Is Experimental Economics Living Up to Its Promise?

By Alvin E. Roth

This draft, January 20, 2010

Forthcoming in Fréchette, Guillaume and Andrew Schotter (editors) *The Methods of Modern Experimental Economics*, Oxford University Press

Introduction

The question that is the title of this essay already suggests that experimental economics has at least reached a sufficient state of maturity that we can try to take stock of its progress, and consider how that progress matches the anticipations we may have had for the field several decades ago, when it and we were younger. So it will help to begin by reconstructing what some of those anticipations were.

When I surveyed parts of experimental economics in Roth (1986-7,1988), I hoped that experimentation would facilitate and improve three kinds of work in economics, which I called *Speaking to Theorists*, *Searching for Facts*, and *Whispering in the Ears of Princes*. By *speaking to theorists* I meant testing the empirical scope and content of theories (including especially formal theories that might depend on factors hard to observe or control outside the lab), and in particular testing how well and on what domains their quantitative and qualitative predictions might serve as (at least) useful approximations. By *searching for facts* I meant exploring empirical regularities that may not have been predicted by existing theories, and might even contradict them, but whose contours, once they had begun to be mapped by experiments, could form the basis for new knowledge and new theories. And by *whispering in the ears of princes* I meant formulating reliable advice, as well as communicating, justifying, and defending it.

Of course, whether experimental economics is living up to its promise could also be a question about how experimental economists are doing at developing a body of experimental methods

and knowledge, and creating a lively, self-sustaining and productive community of economic research, held in high regard by the larger community. I'll address this last question first.

How are we doing at building a research community?

There are now many experimental economists and laboratories around the world. A list compiled at the University of Montpelier locates more than 135 labs in 22 countries, including concentrations of 8 in France, 15 in Germany, 12 in Italy, 7 in Spain, 11 in the U.K. and 52 in the U.S.[1] There is also a professional society devoted to experimental economics, the Economic Science Association, which sponsors regular meetings, a journal, and an active internet discussion list (esa-discuss@googlegroups.com) that allows participants to quickly query the larger community with questions of all sorts, including questions about prior work, and about solutions to particular problems of experimental design or analysis.

Regarding acceptance by the larger community of economists, experiments are now regularly reported in the top general-interest journals, experimenters are employed by highly ranked departments, and the Nobel memorial prize in economics has been awarded to a number of experimenters since its inception in 1968. A look at some of the Nobelists can serve as a metaphor for the progress of the field.

Maurice Allais won the prize in 1988. Although he is well known for the hypothetical choice experiment called the Allais paradox (Allais, 1953), experiments were not a persistent part of his work, nor part of the work that the Nobel committee cited him for, on general equilibrium theory. Reinhard Selten won the prize in 1994. Experiments are a large and continuous part of his work, and indeed he's one of the earliest experimenters who conducted experiments throughout his career (see e.g. Sauermann and Selten 1959, and Selten and Chmura 2008). But his experimental work isn't particularly related to the work he was cited for, which was the development of the theory of perfect equilibrium, a concept that in fact often performs poorly in experiments. Daniel Kahneman and Vernon Smith shared the prize in 2002. Their prize was specifically for experimental economics, and experimentation constitutes the lion's share of

[1] http://leem.lameta.univ-montp1.fr/index.php?page=liste_labos&lang=eng , accessed August 2009.

their work. This brings us to the 2009 prize to Elinor Ostrom. Experiments are an important part, but not the most important part, of the work she was cited for. Her experiments complement her field work (see e.g. Ostrom 1998; Dietz, Ostrom, and Stern 2003)).[2]

There is a sense in which this history of Nobel prizes parallels how experimental economics has grown. Early experiments were done only sporadically (e.g. the Prisoner's dilemma game, which has spawned as many experiments as anything in the social sciences, was formulated for an experiment concerning Nash equilibria conducted in 1950 at the Rand Corporation, and subsequently reported by Flood (1952, 1958), who did not persist in doing experiments). Experiments became increasingly important to a relatively small group of specialists, but only more slowly achieved widespread recognition in the profession. And today experiments are flourishing in part because of how well they complement, and are complemented by, other kinds of economic research.

There are respects in which experimental economics as a community is still struggling. Chief among these is that the majority of economics departments do not have an experimental economist on their faculty, let alone a dedicated laboratory. (This may reflect low supply as much as low demand.) How this is likely to change is yet to be seen: maybe more economists will devote themselves primarily to experiments, or maybe more economists who don't primarily identify themselves as experimenters will occasionally do experiments when they need to. In the meantime, while the experimental revolution in economics is pretty well won in the journals, it still has some way to go as measured by employment in economics departments.

How have we done at speaking to theorists and searching for facts?

Moving back to substantive questions, how well has experimental economics begun to live up to its promise for changing the way theories are discussed, tested, and proposed? How productive has been the algorithm embodied in what we might call *the experimental cycle of*

[2] John Nash, who shared the 1994 prize with Selten and Harsanyi, and Thomas Schelling, who shared the 2005 prize with Aumann, also have been involved in some significant experiments, although experiments didn't play a large role in their careers. But Schelling (1957, 1958, 1960) reported important early experiments.

creative destruction that proceeds from Theory Testing to Exploring Unpredicted Regularities to Theory Building and back to Theory Testing?

Of course, the same 'algorithm' could be said to apply to any program of theory testing, but experiments speed it up. If the test of theories waits on appropriate data to emerge through processes uncontrolled by the investigators, progress will be slow, and in economics slow progress sometimes takes generations. But if you think some theory or some experiment is unreliable, or is being over-generalized or misinterpreted, you can quickly do an experiment motivated by your hypothesis, so the conversations among economists that are conducted with the help of experiments can move (relatively) fast. To my eye, this process has been quite productive, and I'll illustrate what I mean by reviewing (briefly and from a very high altitude) two experimental programs that began with theory testing, having to do with individual choice, and bargaining.

Individual choice behavior

I won't discuss here the long process that led to utility theory and then subjective expected utility theory becoming the canonical models of individual choice in economics, except to recall that there were a few early experiments that played a role. See Roth (1993 or 1995a) on the early history of experimental economics for accounts of Bernoulli's 1738 hypothetical choice experiment on the Petersburg Paradox, and Thurstone's 1931 experiment on indifference curves. Even this line of experiments elicited a methodological critique of experiments in economics, by Allen Wallis and Milton Friedman (1942).

But once utility theory had taken pride of place among economists' models of individual choice behavior, it started to be tested by experiments, including several early ones published in Economics journals that foreshadowed the more sustained investigations that took place in the 1970's by psychologists as well as economists. I have in mind here not only Allais' (1953) 'paradoxical' demonstration of what later came to be generalized and called the common ratio effect (see e.g. Camerer, 1995), and Ellsberg's (1961) demonstration of ambiguity aversion, but also demonstrations like that of May (1954) that even ordinal preferences could be intransitive.

In the 1970's, as choice experiments found yet more reproducible violations of the predictions of expected utility theory, some of the unpredicted regularities also came to be more thoroughly explored by multiple investigators, among whom Amos Tversky and Danny Kahneman were prominent. Their proposal for Prospect Theory (Kahneman and Tversky 1979) was meant to offer a replacement for expected utility theory that captured some of these regularly observed departures from utility theory, such as the overestimation of small probabilities, as well as other behavioral regularities that were not necessarily in conflict with utility theory, such as dependence of choices on reference points, and different patterns of risk

aversion and “loss aversion” with respect to gains and losses as measured against those reference points.

