Paradigmatic Experiments: The Ultimatum Game from Testing to Measurement Device

Francesco Guala[†][‡]

The Ultimatum Game is one of the most successful experimental designs in the history of the social sciences. In this article I try to explain this success—what makes it a "paradigmatic experiment"—stressing in particular its versatility. Despite the intentions of its inventors, the Ultimatum Game was never a good design to test economic theory, and it is now mostly used as a heuristic tool for the observation of nonstandard preferences or as a "social thermometer" for the observation of culture-specific norms.

1. Introduction. Imagine that you have just been given $10. The sum will have to be shared with an anonymous, invisible partner, and you will have to agree with her on how to divide it. The room for discussion is almost inexistent: you will only be able to offer one division of the cake, and your partner will only be able to accept or reject it. If she rejects it, you will both lose the opportunity of sharing the $10; if she accepts, you will both walk out with your share, as determined by the proposed division.

This is essentially the strategic situation known as the Ultimatum Game (UG). Most likely, you have encountered the UG before. If, as Robert Axelrod once remarked, the Prisoner's Dilemma game has become the *E. coli* of the social sciences, then the UG must be its *Drosophila melanogaster*. Axelrod's pun should be taken seriously. It suggests that experimental game theory has reached a level of maturity similar to that of firmly established experimental sciences like biology. It also suggests that this maturity is most clearly displayed by the emergence and consolidation

[†] To contact the author, please write to: Department of Sociology and Philosophy, University of Exeter, Amory Building, Exeter EX4 4DT, UK; e-mail: f.guala@ex.ac.uk.

[‡] Marcel Boumans, Margaret Schabas, Bob Sugden, and various members of the audience at Philosophy of Science Association 2006 meeting offered very useful comments and suggestions. All remaining mistakes are mine.

Philosophy of Science, 75 (December 2008) pp. 658–669. 0031-8248/2008/7505-0014$10.00

of experimental prototypes or “paradigmatic experiments” similar to the “model organisms” of experimental biology.

Paradigmatic experiments have several distinctive epistemic features. First, years of experimenting ensure that we know them very well, indeed better than any other system in the domain of science. Second, they are used to make a variety of inferences to other systems that are less readily available for experimental investigation. Finally, they are usually versatile systems that can be used for different purposes at different times and in different contexts of scientific inquiry. In this paper, I will focus mainly on this latter characteristic, although the other ones will also feature in the story of the UG.

2. Theory Testing. The UG was first designed and run by a group of German experimental economists led by Werner Güth in the early 1980s (Güth, Schmittberger, and Schwarze 1982). Güth and colleagues intended to investigate sequential bargaining, and they chose the UG mainly because it is the simplest possible sequential bargaining problem that one can conceive of.[1] Simplicity was sought in order to minimize the cognitive cost of computation. It is well known that people struggle when they have to analyze complex dynamic games, particularly when they have to engage in so-called backward induction. In the UG, no one can fail to realize that the game will be over after the Responder’s move, so the noise in the data due to misunderstandings of the decision situation should be minimal.

Güth’s team also intended to check some of the results of previous experiments. Fouraker and Siegel, two psychologists, had performed in the 1950s a series of classic studies of bargaining. Güth and colleagues were surprised to find that unfair offers were rarely rejected in Fouraker and Siegel’s experiments. They conjectured that this could be due to some details of Fouraker and Siegel’s design, and they decided to control for these by slightly changing the experimental setup (Güth et al. 1982, 370).

As a matter of fact, unfair offers are rejected in the UG. In their first study, Güth and colleagues found that on average Proposers offer 35% of the cake. There was a substantial mode at 50%, and very few rejections by Responders (about 10%). When the game was repeated a week later (“experienced players” condition), the average offer went down to 31%. The mode at 50% disappeared, with most offers lying in the range of 20%–30%. There were also more rejections with repetition (about 30%). These results have since been replicated several times, and they constitute

[1] Bargaining is customarily represented in game theory as a sharing problem, where the surplus from exchanging two goods has to be allocated among the parties. The $10 in the UG somehow stand for this surplus.

what is sometimes called the “UG anomaly.” But anomaly with respect to what?

