by Walrasian tatonnement. The experimental design used a double auction instead. There is little resemblance between the two.

In a market for securities with uncertain payoffs, the price predictions should depend on the traders' attitudes towards risk. The experimental design made no attempt to control the risk attitudes. Subjects brought their own home-grown risk attitudes to the lab.

There also many respects in which the model is often more general than the particular lab environment of an experiment. For example, the PI model is not specific to any probability distribution but the lab environment used a binary distribution for dividends.

To be completed.

References

Friedman, Daniel and Shyam Sunder, Experimental Methods: A Primer for Economists. Chapters 1 and 2 (pp. 1-20). Cambridge, UK and New York: Cambridge University Press, ISBN No. 0-521-45068-3 (1994).

Sunder, Shyam, "Experimental Asset Markets: A Survey," Pages 191-193 of Chapter 6 in A. Roth and John Kagel, eds., Handbook of Experimental Economics, Princeton, NJ: Princeton University Press (1995).

Figure 1.

Single Theory Experiment

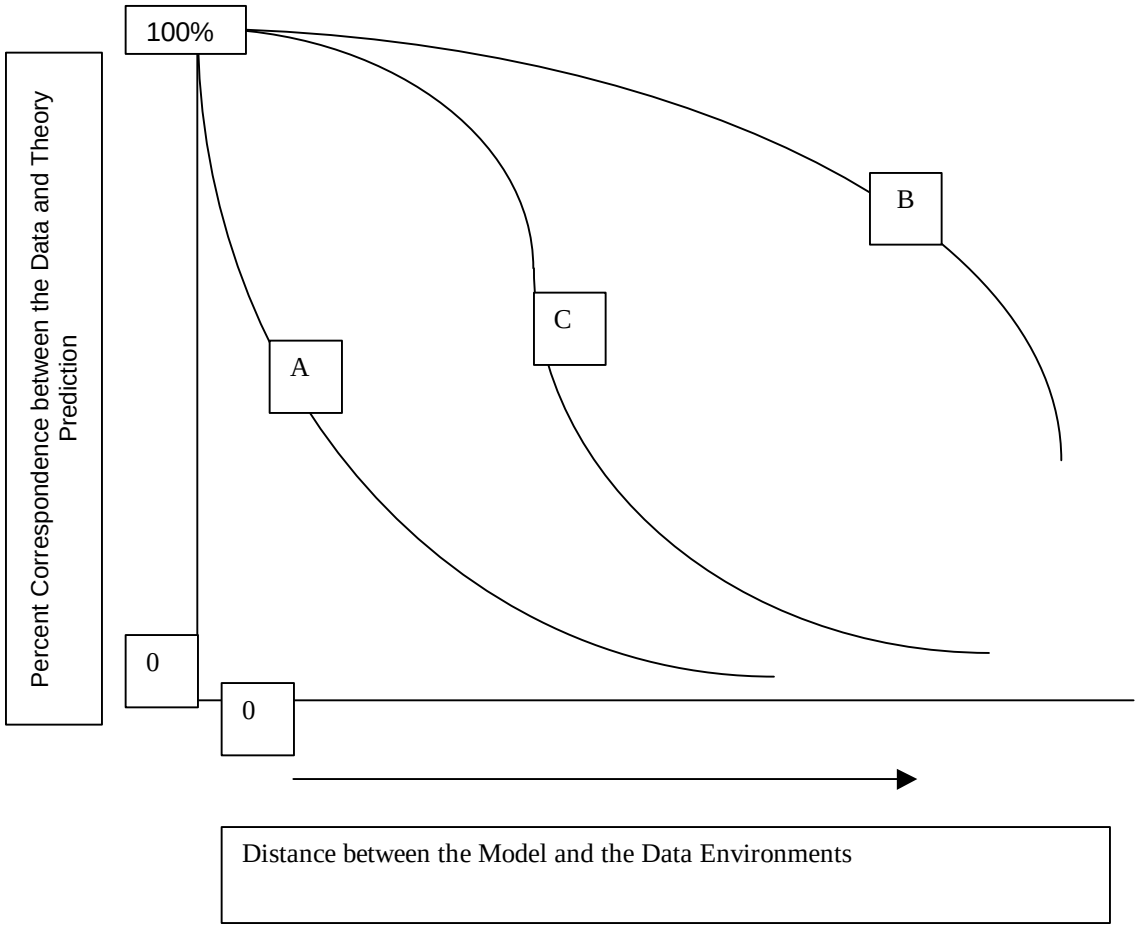


Figure 2: Multi-Theory Experiments

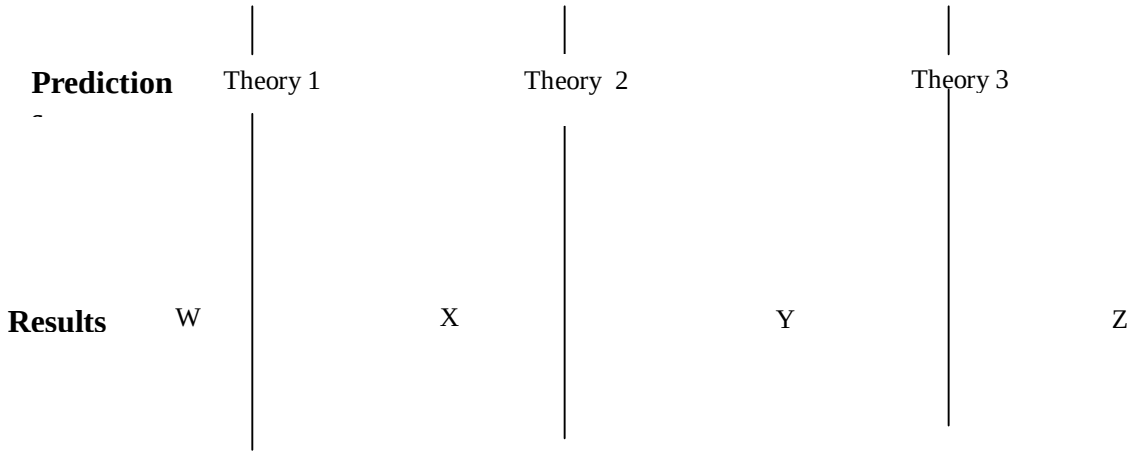
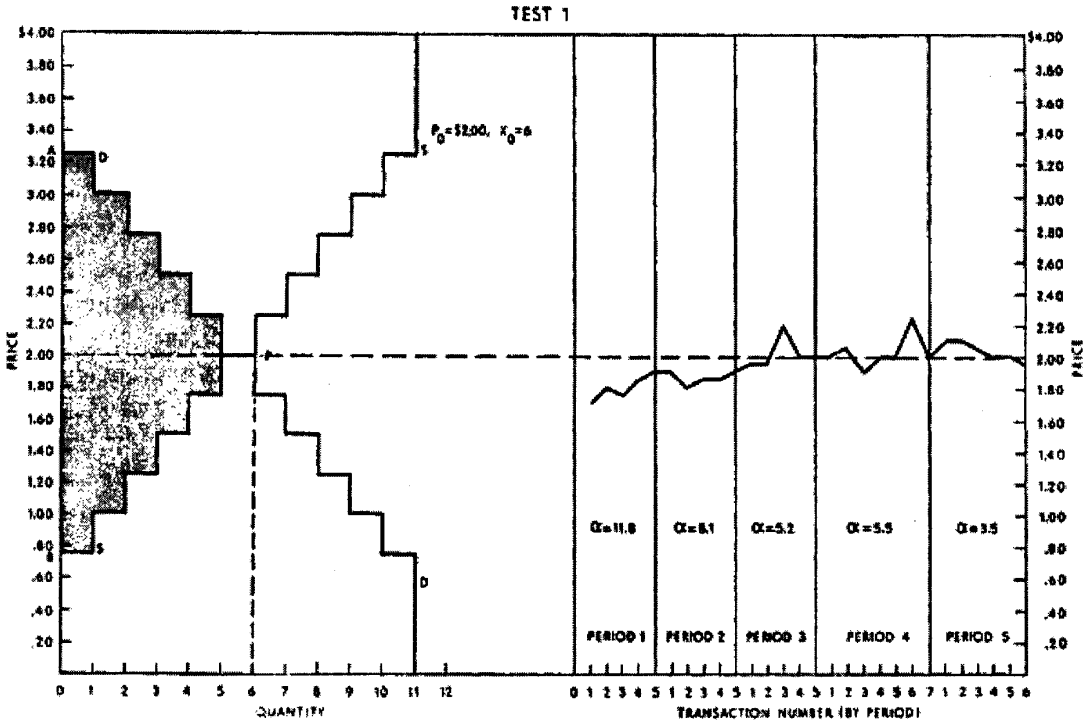