Together with subsequent refinements of the theory meant to allow it to be used to make predictions (Tversky and Kahneman 1992), prospect theory came to seem to some investigators like a plausible replacement for utility theory. That is, prospect theory has a lot in common with utility (e.g. in treating individuals as having preferences), but adds parameters meant to allow it to accommodate reproducible violations of utility theory and other regularities that had been observed in experiments. [I pass silently over experiments that raised more fundamental questions about when modeling individuals as having well-defined and stable preferences is a useful approximation.]

Contemporary experiments have started focusing the sort of critical attention on prospect theory that early experiments focused on utility theory. For example, some of the regularities encoded by prospect theory turn out to be sensitive to how choices are elicited. Harbaugh, Krause, and Vesterlund (forthcoming) find that asking subjects to name a price they are willing to pay for a lottery induces quite different patterns of risk aversion for gambles over large and small gains and losses than does asking them to choose between a lottery and a riskless payment. In a similar way, Ert and Erev (2009) find that patterns that were interpreted as loss aversion when subjects were asked if they wished to participate in lotteries involving potential gains and losses disappear when subjects are instead asked to choose between the lottery or receiving zero for certain.

In a different kind of investigation of prospect theory, a series of papers by Ido Erev and colleagues (e.g. Barron and Erev 2003, 2005 and Hertwig, Barron, Weber and Erev 2004) show that how subjects react to small probabilities depends on how they learn them.[3] When subjects have lotteries described to them with numerical probabilities (as in the experiments of Kahneman and Tversky that motivated prospect theory), they tend to over-weight small probabilities. However, when subjects learn about lotteries by experiencing them multiple times (but without having the probabilities described), they tend to under-weight small probabilities.

Thus some of the behavioral regularities encoded in prospect theory may be more closely related to the way the experiments investigating utility theory were performed than was earlier appreciated.

[3] The 2003 and 2004 papers address this issue directly; the 2005 paper looks at models of learning that might help account for it.

To put his work in perspective, note that a major focus of mainstream behavioral economic research involved experiments designed to find and study counter-examples to rational decision theory, and specifically examples in which expected utility theory can be shown to make a false prediction. This led to a concentration of attention on situations in which utility theory makes a clear, falsifiable prediction; hence situations in which all outcomes and their probabilities are precisely described, so that there is no room for ambiguity about subjects' beliefs. One consequence of this is that decisions in environments in which utility theory does not make precise predictions received less attention. Environments in which participants are free to form their own beliefs fall into this category, since sometimes any decision is consistent with utility theory when beliefs cannot be observed or controlled. Decisions made in environments in which probabilities are not precisely described, but are left for subjects to learn from experience, are of this kind.

Once parts of the experimental cycle of creative destruction became less exclusively focused on utility theory, it was no longer an essential feature of good experimental design that the predictions of utility theory could be unambiguously determined for each task studied. So a much wider range of experiments opened up, some of them addressed to situations of great economic importance that had previously received less attention, like how subjects' choice behavior reflects their experience. The experiments concerning choices based on experience weren't designed to test utility theory (since they didn't pin down subjects' beliefs about the lotteries they experienced), but allowed the discovery of behavioral regularities that couldn't have been guessed from experiments that introduced the lotteries with numerical probabilities.

Even before debates about whether and how utility theory might be replaced or improved are resolved, we have gained robust new knowledge from this heritage of choice experiments. To the extent that utility theory is a useful approximation, it can only make it more useful to know that it is likely to be less accurate when choices involve small probabilities. (Whether small probabilities are likely to be over or under weighted compared to the frequency of the events they determine may depend on how information about those probabilities is acquired…) So, to the extent that we want to use utility theory as an approximation, we have become more aware that it is more of an approximation in some kinds of situations than in others.

Bargaining behavior

A similar story of theory and experiments, followed by more theory and more experiments, could be told about many topics. In the case of bargaining theory the experimental cycle of creative destruction eventually gave rise to new theories that aim to organize data over a broader class of phenomena than merely bargaining.

In the 1970's, Nash's 1950 model was economists' canonical model of bargaining. A central assumption of Nash's model (and a number of related models, see Roth, 1979), was that the outcome of bargaining could be predicted from the information contained in the expected utility payoffs to the bargainers from each potential outcome. In particular, Nash's model predicted that the outcome of bargaining in otherwise symmetric situations would be determined by differences in the bargainers' risk aversion. It was tested by experiments that assumed that all subjects were risk neutral, and under this assumption some important aspects of its predictions could be rejected (see e.g. Rapoport, Frenkel, and Perner 1977). But economists were largely unpersuaded by these experiments, basically because of the feeling that a model whose predictions were based entirely on differences in risk aversion could not be adequately tested by assuming that there were no differences in risk aversion. To put it another way, the theory's predictions depend on knowing what the utility payoffs are, and without some controls, it might not be adequate to identify money payoffs with utility payoffs.

To test the theory in an environment that allowed unobserved risk aversion to be controlled, Roth and Malouf (1979) introduced the technique of binary lottery games, which has since been used fairly widely to test theories that depend on risk aversion as modeled by expected utility functions. In these experiments, payments are made in lottery tickets (i.e. in probability of winning a lottery), and the outcome of the lottery is binary, the player always wins one of two prizes (sometimes one of the prizes is zero). An expected utility maximizer would be risk neutral in lottery tickets because expected utility is linear in probability, and so in an experiment that uses binary lottery payoffs, the predictions of a theory that depends on payoffs measured in expected utility can be made unambiguous, so that they can be tested.

Note that the use of binary lottery payoffs is to control for the predictions of the theory being tested, and not to control the behavior of the experimental subjects. The subjects of the experiment may not themselves be utility maximizers in the manner that a theory predicts, but binary lottery payoffs allow the experimenter to know exactly what it predicts, so that the observed behavior can be compared to the predictions. Note that there is good reason to expect that binary lottery payoffs would *not* influence subjects' behavior in the way it would if they were ideal expected utility maximizers, since the binary lottery design depends on the theory's linear treatment of probabilities, and, as noted above, subjects seem to treat (at least) small probabilities nonlinearly.

A quick digression is in order here, having to do with how I saw the promise of experimental economics, thirty years ago. We sent Roth and Malouf (1979) to the journal *Psychological Review* rather than to an economics journal because I had what turned out to be a mistaken idea of how the interaction between economics and psychology might develop. I thought there might (continue to) be a division of labor, in which economists would theorize and

psychologists would experiment, and that, in areas of potential mutual interest to economists and psychologists, what had kept economists from properly appreciating psychologists' experiments, and psychologists from properly appreciating economists' theories, might be lack of familiarity. So I thought that if we put an economic experiment in *Psych Review*, with an experimental design that would address economists' concerns about experimental tests of theories stated in expected utilities, psychologists would pick it up. In fact, psychologists weren't interested in testing the theories that attracted economists, or, when testing them, weren't interested in controlling for what economists regarded as plausible alternative hypotheses. So it turned out that if we wanted economics to have a robust experimental component, we would have to do experiments ourselves. (Although Roth and Malouf (1979) eventually became well cited by economists, I don't think it ever appealed to psychologists.)