3. Testing and Controlling Preferences. According to standard game theory, Proposers should offer close to nothing and Responders should accept. The idea is that Respondents face a seemingly trivial decision problem: either get nothing or get whatever Proposers have offered. Let us suppose that the minimum amount that can be offered is $1. One dollar is better than nothing, so Responders should accept. Proposers know this, and thus they should offer $1. Under common knowledge of the game and of rationality, the $1/$9 split is the only equilibrium of the UG.

This “standard prediction” is actually derived from a fairly complex machinery: roughly, from a theory of strategic play (the “core” of game theory) plus a set of assumptions about people’s preferences and beliefs. According to the self-interest assumption, people prefer more money to less and they do not care about others’ payoffs. The UG therefore is an anomaly with respect to a prediction derived from a composite model (the model of “self-interested rationality,” for brevity), and it raises Duhem-Quine issues of the usual kind. Should the theory of strategic play (game-theoretic rationality) be blamed for the anomaly? Is the selfishness assumption inadequate to capture behavior in the UG? Do people’s beliefs diverge from the standard assumptions, or is some other part of the model inadequate to capture what goes on in an experimental UG?

The theory of rational play at the core of game theory is an “if . . . , then . . .” theory: it says that *if* their preferences and beliefs are so and so, *then* people will behave in such and such a way. If people’s preferences are not self-interested in the way postulated by the standard model, then the theory of rationality cannot be tested in a *single* game like the UG. “Kosher” experimental game theory thus should begin with a measurement of preferences in nonstrategic circumstances and should then proceed by checking whether the theory of strategic play issues correct predictions using the preliminary measurements as input data in the testing procedure.

However, most experimental game theory is not “kosher” in this regard. Most experiments start by *postulating* the content of individual preferences rather than trying to establish them empirically. The UG, as we have seen, is no exception in this regard. But if the UG does not really test the theory of strategic play (the “core” of game theory), why is it such a successful, widely replicated design?

Like the Prisoner’s Dilemma, the UG is so simple that no one really doubts that experimental subjects are rational—in the minimal sense that their actions follow from their preferences and beliefs. Rather, the UG is interesting because it makes *nonstandard preferences* observable. By simple intuition or introspection, it seems plausible that Respondents may prefer to give up some money to punish unfair offers and that Proposers anticipate that an unfair offer will hurt the feelings of the Respondent. Respondents have "other-regarding" preferences, possibly influenced by a concern for fairness. In fact, if we look at the theoretical developments prompted by anomalies like the UG, we can see that most of the theoretical action has been in the area of modeling nonstandard utility functions in which individual preferences are other-regarding rather than strictly self-interested, as in the standard model.[2]

The UG had the merit of making the other-regarding motives that influence behavior observable experimentally in a most vivid fashion. Whether such motives can be accommodated or not in the framework of game theory is a different question, one that I will not address here.[3] But answering that question was not essential for the UG's success, which was accomplished quite independently anyhow.

4. Institutions. Early variations on the theme of the UG tested the robustness of the anomaly and often were explicit attempts to make it go away. With hindsight, we can say that these attempts were unsuccessful. The UG's status emerged unscathed, as documented by all surveys published since the late 1980s. By the early nineties, however, the history of the UG took a new twist, with the publication of an important article by an international group of game theorists led by Alvin Roth.

Roth and colleagues (1991) aimed at testing two hypotheses: (1) Does competition matter? (2) Are there significant differences in bargaining behavior across different cultures? To test the first hypothesis, they designed, alongside the standard UG, a market where several buyers compete for the acquisition of an item owned by a single seller. Both the UG and the market game have extremely asymmetric equilibria (where one player grabs almost the entire cake). Yet the self-interested equilibrium is consistently achieved only in the market setting. Competition has the effect of "washing out" fairness considerations.

The Roth experiment is one of the most cited articles in experimental economics. It encapsulates in a single paper what are widely considered the three most important results of experimental economics so far: (1) that the model of rational selfish *homo oeconomicus* cannot explain a great number of observations; (2) that the rational self-interest model is nev-

[2] The most popular contributions are Rabin (1993), Fehr and Schmidt (1999), and Bolton and Ockenfels (2000). The experimental evidence indicates unequivocally that purely consequentialist models like the Bolton-Ockenfels and Fehr-Schmidt ones are inadequate, whereas more complex models like Rabin's do better albeit at the cost of predictive indeterminacy in a large class of games.