CHART 1


TEST 1
PRICE
QUANTITY
TRANSACTION NUMBER (BY PERIOD)
P_0=$2.00, X_0=6
PERIOD 1
PERIOD 2
PERIOD 3
PERIOD 4
PERIOD 5

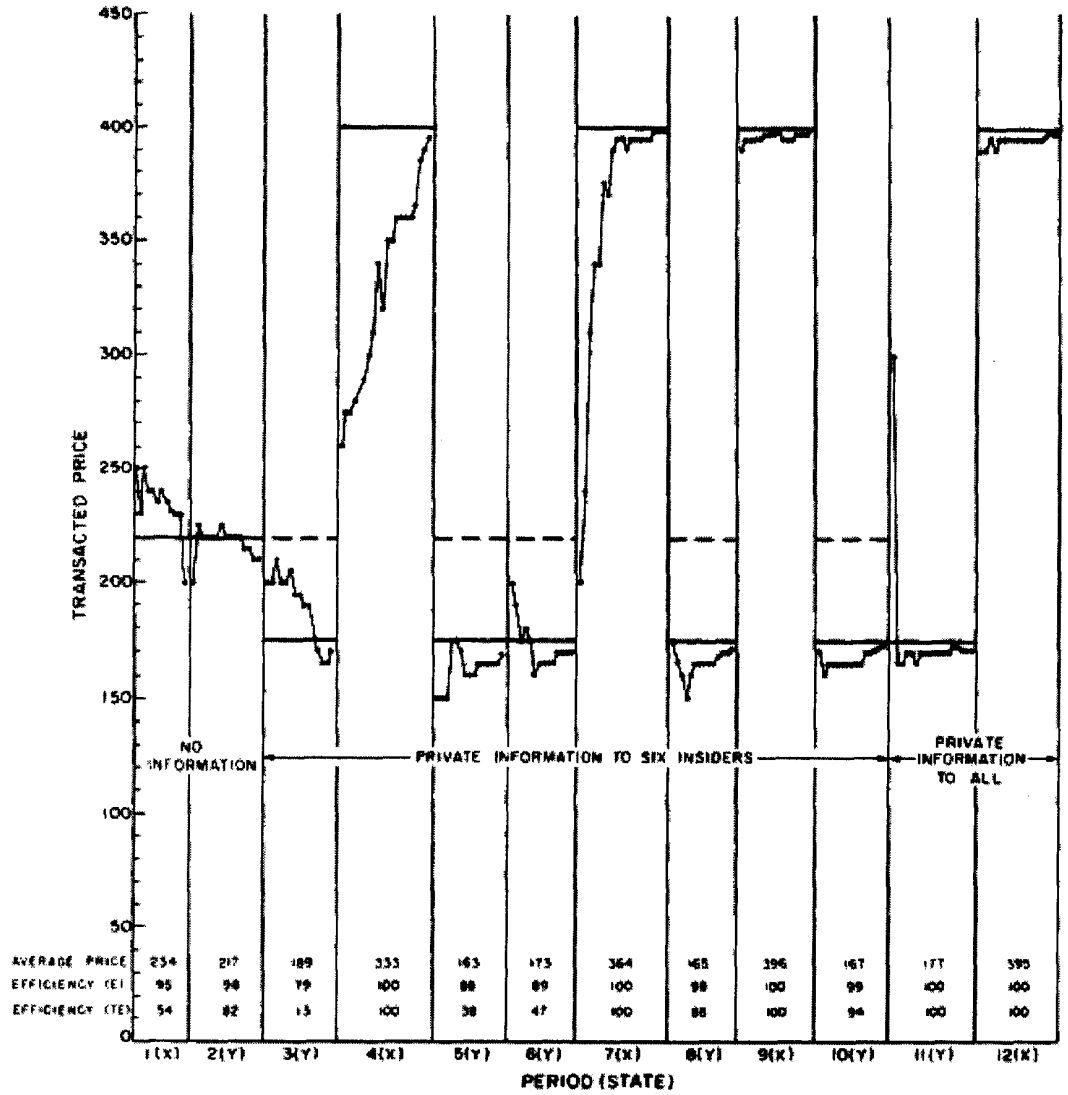


FIG. 4.—Market 3. Time series of contract prices