In any event, that bargaining experiment, and some subsequent ones published in economics journals (Roth and Murnighan, 1982; Roth and Schoumaker, 1983) helped remove Nash's model from its position as economists' default model of bargaining. For a time at least, Rubinstein's (1982) alternating offer/perfect equilibrium model of bargaining over a shrinking pie became the new most popular model, with impatience (and first mover status) rather than risk aversion playing the lead role in differentiating between bargainers. And, as in the case of individual choice theory, once parts of the experimental cycle of creative destruction became less focused on bargaining theories based on expected utility theory, it was no longer an essential feature of good experimental design that the predictions of utility theory could be unambiguously determined for each task studied. Instead, tests of bargaining theory came to involve tests of perfect equilibrium, initially under the assumption that monetary payoffs were a good proxy for players' ordinal preferences.

But perfect equilibrium behavior was not observed in one-period ultimatum games (Guth, Schmittberger and Schwarz 1982), nor (after some preliminary suggestion otherwise by Binmore and Shaked 1985) in multi-period alternating offer bargaining games (Neelin, Sonnenschein and Spiegel 1988). Ochs and Roth (1989) observed that the perfect equilibrium predictions (under the assumption that money was a good proxy for bargainers' ordinal preferences) did not do well, but also that bargainers' preferences seemed to be more complex. In particular, we noted that in our experiment, and in the earlier experiments reported by others, unequal offers were often rejected and then followed by "disadvantageous counterproposals," that, if accepted, would give the bargainer a smaller monetary payoff than if he had accepted the original offer. This seemed to indicate a preference for more equal divisions, even at the cost of lower monetary payoffs.

Bolton (1991) followed with an experiment and a theory of bargaining that explicitly incorporated a preference for "fairness" in bargainers' utility functions, and this was followed by more comprehensive theories of other-regarding preferences meant to explain not only bargaining games but also market games run under comparable conditions, in which the

"pecuniary" perfect equilibrium (in terms of self-regarding monetary payoffs) performed better (see Roth, Prasnikar, Okuno-Fujiwara, and Zamir 1991). These new theories (Bolton and Ockenfels 2000 and Fehr and Schmidt 1999) kept the perfect equilibrium assumption, but replaced self-regarding utility functions with utility functions in which agents cared about the distribution of monetary payoffs among all players. (A different direction was taken by theories of learning proposed to explain the same kinds of experimental behavior, by replacing the perfect equilibrium assumption with a model of bounded rationality, while keeping the assumption that players were concerned only with their own payoffs; see Roth and Erev 1995.)

Each of these theories has inspired new experiments, and new theories (e.g. learning theories with more or different parameters, and other-regarding preference theories that incorporate preferences concerning intentions and expectations as well as payoffs), and these theories are in turn being tested by further experiments. So here too, the creative cycle of experimental destruction is operating with vigor.

How are we doing at whispering in the ears of princes?[4]

Economists give all sorts of advice, and so we whisper in the ears of many different kinds of princes. I'll focus here on market design, an area in which, recently, economists have succeeded in participating from the initial conception and design of markets all the way to their eventual adoption and implementation.

Experiments are a very natural part of market design, not least because to run an experiment, an experimenter must specify how transactions may be carried out, and so experimenters are of necessity engaged in market design in the laboratory. And experiments can serve multiple purposes in the design of markets outside of the lab. In addition to the ordinary scientific uses of experiments to test hypotheses, experiments can be used as testbeds to get a first look at market designs that may not yet exist outside of the laboratory (cf. Plott, 1987), and experiments can be used as demonstrations and proofs of concept.

My sense is that, in the first attempts to employ experiments in practical market design, experimental studies by themselves were expected to bear almost all of the weight of the argument for a particular design. More recently, experiments have served a more modest but effective role, as complements to other kinds of work, in bringing market designs from conception to implementation.

[4] This section borrows from my forthcoming chapter on Market Design in volume 2 of the Handbook of Experimental Economics, in which more detail will be found.

An early example of very creative experimental work with market design as its goal took place in the 1970's and 80's, when the topic of allocating slots at the nation's busiest airports was raised. Experiments by Grether, Isaac, and Plott (1979), and by Rassenti, Smith and Bulfin (1982) helped direct attention to important design issues about the auction of takeoff and landing slots that remain relevant today. But they remain relevant today partly because these early efforts were unsuccessful at persuading policymakers to adopt auctions, and the political and other problems involved in doing so have yet to be solved.

Another effort to have experimental results translated directly into design decisions came in the early 1990's, when Congress passed legislation requiring the Federal Communications Commission (FCC) to design and run an auction for radio spectrum licenses. The FCC called for extensive public comment and discussion, and many economists were hired by telecommunications firms and the FCC itself to participate in the process. This eventually became one of the success stories for the involvement of economists from the initial design to implementation of a market. Plott (1997) gives an account of some of the ways experiments and experimenters played a role in that process, see also Ledyard, Porter, Rangel (1997). But many of the most influential economists in the process were not experimenters, but rather auction theorists such as Paul Milgrom and Robert Wilson.

Vernon Smith (2008), in a chapter on the FCC auctions, attributes what he feels is a lack of success by experimenters at influencing policy to mistaken positions taken by policymakers and these other economists due to: "entrenched resistance," (p131), "casual empiricism" (139), "mistakes," (139), "elementary errors," (140), "remarkably casual empiricism" (145), "early designers were all inexperienced" (148), and "both users and designers have become accustomed to the fantasy that strategizing can be controlled by ever more complex rules without significantly increasing implementation costs for everyone."(148)

I think Smith may underestimate the influence that experiments had in helping to shape some of the discussions about the design of the FCC auctions, but he is certainly correct that none of the particular design proposals advanced by experimenters were adopted.[5]

[5] Milgrom (2007) adds an interesting dimension to the discussion of how experiments were used in the policy discussion, writing of one of the consulting reports (Cybernomics, 2000) that presented a particular proposal based on an experiment (p953): "Cybernomics presented its results to the FCC in a report and at a conference, where they were represented by two highly regarded academic experimenters: Vernon Smith and David Porter. … The Cybernomics report is not detailed enough to enable a fully satisfactory assessment of its results. The FCC contract did not require that detailed experimental data be turned over to the sponsors. When the FCC and I later asked for the data, we were told that they had been lost and cannot be recovered."

But the design of spectrum auctions is an ongoing process (although design changes now come slowly), and experiments continue to play a role in the discussion in the scientific literature, see e.g. Kagel, Lien, and Milgrom (2009), or Brunner, Goeree, Holt, and Ledyard (forthcoming) for contemporary discussions of combinatorial auctions as compared to the simultaneous ascending auctions that have become the standard design. Kagel et al. point in particular to how the development of appropriate theory helps in the design of an experiment investigating a domain like combinatorial auctions, in which the space of potential combinations and valuations created by even the simplest experimental environment is much bigger than can be meaningfully explored without some guidance about where to look.

In general, experiments have begun to play a more modest but more effective role in helping market designs by economists become implemented in functioning markets. I've worked on the design of labor markets for doctors, schools choice systems, and kidney exchange, and I'll concentrate here on the design of medical labor markets, since experiments have so far been best integrated in that work (see Roth 2002, 2008b). In particular, experiments have played roles in diagnosing and understanding market failures and successes, in exploring new market designs, and in communicating results to policy makers.