[3] See Guala 2006.

ertheless able to account for behavior in a wide range of experimental situations; and finally, by combining these two results, (3) that "institutions matter."

It is worth distinguishing between two kinds of institutions: informal *norms*, such as the norms of fairness, equality, cooperation, and reciprocity that govern our behavior in a variety of choice situations (like the UG), and formal *rules* of exchange, aggregation, information transmission, and so forth that are characteristic of specific relatively formalized market institutions. The importance of formal rules of exchange emerged in the Roth experiment from comparing UG bargaining and market competition. The importance of cultural norms, in contrast, was highlighted by replicating the experiment in four different countries (United States, Japan, Yugoslavia, and Israel). Roth and colleagues found some significant differences in behavior: Japanese and Israeli Proposers tend to make lower offers (mode at 40%) than Americans and Slovenians, but, interestingly, unfair offers in Tokyo and Jerusalem are as likely to be accepted as the 50/50 splits offered in Pittsburgh and Lubljana. While testing a theoretical hypothesis, for the first time the UG was also used as a *measurement device*.

5. Measurement. In 2000, the anthropologist Joe Henrich reported a series of "surprising" observations collected using the UG among the Machiguenga, a group of Peruvian slash-and-burn horticulturists. The Machiguenga proposed *more unequal* splits of the cake, and they rejected unfair offers *less* often than Western subjects. In other words, they behaved more (but not entirely) like *homi oeconomici*. The Machiguenga results could appear anomalous only against a background of established experimental knowledge, that is, once the UG results in Western societies had become "standardized" and widely accepted regularities. In the early 1980s, Machiguenga behavior would probably have been considered unsurprising or much *less* anomalous (because closer to the theoretical prediction) than it appears now.

Henrich concluded that different peoples have different cultural expectations and norms of fairness and that we need to know more about how these norms are created and sustained in each social context. With this idea in mind, a group of economists and anthropologists set out to compare behavior in some classic experimental designs (especially UG and Public Goods games) across 15 "small-scale societies" in South America, Africa, and Asia. This project, funded by the MacArthur Foundation, is the most ambitious and exciting attempt to use experimental economics for heuristics and measurement purposes ever attempted (see Henrich et al. 2004).

The MacArthur project has a number of interesting methodological features. Experimental investigation is combined with in-depth ethnographic information concerning the social context (political and economic structure, religious beliefs, rituals, food-sharing practices, etc.). This information sheds light on "surprising" results (deviations from the standard prediction) in a way that is usually precluded by standard experimental investigation. But, above all, it highlights the importance of standardization for the consolidation of paradigmatic experimental designs, which can then be used as portable instruments to measure nonexperimental phenomena in a variety of settings.

6. Standardization. Standardization is the process of consolidation of paradigmatic experimental designs. Standardization is driven by many factors. These may vary in importance at different stages of a research program.

(1) *Imitation*: as Thomas Kuhn emphasized, the trade of a scientific discipline is often learned by replicating the exemplars of a previous generation of scientists.
(2) *Robustness tests*: the very logic of controlled variation requires that new designs are compared with a benchmark. The benchmark can be a theoretical prediction, but often it is simply an earlier, "standard" experimental result that as a consequence gets replicated over and over again.
(3) *Disciplinary cohesion*: many disciplines at some point endorse and enforce some "good practices" of experimental research. Design practices such as the use of monetary incentives, anonymity, repetition, and the ban on deception are relevant examples in the case of experimental economics.[4]

Via these processes, standardization leads to the stabilization of phenomena and the related stabilization of an experimental design.[5] Once a design has been standardized, some of its details lose their original methodological justifications and are retained primarily for pragmatic purposes: they are means to achieve comparability across diverse subject populations.

[4] Some of these practices are formalized in the so-called precepts of experimental economics (Smith 1982). Others are justified less formally. Guala 2005, Chapter 11, includes a more detailed illustration and discussion.

[5] Sugden (2005) captures both aspects nicely in his concept of *experimental exhibit*: a phenomenon attached to a standard design, like a ship in a bottle. Boumans (2005) correctly highlights the other side of the coin: sometimes the robustness of a phenomenon leads to the standardization of a design—the design that best captures the phenomenon in its "pure" manifestation.

7. External Validity. It is often said that we face a UG every time we enter a chain store. We are offered one price, and there is hardly any opportunity to haggle. Yet, in a market economy, we rarely perceive the prices as "take it or leave it" offers. We usually have other options, like simply walking into a different store that sells similar goods and seeing if it offers more attractive prices. When we do accept without a grudge, it is because we believe (or assume) that the price results from a fair process of market competition and that it is more or less the best that can be offered given production and other costs. (Of course this is not always the case, which is why consumers' associations, watchdogs, and consumers' boycotts have a role to play in a market economy.)

Fairness considerations emerge differently in different contexts, and the UG can at best be considered a stylized representative exemplar for a whole range of situations where fairness norms play a more or less direct role in economic and social behavior. This is especially relevant for the MacArthur Foundation project. Some of the small societies had little experience of market transactions, and sometimes they were even unfamiliar with money. When the experimental task "tapped" onto some local institution, it was mostly by chance rather than by experimental intention and design.[6]

External validity is the problem of generalizing experimental results from laboratory conditions to other situations of interest. It is generally considered the Achilles' heel of much experimental research in the social sciences, although its relevance to natural science experiments is a largely unexplored—and possibly underestimated—issue. I have argued elsewhere (Guala 2005, Chapters 7–10) that external validity must be resolved on a case-by-case basis. There is no point in questioning the external validity of an experiment *in general*. One must first ask what the target of the experiment is—what nonlaboratory system or phenomenon we intended to study in the first place. Often experiments are even performed without an external target in mind, and thus external validity is not their primary goal or concern.

Intuitively, standardization seems to be the enemy of external validity. If the social and economic world is diverse and complicated, flexibility rather than rigidity of design is required. We need to be able to tailor our designs to the specific research questions that are prompted by the problem we are studying rather than the other way around. So how can the consolidation of a rigid format for the UG help and explain its success? Because, I will argue, the methodological criteria for good measurement are different from those required for good testing.

[6] For a clear example, see Ensminger 2004 on the *harambee* institution among Orma people in Kenya.

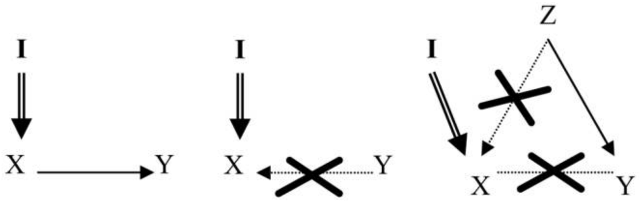

Figure 1. Intervention and causal discovery.

8. Observing and Intervening. Philosophers belonging to so-called New Experimentalism have identified material intervention as the *differentia specifica* between experimental and nonexperimental science. According to one version of this view developed by James Woodward (2003), intervention is functional to the discovery of *robust causal relations*. An intervention, according to Woodward, is a "surgical" manipulation of a variable that leaves the rest of the experimental system intact. Consider Figure 1. Suppose that we want to know whether X causes Y, Y causes X, or the two are only spuriously correlated. In a "surgical" intervention (I), X is manipulated in such a way that all the links with its causal parents are severed, but none of its putative effects is directly affected by the manipulation. In other words, any change in Y due to I must take place via X. If the intervention is successful, and X causes Y, then variations in X will be reflected in variations in Y. If, in contrast, Y causes X (middle case) or the two are only spuriously associated (right-hand case), the correlation between X and Y will be disrupted.

In using the UG as a measurement instrument, we are not comparing two (or more) states of a variable that we can set at will. We are surely comparing *something*—the number of rejections to unfair offers in societies A and B, for example—but the value of the variable is assumed to be determined by some underlying factors upon which we have no direct control. Indeed, it is extremely important that the experimental procedures do not interfere with the operation of these underlying factors, for otherwise the measurement would be considered "artifactual" and useless.