I'll briefly give examples from the design process for two medical labor markets. The first is the redesign of the labor clearinghouse through which American doctors get their first jobs, the National Resident Matching Program (see Roth and Peranson 1999), and the second involves the reorganization of a labor market for older physicians seeking gastroenterology fellowships, the entry level positions in that subspecialty (see Niederle and Roth, 2010).

Designing labor markets for doctors

New medical graduates

By the time I was asked in 1995 to direct the redesign of the big American clearinghouse that places most doctors in their first jobs, the National Resident Matching Program had been in operation for almost half a century, and I had studied it, and related clearinghouses around the world, both empirically and theoretically. The body of theory that seemed most relevant to the redesign of the NRMP was the theory of *stable matchings* (summarized at the time in Roth and Sotomayor 1990), since Roth (1984) had showed that the early success of the NRMP in the 1950's arose when it adopted a clearinghouse that produced matchings that were stable in the sense of Gale and Shapley (1962). Subsequent studies suggested that the stability of the outcomes played an important role in the success of other labor market clearinghouses (see e.g. Roth (1990, 1991, 2008a). Except for the last two lines of Table 1, which concern the experiment I'll come to in a moment, the table reports some of the relevant field observations. For each of the clearinghouses listed, the first column of the table reports whether it produced

a stable outcome, and the second column reports whether the clearinghouse succeeded and is still in use.

Table 1

Market	Stable	Still in use (halted unraveling)
• NRMP	yes	yes (new design in ’98)
• *Edinburgh ('69)*	*yes*	*yes*
• *Cardiff*	*yes*	*yes*
• *Birmingham*	*no*	*no*
• *Edinburgh ('67)*	*no*	*no*
• *Newcastle*	*no*	*no*
• *Sheffield*	*no*	*no*
• Cambridge	no	yes
• London Hospital	no	yes
• Medical Specialties	yes	yes (~30 markets, 1 failure)
• Canadian Lawyers	yes	yes (Alberta, no BC, Ontario)
• Dental Residencies	yes	yes (5) (no 2)
• Osteopaths (< '94)	no	no
• Osteopaths (≥ '94)	yes	yes
• Pharmacists	yes	yes
• Reform rabbis[6]	yes	yes
• Clinical psych[7]	yes	yes
• Lab experiments	yes	yes
• “	no	no

[6] First used in 1997-98.
[7] First used in 1999.

From the empirical observations, stability looks like an important feature of a centralized labor market clearinghouse. Because the clearinghouses involved are computerized, their rules are defined with unusual precision, which makes questions about stability much easier to answer than in decentralized markets. Nevertheless, the empirical evidence is far from completely clear, not least because there are other differences between these markets than how their clearinghouses are organized. E.g. there are differences between Edinburgh, in Scotland, and Newcastle, in England, other than whether their medical graduates were matched using a stable matching mechanism.

There are even more differences between the markets faced by medical graduates looking for jobs in Britain's National Health Service and those faced by new American doctors seeking employment in the decentralized U.S. market. The differences between those markets were very clear to American medical administrators, who therefore had reason to question whether the evidence from the British markets was highly relevant for the redesign of the American clearinghouse. And the question of whether a successful clearinghouse had to produce stable matchings had important policy implications, concerning for example whether the shortage of young doctors at rural hospitals could be addressed by the redesign of the clearinghouse (Roth, 1986 showed that under-filled hospitals would be matched to the same set of new doctors at every stable matching).

There was thus a need for experiments to help investigate if the difference between matching mechanisms could account for the differential success of clearinghouses that had been observed to fail or to succeed in the field. That is, an experiment would allow these different mechanisms to be examined without the confounding effect of differences between different regions of the British National Health Service, for example.

Kagel and Roth (2000) reported an experiment that compared the stable algorithm used in Edinburgh and Cardiff with the unstable "priority" algorithm used in Newcastle and in slightly different versions in Birmingham and Sheffield. The point of the experiment was not, of course, to reproduce the field environments, but rather to create a simpler, more controlled environment in which the clearinghouse algorithm could be changed without changing anything else.

The experiment examined laboratory markets consisting of 6 firms and 6 workers (half "high productivity" half "low productivity"). Subjects received about $15 if they matched to a high productivity partner, and around $5 if they matched to a low productivity partner, and there were three periods in which matches could be made:-2, -1, 0, with the final payoff being the value of the match minus $2 if made in period -2, or minus $1 if made in period -1. That is, there was a cost for matching early, before period 0.

However the experimental markets initially offered only a decentralized match technology: firms could make one offer in any period if they were not already matched. Workers could accept at most one offer. This decentralized matching technology suffers from congestion: firms would like to make more offers than they are able to at period 0, and a firm that waited until period 0 to make an offer would run a risk that its offer would be refused by a worker who had received a preferable offer, and it would be unmatched. So firms learned from experience that they had to make offers early, even though this was costly. (In this simple experiment, the costs of going early were simply the fines imposed by the experimenters.)

After experiencing ten markets using this decentralized technology, a centralized matching technology was introduced for period 0 (periods -2 and -1 were organized as before). Participants who were still unmatched at period 0 would submit rank order preference lists to a centralized matching algorithm. The experimental variable was that the matching algorithm would either be the unstable priority algorithm used in Newcastle, or the stable matching algorithm used in Edinburgh.

The experimental results reproduce what we see in the field: the stable matching mechanism reverses unraveling, the unstable one does not. In addition, the experiment allows us to observe more than the data from the field. We can see not only who matches to whom, but also the pattern of offers and acceptances and rejections, which turns out to be quite revealing. In particular, the introduction of the stable matching mechanism, which reversed the unraveling, did so not by making firms unwilling to make early offers, but by making it safe for workers to decline them. This experimental observation informed and was confirmed in subsequent field and experimental studies of the market for lawyers (see Avery et al. 2001, 2007, and Haruvy et al 2006), and played a role in the subsequent design of the Gastroenterology labor market described below.

Notice how the laboratory experiments fit in Table 1's list of observations, and complement the variety of matching mechanisms observed in the field. The lab observations are by far the smallest but most controlled of the markets on the list (which otherwise range over two orders of magnitude in size, from the large American market for new doctors, which fills more than 20,000 positions a year, to the smallest British markets and American fellowship markets, some of which fill fewer than one hundred positions a year). The laboratory markets also offer the smallest incentives, far smaller than the career shaping effects of a first job.

So, by themselves, the laboratory experiments would likely not be seen as providing strong evidence that the large American medical clearinghouse needed to produce stable matchings. But, by themselves, the field observations left open the possibility that the success and failure of the various clearinghouses is unaffected by the stability of the matching mechanism, and

that the apparent connection is only coincidental. The field observations also leave open the possibility that the experience of the British markets in this regard depends in some way on the complex ways in which British medical employment differs from that in the United States.

Taken together, the field evidence plus the laboratory evidence give a much clearer picture. In the laboratory experiments, the success of the stable mechanism and the failure of the unstable mechanism can be unequivocally attributed to the difference between the two mechanisms, since, in the lab, the markets are controlled so that this is the only difference between them. The laboratory outcomes thus add weight to the hypothesis that this difference is what caused the same outcomes in the field, in Edinburgh and Newcastle, even though there are other differences between those two cities. And seeing this effect in the simple laboratory environment shows that the choice of algorithm has an effect that is not simply a function of some of the complexities of the British medical market. Together with the large body of theoretical knowledge about stable mechanisms, the laboratory experiment and field observations thus provided quite helpful guidance about how the redesign of the clearinghouse should proceed, and supported the hypothesis that stability is an important ingredient of a successful labor market clearinghouse of this kind. The current NRMP clearinghouse employs the stable Roth and Peranson (1999) algorithm.

So the experiments fit very naturally on the list of markets studied in Table 1. They are the smallest but clearest, and they illuminate and are illuminated by the similar results observed in the larger, naturally occurring markets on that list.[8]

Gastroenterology fellows

In a similar way, helping gastroenterologists redesign the labor market for new gastroenterology fellows in 2006 required a mix of field and experimental studies.

A gastroenterology fellowship is the entry-level job for the internal medicine subspecialty of gastroenterology, and doctors can take this position after they have become board certified internists by completing a three year residency in internal medicine. So, when the gastroenterology labor market started to unravel in the 1980s, gastroenterologists were already familiar with labor market clearinghouses, since they had all participated in the resident match, the NRMP. A fellowship match program was set up in 1986, but in 1996 it suddenly began to fail, and soon completely collapsed, with fellowship programs once again hiring fellows outside of the match.

[8] Other experiments illuminated some of the outlier results, e.g. regarding the single-medical-school markets at the London Hospital and Cambridge. See Unver (2001, 2005).

There was considerable disagreement about the cause of this failure, and a combination of field studies and an experiment helped clarify this (see Niederle and Roth 2003,4; and McKinney, Niederle, and Roth 2005). The field evidence consisted of one set of observations of a complex historical event leading to the failure of the clearinghouse, which was consistent with many hypotheses. These could be investigated in laboratory attempts to make a clearinghouse fail under similar circumstances. To make a long story short, part of what happened in 1996 is that there was an announced and widely anticipated reduction in the number of fellowship positions (together with an increase from two to three years needed to become a board certified gastroenterologist). This reduction in the number of positions was accompanied by an even larger, and unexpected reduction in the number of doctors applying for those positions. As it happened, despite the reduction in the number of positions, 1996 turned out to be the first year in which the number of positions exceeded the number of applicants. It now appears that the collapse of the clearinghouse began when fellowship programs (alarmed by the smaller than expected number of applicants they received) made early offers to applicants, who accepted them without waiting for a match.

Of course, there are other ways the historical story could be parsed. But McKinney, Niederle and Roth (2005) found in the laboratory that anticipated shifts in supply and demand, visible to both sides of the market, did not cause declines in match participation anywhere near the magnitude caused by unanticipated shocks, particularly when these are more visible to one side of the market than to the other.[9] In particular, we looked at shifts in demand that were either visible to both firms and workers, or only to firms (as when an unexpected change in demand is visible to firms who receive few applications, but not to workers). Demand reductions of both kinds caused firms to try to make more early hires, but when workers knew that they were on the short side of the market they were more likely to decline such offers than when they were unaware of the shift in demand. In the lab it was clearly the combination of firms making early offers outside of the match, and workers not feeling safe to reject them and wait for the match that caused the market to unravel. The experimental results also clearly suggested that, after such a shock, it would be possible to re-establish a functioning match.

This experiment, like that of Kagel and Roth (2000), also suggested that when there was not much participation in the match there would be pressure for the market to unravel, with participants making offers earlier and earlier. But this is an observation that is clearly built into the experimental design, almost as an assumption, since in the experiment, early offers were

[9] Subjects in the roles of workers and firms first participated in 15 3-period decentralized markets with a congested (1-offer per period) match technology and a cost for matching early, then in 15 markets with the same number of firms and workers in which a centralized clearinghouse was available to those who remained unmatched until the last period, then in 15 further markets in which a change was made either in the number of firms or workers, a change that was observable to firms but only observable to workers in some treatments.

one of very few strategic options available. So the experiment by itself didn't provide much evidence that unraveling was going on in the gastroenterology market. Establishing that depended on field data, both from employer surveys and analysis of employment data, which showed that, ten years after the collapse of the match the market continued to unravel, with employers making exploding offers earlier each year than the previous year, not all at the same time, and months ahead of the former match date (Niederle, Proctor, and Roth, 2006). This also had the consequence of causing a formerly national market to have contracted into much more local, regional markets.

Taken together, the field and experimental evidence made what proved to be a convincing case that the absence of a match was harmful to the market, and that the collapse following the events of 1996 had been due to a particular set of shocks that did not preclude the successful operation of a clearinghouse once more.

But a problem remained before a clearinghouse could be restarted. The employers were accustomed to making early exploding offers, and program directors who wished to participate in the match worried that if their competitors made early offers, then applicants would lose confidence that the match would work and consequently would accept those early offers, because that had been the practice in the decentralized market. That is, in the first year of a match, applicants might not yet feel that it is safe to reject an early offer to wait for the match. Program directors who worried about their competitors might thus be more inclined to make early, pre-match offers themselves.

There are decentralized markets that have avoided the problem of early exploding offers, in ways that seemed to suggest policies that might be adopted by the Gastroenterology professional organizations. One example is the market for Ph.D. students, in which a policy of the Council of Graduate Schools (adopted by the large majority of universities) states that offers of admission and financial support to graduate students should remain open until April 15. Specifically, the policy states in part:

> "Students are under no obligation to respond to offers of financial support prior to April 15; earlier deadlines for acceptance of such offers violate the intent of this Resolution. In those instances in which a student accepts an offer before April 15, and subsequently desires to withdraw that acceptance, the student may submit in writing a resignation of the appointment at any time through April 15."

This of course makes early exploding offers much less profitable. A program that might be inclined to insist on an against-the-rules early response is discouraged from doing so in two ways. First, the chance of actually enrolling a student who is pressured in this way is

diminished, because the student is not prevented from later receiving and accepting a more preferred offer. Second, a program that has pressured a student to accept an early offer cannot offer that position to another student until after the early acceptance has been declined, at which point most of the students in the market may have made binding agreements. In the market for new Ph.D. students, this policy has helped to make early exploding offers a non-issue.

But gastroenterologists were quick to point out that there are many differences between gastroenterology fellowships for board certified internists and graduate admissions for aspiring PhDs. Perhaps the effectiveness of the CGS policy depended in some subtle way on the many and complex differences between these two markets. So experiments still had another role to play before a marketplace could be built that would reverse the previous decade of unraveling. And here the role of experiments was (once again) to help bridge the gap, in the laboratory, between two rather different markets, the gastroenterology market, and the market for admissions of Ph.D. students to graduate programs. (Recall our earlier discussion of the differences between British and American markets for new doctors.)