Of course, when the UG is used as a measurement device, there is a sense in which we are "testing" something. We are testing the hypothesis that two sets of observations (e.g., collected in the United States and among the Machiguenga) come from the same population or data-generating process. It is precisely in order to test this proposition that we need to keep the experimental design as stable (or "standardized") as possible. Otherwise, it will not be possible to attribute differences in the observed behaviors to differences in something *other* (whatever that may be) than the experimental design itself. An analogy with thermometry may be useful: suppose that you want to demonstrate that water boils at different temperatures at sea level and on a high plateau. A standardized, portable thermometer is needed, one that is robust to variations in the external environment when it is transported from one measurement site to another. Since we do not want to interfere with the variable that we are measuring (the temperature), only manipulations that are functional to obtaining an objective and standardized measure that we can use for prediction and comparison will be allowed.

Historically, temperature has been measured for a long time quite independently from an even remotely correct understanding of the nature of heat (Chang 2004). The first step in the development of thermometry is the construction of a *thermoscope*. A thermoscope measures temperature on an ordinal scale and can be used to determine one or more *fixed points*—for example, the freezing and the boiling points of water. Once these have been fixed, one can construct a *thermometer* that measures temperature on a cardinal scale. Historically, a "good" thermoscope was judged by its conformity with ordinary sensations of heat and cold: it must expand when we feel hotter and contract when we feel colder. This is to make sure that we are measuring something that matters (to us), but of course this cannot guarantee that we are hitting on a physically significant robust relationship.

The same applies to economic experiments when they are used as measurement devices. The causes of variations in UG behavior are still controversial, and indeed it is possible that different causes may be at work in different social contexts. And yet the UG can be used for measurement purposes if proper standardization and care in running the measurement sessions has been achieved. When this is the case, we can conclude with a high degree of confidence that whatever difference is observed in the data is due to differences in the populations or data-generating processes. The inference, to use Deborah Mayo's (1996) expression, is warranted because it has been *severely tested*.

A severe test produces a certain kind of data, D_1, if the hypothesis under test is true, and another kind of data, D_2, if the hypothesis is false. In more familiar statistical terms, it minimizes the probability of making errors of types I and II. Severe testing, according to Mayo, is the hallmark of experimentation because good experiments typically support inferences that have been severely tested by the experiment itself.[7] Woodward's characterization of experiments as aimed at causal discovery is largely consistent with this view. Let's go back to Figure 1. If the intervention is successful, and X causes Y (left-hand case), then variations in X will be

[7] Or, more precisely, inferences to hypotheses that have been severely tested.

reflected in variations in Y. If, in contrast, Y causes X or the two are only spuriously associated (middle and right-hand case), the correlation between X and Y will be disrupted. In a genuine, competently performed experiment, in other words, one kind of data will be observed if "X causes Y" is true and another kind if it is not.[8] Whatever causal conclusions we end up drawing, they will be severely probed or tested according to Mayo's criterion.

9. Experiments and Field Data. A standard design allows the differences among various settings to emerge. It matters little what these differences are. They could be due to local uncontrolled factors (habits, norms, psychological dispositions), but they would still be interesting because these differences would be precisely what we were looking for. Indeed, one of the striking lessons taught by cross-cultural studies is how difficult it is to generalize across the various populations.

The MacArthur Foundation studies show that more or less cooperative behavior is correlated with different factors in different societies (wealth, gender, age, political structures, market exposure, etc.). And, even more important, they demonstrate by example that such relationships would *never* have been discovered without the in-depth ethnographic knowledge provided by the anthropologists in the field. In his study of the Hadza, a group of hunter-gatherers in Tanzania, Marlowe (2004), for example, reports a surprisingly low degree of sharing and cooperation in an otherwise distinctively egalitarian society. The proposed explanation appeals to a subtle distinction between what can be seen (e.g., big game) and what can be easily hidden (e.g., money). The Hadza norms impose the sharing of what can be seen, but they do not effectively prohibit the individual ownership and consumption of smaller goods. Scrounging is tolerated when it cannot effectively be sanctioned, whereas social control is implemented for big hunts.

Patton (2004) reports a higher degree of cooperative behavior among the Achuar than among the Quichua, two closely related groups living in the Ecuadorian Amazon forest. This relationship holds both in the experimental data and in the patterns of meat sharing observed in the two communities. However, in-group differences show that the greater generosity of the Achuar results almost entirely from a small elite of leaders who use it for political purposes, especially to create stable and strong coalitions among families and villages. The Quichua, who have enjoyed a less stable political structure and where coalitions tend to be more ephemeral, have simply failed to generate such a class of political sharers.