Niederle and Roth (2009) bridged this gap by studying in a simple laboratory environment the effect of the CGS policy of empowering students to accept offers made before a certain time and then change their minds if they received offers they preferred.[10] In the lab, early inefficient matches that were common when subjects could not change their minds about early offers and acceptances essentially disappeared when the policy allowing changes of mind was in place. And the rise in efficiency came about not because early offers were made,

[10] Each market involved 5 firms and 6 applicants, and consisted of 9 periods in which firms could make offers. (So this decentralized market was not congested, i.e. there was enough time for all offers to be made.) Firms and applicants had qualities, and the payoff to a matched firm and applicant was the product of their qualities. Firms' qualities (1,2,3,4, and 5) were common knowledge, but applicants' qualities were stochastically determined over time: In periods 1, 4 and 7 each applicant received an integer signal from 1 to 10 (uniform iid). The quality of each applicant was determined in period 7 through the relative ranking of the sum of their three signals: The applicant with the highest sum had a quality of 6, the second highest a quality of 5, the lowest a quality of 1 (ties were broken randomly). So efficient matches (which assortatively match the applicants and firms in quality order) can only be made if matching is delayed until applicants' qualities have been determined by the final signal in period 7. But lower quality firms have an incentive to try to make matches earlier, since this gives them their only chance at matching to higher quality workers.

accepted, and then subsequently rejected, but rather because this possibility discouraged early offers from being made. The fact that this could be observed in the transparently simple laboratory environment showed that the policy did not depend for its effectiveness on some subtle feature of the complex Ph.D. admissions process.

The four gastroenterology organizations adopted the policy, as proposed in Niederle, Proctor and Roth (2006), and the gastroenterology match for 2007 fellows was held June 21, 2006. It succeeded in attracting 121 of the 154 eligible fellowship programs (79%). 98% of the positions offered in the match were filled through the match. Niederle, Proctor and Roth (2008) show that in the second year of the new centralized match the interview dates were successfully pushed back and are now comparable to those of other internal medicine specialties that have used a centralized match for many years.

To summarize the role of experiments in practical market design, simple experiments can help us understand complex markets. They fill a gap left even by field studies of similar markets, since comparison between one complicated market and another is often quite properly viewed with suspicion; e.g. the market for new American doctors is indeed very different from the comparable markets for British doctors, and the market for gastroenterology fellows is very different from the market for Ph.D. students.

Of course, experiments are also very different from complex markets like those for doctors, but they are different in simple, transparent ways. As such, they can sometimes have more *ecological validity* than observations from natural markets that are equally as complex as, but different from, a particular market. By the same token, evidence about complex markets drawn entirely from simple experiments is also likely to be less convincing than a wide array of observations from markets of different transparency and complexity.

Experimental economics has thus become more effective and important for practical market design as it has become less 'heroic' and stand-alone. Experiments (in the lab or in the field) seem to have the greatest effect when they are used together with other kinds of investigations, both observational and theoretical.

How are we doing at generating productive new areas of research?

One question that seems natural for diagnosing the current state of experimental economics is whether it is continuing to generate new areas of investigation. Certainly the field has been fertile in the last few decades; for example, experiments are at the heart of the great growth in

study of what is often called behavioral economics, but might better be described as Economics and Psychology (E&P). This line of work includes not only the venerable study of failures of utility theory and of heuristics and biases, in the style of Kahneman and Tversky, but also some very new components, such as the emerging "neuroeconomics" conversation between experimental economists and neuroscientists.[11]

Loosely speaking, the main project in E&P has been to better understand individuals' thought processes as they make choices. The attraction of neuroeconomics, with tools like fMRI scanners, is to find real time correlates of these choice processes in the brain. Time will tell how productive this kind of research will be for economists' agenda (a subject on which there has been much discussion, to which I will return briefly in the next section).

But I take it as a sign of the healthy state of experimental economics that there are other, related but different research programs underway, that seem to be growing quickly. One of them is the emerging study of what might be called Economics and Biology (E&B). It too has historical antecedents[12], and, like E&P (which purists might want to argue is a subset of E&B), it has many parts, including concerns with nutrition and disease. But for comparison purposes, it may be helpful to focus on a topic shared by E&P and E&B, namely individual choices and preferences. One difference between the two approaches is that E&P has a strong emphasis on thought processes *while* making a choice, while E&B directs our attention to some of *the longer term determinants* of individuals' choices, actions, and preferences.

At the level of the whole human organism are studies of gender, and how men and women may behave differently in economic environments. For example, the experimental studies of Gneezy, Niederle and Rustichini (2003) and Niederle and Vesterlund (2007) address the question of whether men and women behave differently in competitive environments, and whether they make different choices about whether to engage in competition.

Closer to the level of brain chemistry, a small literature is growing up around "endocrinological economics," having to do with the effect of hormones on behavior, including for example studies of risk taking behavior and the menstrual cycle, but also studies of how exposure to

[11] New opportunities for experiments come about both as new subject matter is considered, and as new possibilities for running experiments arise. In just this way, neuroeconomics reflects both interest in the brain, and the increasing availability of tools for measuring aspects of brain activity. Other new possibilities for running experiments are likely to be increasingly exploited in the future. The ubiquity of the internet will surely lead to increasing numbers of experiments using it, not merely on websites set up by experimenters, but also on sites set up for other purposes, like sales, social networking, labor market matching, dating, advice, etc.

[12] For example, William Stanley Jevons, better remembered today for his 1876 Money and the Mechanism of Exchange, reported a set of three experiments about efficient weight lifting (and throwing) in an 1870 article in *Nature* (which I first learned of in Bardsley et al.). At that time, of course, a lot of economic activity was muscle powered.

testosterone in the womb (e.g. as measured by the ratio of the length of the second and fourth fingers) is correlated with risk taking behavior of adult men and women.[13] A related line of work measures the *heritability* of preferences such as risk taking propensities, e.g. through studies of twins (see e.g. Cesarini et al. 2009).

It will likely be quite some time before the impact of this developing literature on economics, or on psychology and biology, can be assessed. But the growing collaboration of economists with other kinds of scientists (such as neuroscientists and biologists generally) interested in using economic experiments to gain insight into aspects of human biology is a sign of how useful and flexible the experimental economics laboratory is. And the fact that experimental economics is full of new research directions, of which this is just one example, is an indication of the thriving life of experimentation within economics.

Critiques and criticisms of experiments (from within and without):

Perhaps one of the clearest signals that experimental economics is coming of age is that we have recently seen a wide variety of commentaries, criticisms and critiques, both from those who do experiments and recommend them, and from those who don't and don't; as well as some crossovers, and comments from those who favor one style of experimental research over another.[14] Only recently has experimental economics become so widely perceived as successful that it can be in a position to experience any sort of *backlash*.[15] Experimental economics is even approaching the decadent phase in which it becomes the subject of philosophy of science; see

[13] One indicator of the vibrant nature of the experimental community is that in August 2009 Burkhard Schipper sent out a request on the ESA-discuss chat group for papers on endocrinology and economics, and later in the month circulated the bibliography of papers he had received, consisting of about thirty economics papers written since 2002, with titles containing "testosterone", "oxytocin", "dopamine", "second-to-fourth digit ratio"…(see Apicella et al. 2008, Baumgartner et al. 2008, Burnham 2007, Chaplin and Dean 2008, Chen et al. 2009, Coates et al. 2009 and Pearson and Schipper 2009 for a sampling of recent papers whose titles are evocative of this area.)

[14] Recall from Wallis and Friedman (1942) that methodological critiques of experiments in economics aren't entirely new. But when there wasn't much experimental economics, there wasn't much to analyze, criticize, or debate. In the 1990's there was some internal discussion of methods and goals. See e.g. Roth (1994) on what was then sometimes the practice of labeling as an experiment each session of a larger experimental design, with some resulting irregularities in how experiments were reported. Today the practice of reporting whole experimental designs seems to be much more widely followed. See also some discussion of the differing emphases and interpretations that arose from attempts to distinguish between 'experimental' and 'behavioral' economics, e.g. in Binmore (1999) and Loewenstein (1999).

[15] Not so long ago, there was hardly any criticism of experimental economics, there were just statements to the effect that "economics is not an experimental science." (If you google that phrase, you can still find some examples, about half of them with the original meaning, and the other half by experimenters reflecting on the change in the views in the profession.)

e.g. Guala, 2005 and Bardley et al. 2010, not to mention conferences and volumes on its methods and vital signs.[16]

The tone of the criticisms and critiques ranges from temperate to hot, from thoughtful to polemical, sometimes in the same article. Like articles reporting experiments, articles criticizing or praising them can suffer if the conclusions are too broad, even if they also contain valid points. Nevertheless, there is always the potential opportunity to learn something about our craft, and its place in economics, by paying some attention to the criticisms of experiments, and the arguments made in their defense.

Most of the critiques touch on both a methodological theme and a substantive one. See for example the discussion of neuroeconomics by Bernheim (2009), Gul and Pesendorfer (2009), Rustichini (2009) and Sobel (2009), or the discussions of inequity aversion and experimental methods and reporting by Binmore and Shaked (2010a,b), Fehr and Schmidt (2010), and Eckel and Gintis (2010).

One of the most broadly controversial lines of criticism among experimenters has to do with the comparative advantages of experiments conducted in the laboratory and in the field. The papers by Levitt and List (2007a,b) are widely read as proposing that laboratory experiments are in some important senses simply inferior to field experiments, particularly regarding their ecological validity, i.e. the ability of their conclusions to generalize to naturally occurring target markets. In my remarks above on market design I've already made clear that I don't agree, although I certainly think that studies in the lab and in the field (whether experimental or not) can complement each other.

It's worth noting in this regard that field experiments have come in for some criticism of the same sort from applied econometricians, see e.g. the recent papers by Deaton (2009) and Heckman and Urzua (2009), and the reply by Imbens (2009), which discuss the extent to which the local average treatment effects that are revealed by randomized field experiments can be

[16] Some of the recent philosophy of economics (as it is called when addressed to philosophers of science, and *methodology* when addressed also to economists) has been written by distinguished experimenters; see e.g. the special issue of the *Journal of Economic Behavior and Organization* "On the Methodology of Experimental Economics (Rosser and Eckel (2010) beginning with a target article by Smith (2010) followed by many replies and a final article by Croson and Gächter (2010). See also Bardley et al. (2010), and Starmer (1999), but also philosophical investigations of the uses of theoretical models that resonate with the use of experiments as models, such as Sugden (2009). In this connection see also Mäki (2005). (I write as a frequently disappointed consumer of philosophy of science who nevertheless returns sporadically in the hope that it will teach me how to do science better. I am often reminded of the quip attributed to the late Richard Feynman that philosophy of science is as useful to scientists as ornithology is to birds. Needless to say, this may just mean that birds are misguided if they look to ornithology for advice; ornithology is interesting in its own right, even if not so useful to practicing birds. In this respect, it is another sign of the progress of experimental economics that it now commands some attention from those who study what economists do.)

appropriately generalized from one domain to another.[17] In this connection, see also Falk and Heckman (2009), whose paper's title conveys its central message "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." Their concluding paragraph is one with which I agree:

> "Causal knowledge requires controlled variation. In recent years, social scientists have hotly debated which form of controlled variation is most informative. This discussion is fruitful and will continue. In this context it is important to acknowledge that empirical methods and data sources are complements, not substitutes. Field data, survey data, and experiments, both lab and field, as well as standard econometric methods can all improve the state of knowledge in the social sciences. There is no hierarchy among these methods and the issue of generalizability of results is universal to all of them."

I suspect that the reason they found it necessary to argue in favor of what should be such an uncontroversial conclusion is that these methodological debates have been conducted with much more heat than are the usual scientific disagreements. They sometimes have the flavor of accusation, as if those who disagree must be bad scientists who ignore critical evidence. I'm puzzled by this, but I think it may have something to do with the general problem of drawing conclusions from evidence, in a way that is made particularly stark by experimental evidence.

I've elaborated on this for many years in my experimental economics classes with the following dramatization designed to show how different people can view the same evidence differently.[18]

> *Suppose I show you what appears to be a deck of playing cards, all face down, and propose that we investigate the hypothesis that none of the cards is blue, by turning them face up one at a time. The first 20 cards are all red and black, Hearts and Clubs and Spades and Diamonds. After each one is turned up, we both agree that our confidence that none of the cards is blue has increased. The 21st card is the four of Forest, which is, of course, green. I argue that since yet another card that isn't blue has been observed, my confidence that none of them are blue has increased. But you argue that if the deck contains the four of Forest, you suddenly realize it might contain the six of Sky, which is, of course, blue. So, on the basis of the same evidence, and even though none of the cards turned over so far is blue, your confidence in the no-blue hypothesis is diminished (because the hypothesis you really believed going in was the unstated one that we were dealing with a standard deck of playing cards, i.e. that there were only red and black cards).*

[17] See also Banerjee and Duflo (2008), and Cohen and Easterly (2009).

[18] I suspect that something like this may be an old philosophy of science chestnut, but correspondence with philosophers who specialize in "confirmation theory" and related matters hasn't turned up a source, although it has turned up a number of related examples illustrating the difficulty of interpreting evidence.

The fact that we disagree based on the same evidence might be enough to disrupt our future card games, especially if we draw our conclusions with unjustified confidence or generality. But, fortunately, there is a way forward towards resolving such disagreements, which is to gather more evidence, including designing and conducting more experiments. So the recent methodological disputes will probably be good for business in the longer term.

In conclusion

Experimental economics is thriving. It is living up to its promise, although we are also learning, by doing experiments, more about what experiments promise for economics.

Experiments are becoming better integrated with other kinds of economic research, including a vigorous conversation with theorists that motivates new theories as well as tests existing ones. We are learning more about how experiments are complements to other kinds of empirical investigation. The lab has unique advantages, as does field data, and theoretical models. (Just as modern warfare cannot be won by air power alone, but no commander would voluntarily give up his air force, experiments are often not enough on their own to carry the day, but no science should try to do without them.) Experiments are essential when control is needed.

Let's celebrate our successes, and learn from our experience, but not spend too much time looking backwards and inwards. There's good work to be done.

Bibliography

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.

Apicella, C.L., Dreber, A., Campbell, B., Gray, P.B., Hoffman, M., and Little, A.C. (2008). Testosterone and financial risk preferences, *Evolution and Human Behavior* 29, 384-390.

Avery, Christopher, Christine Jolls, Richard A. Posner, and Alvin E. Roth, "The Market for Federal Judicial Law Clerks", *University of Chicago Law Review*, 68, 3, Summer, 2001, 793-902.

Avery, Christopher, Christine Jolls, Richard A. Posner, and Alvin E. Roth, "The New Market for Federal Judicial Law Clerks", *University of Chicago Law Review*, 74, Spring 2007, 447-486.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U. and Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans, *Neuron* 58, 639-650.

Banerjee, Abhijit V. and Esther Duflo (2008), "The Experimental Approach to Development Economics," November, NBER working paper No. 14467.