[8] See also Woodward 2000.

Knowledge of the fine structure of the society and its norms is crucial because "macro" correlations are often misleading or useless. Each case is different; every explanation is "local".[9]

10. Social Thermometers. A thermometer is the beginning of a diagnosis, not the end of the story. It measures a symptom (an important and ubiquitous symptom), not a cause. This does not mean that it is less valuable—on the contrary, life would be much worse without thermometers. Like a good thermometer, the UG has become a portable tool that we can transport from culture to culture to measure a set of phenomena that we believe are associated with social norms. What these phenomena are exactly is still controversial. Unlike in the case of temperature, where the thermometers could be compared with unaided sensations of heat and cold, we do not have such direct benchmarks for the UG. We do "feel" that the UG can be used to measure what is perceived as fair or unfair, but our own concept of fairness is not necessarily a good guide to the ones that are prevalent in the societies we are studying. Our conception of fairness can work as a benchmark, however, and a benchmark is precisely what we need for a measurement enterprise. The UG emerged by a process of social selection in experimental game theory, as a robust design that "taps on" something that seems to matter for us. There is no guarantee that it is the best measurement tool to study what matters to other people living in different societies. But it is a starting point, and social science needs badly some stable platform in order for rigorous research to take off.

REFERENCES

Bolton, Gary, and Axel Ockenfels (2000), "ERC: A Theory of Equity, Reciprocity and Cooperation", *American Economic Review* 90: 166–193.

Boumans, Marcel (2005), *How Economists Model the World into Numbers*. London: Routledge.

Chang, Hasok (2004), *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Ensminger, Jean (2004), "Market Integration and Fairness: Evidence from Ultimatum, Dictator, and Public Goods Experiments in East Africa", in Henrich et al. 2004, 356–381.

Fehr, Ernst, and Klaus Schmidt (1999), "A Theory of Fairness, Competition and Cooperation", *Quarterly Journal of Economics* 114: 817–868.

Gil-White, Francisco (2004), "Ultimatum Game with an Ethnicity Manipulation: Results from Khodovin Bulgan Sum, Mongolia", in Henrich et al. 2004, 260–304.

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

——— (2006), "Has Game Theory Been Refuted?", *Journal of Philosophy* 103: 239–263.

Gurven, Michael (2004), "Does Market Exposure Affect Economic Behavior? The Ulti-

[9] See also Gil-White 2004; Gurven 2004; and Tracer 2004 for similar examples.

matum Game and the Public Goods Game among the Tsimane' of Bolivia", in Henrich et al. 2004, 194–231.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarz (1982), "An Experimental Analysis of Ultimatum Bargaining", *Journal of Economic Behavior and Organization* 3: 367–388.

Henrich, Joseph (2000), "Does Culture Matter in Economic Behavior? Ultimatum Game Bargaining among the Machiguenga of the Peruvian Amazon", *American Economic Review* 90: 973–979.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis, eds. (2004), *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies.* Oxford: Oxford University Press.

Marlowe, Frank (2004), "Dictators and Ultimatums in an Egalitarian Society of Hunter-Gatherers: The Hadza of Tanzania", in Henrich et al. 2004, 168–193.

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

Patton, John (2004), "Coalitional Effects on Reciprocal Fairness in the Ultimatum Game: A Case from the Ecuadorian Amazon", in Henrich et al. 2004, 96–124.

Rabin, Matthew (1993), "Incorporating Fairness into Game Theory and Economics", *American Economic Review* 83: 1281–1302.

Roth, Alvin, Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir (1991), "Bargaining and Market Behavior in Jerusalem, Lubljana, Pittsburgh, and Tokyo: An Experimental Study", *American Economic Review* 81: 1068–1095.

Smith, Vernon (1982), "Microeconomic Systems as an Experimental Science", *American Economic Review* 72: 923–955.

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

Tracer, David (2004), "Market Integration, Reciprocity, and Fairness in Rural Papua New Guinea: Results from a Two-Village Ultimatum Game Experiment", in Henrich et al. 2004, 232–259.

Woodward, James (2000), "Data, Phenomena, and Reliability", *Philosophy of Science* 67 (Proceedings): S163–S179.

——— (2003), *Making Things Happen*. New York: Oxford University Press.