Bardsley, Nicholas, Robin Cubitt, Graham Loomes, Peter Moffatt, Chris Starmer and Robert Sugden, *Experimental Economics: Rethinking the Rules*, Princeton, 2010

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description based decisions. *Journal of Behavioral Decision Making, 16,* 215-233.

Bernoulli, Daniel. 1738. "Specimen Theoriae Novae de Mensura Sortis," Commentarii Academiae Scientiarum Imperialis Petropolitanae, 5, 175-192. English translation in 1954 *Econometrica*, 22, 23-36.

Binmore, Ken (1999) "Why Experiment in Economics?," Economic Journal, 109, 453, Features, Feb., , F16-F24.

Binmore, Ken and Avner Shaked (2010a), "Experimental Economics: Where Next?" *Journal of Economic Behavior and Organization*, 73, 1(January), 87-100.

Binmore, Ken and Avner Shaked (2010b), "Experimental Economics: Where Next? Rejoinder" *Journal of Economic Behavior and Organization*, 73, 1(January), 120-121.

Binmore, Ken, Avner Shaked, and John Sutton [1985], "Testing Noncooperative Bargaining Theory: A Preliminary Study," *American Economic Review*, vol. 75, pp1178-1180.

Binmore, Ken, Avner Shaked, and John Sutton [1988], "A Further Test of Noncooperative Bargaining Theory: Reply," *American Economic Review*, 78, 837-839.

Bolton, Gary [1991], "A Comparative Model of Bargaining: Theory and Evidence," *American Economic Review*, December, 81, 1096-1136.

Bolton, Gary E. and Axel Ockenfels [2000], "ERC: A Theory of Equity, Reciprocity and Competition, *American Economic Review*, 90, 1, March, 166-193.

Brunner, Christoph, Jacob K. Goeree, Charles A. Holt, and John O. Ledyard, "An Experimental Test of Flexible Combinatorial Spectrum Auction Formats" AEJ: Microeconomics, forthcoming.

Burnham, T.C. (2007). High-testosterone men reject low ultimatum game offers. *Proceedings of the Royal Society* (B) 274, 2327–2330.

Camerer, Colin (1995), "Individual Decision Making,"in *Handbook of Experimental Economics*, J.H. Kagel and A.E. Roth, editors, Princeton University Press, 587-703.

Cesarini, David, Christopher T. Dawes, Magnus Johannesson, Paul Lichtenstein, and Bjorn Wallace (2009), "Genetic Variation in Preferences for Giving and Risk-Taking," *Quarterly Journal of Economics*, 124, 809–842.

Chaplin, A., and Dean, M. (2008). Dopamine, reward prediction error, and economics, *Quarterly Journal of Economics* 123, 663-701.

Chen Y., Katuscak, P. and Ozdenoren, E. (2009). Why can't a woman bid more like a man?, mimeo., University of Michigan.

Coates, J.M., Gurnell, M., and Rustichini, A. (2009). Second-to-fourth digit ratio predict success among high-frequency financial traders, *Proceedings of the National Academy of Sciences* 106, 623-628.

Cohen, Jessica and William Easterly (editors, 2009) *What Works in Development?: Thinking Big and Thinking Small*, Brookings Institution Press.

Croson, Rachel and Simon Gächter (2010). “The science of experimental economics,” *Journal of Economic Behavior and Organization*, 73, 1 (January), 122-131.

Cybernomics, Inc. (2000), “An Experimental Comparison of the Simultaneous Multi-Round Auction and the CRA Combinatorial Auction,” Submitted to the Federal Communications Commission, http://wireless.fcc.gov/auctions/conferences/combin2000/releases/98540191.pdf

Deaton, Angus (2009), “Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development,” NBER Working Paper No. 14690.

Dietz, T, Elinor Ostrom, and P.C. Stern (2003) “The struggle to govern the commons,” *Science*, 302, 5652 (Dec 12), 1907-1912.

Eckel, Catherine and Herbert Gintis (2010), “Blaming the Messenger: Notes on the Current State of Experimental Economics,” *Journal of Economic Behavior and Organization*, 73, 1 (January), 109-119.

Ellsberg, Daniel. 1961. "Risk, Ambiguity and the Savage Axioms," Quarterly Journal of Economics, 75, 643-669.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, *112*, 912-931.

Ert, Eyal, and Ido Erev, (2009). “On the descriptive value of loss aversion in decisions under risk,” working paper.

Falk, Armin and James J. Heckman (2009), “Lab Experiments Are a Major Source of Knowledge in the Social Sciences,” *Science*, 326, 23 October, 535-8.

Fehr, Ernst and Klaus M. Schmidt (1999), “A Theory Of Fairness, Competition, and Cooperation,” *Quarterly Journal of Economics*, 114, 3 (August), 817-868

Fehr, Ernst and Klaus M. Schmidt (2010), “On Inequity Aversion: A Reply to Binmore and Shaked,” *Journal of Economic Behavior and Organization*, 73, 1 (January), 101-108.

Flood, Merrill M. 1952. "Some Experimental Games," Research Memorandum RM-789, RAND Corporation, June.

Flood, Merrill M. 1958. "Some Experimental Games," *Management Science*, 5, 5-26.

Gneezy, Uri, Muriel Niederle, Aldo Rustichini, “Performance in Competitive Environments: Gender Differences”, *Quarterly Journal of Economics*, CXVIII, August 2003, 1049 – 1074.

Grether, David M., R. Mark Isaac, and Charles R. Plott (1979), *Alternative Methods of Allocating Airport Slots: Performance and Evaluation*, Prepared for Civil Aeronautics Board Contract Number 79-C-73, Polinomics Research Laboraories, Inc., Pasadena, CA, August.

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

Guth, Werner, Schmittberger, R and Schwarz, B [1982], "An experimental analysis of ultimatum bargaining," *Journal of Economic Behavior and Organization*, 3, 367-88.

Harbaugh, William T., Kate Krause, and Lise Vesterlund, “The Fourfold Pattern of Risk Attitudes in Choice and Pricing Tasks,” Economic Journal, forthcoming.

Haruvy, Ernan, Alvin E. Roth, and M. Utku Ünver, “The Dynamics of Law Clerk Matching: An Experimental and Computational Investigation of Proposals for Reform of the Market,” *Journal of Economic Dynamics and Control*, 30, 3 , March 2006, Pages 457-486.

Heckman, J., and S. Urzua., (2009), “Comparing IV With Structural Models: What Simple IV Can and Cannot Identify,”NBER Working Paper No. 14706

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*, 534-39.

Imbens, Guido W. (2009), “Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009),” April, NBER Working Paper No. 14896.

Jevons, W. Stanley, “On the natural laws of muscular exertion,” *Nature*, 1870, June 30,158-60

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263-291.

Kagel, John H., Yuanchuan Lien, and Paul Milgrom (2009), “Ascending Prices and Package Bidding: A Theoretical and Experimental Analysis,” working paper, April.

Kagel, John H. and Alvin E. Roth, "The dynamics of reorganization in matching markets: A laboratory experiment motivated by a natural experiment," *Quarterly Journal of Economics*, February, 2000, 201-235.

Ledyard, John O., David Porter, and Antonio Rangel (1997), “Experiments Testing Multiobject Allocation Mechanisms,” *Journal of Economics & Management Strategy*, 6, 3 (Fall) 639-675.

Levitt, Steven D. and John A. List (2007a), “What do Laboratory Experiments Measuring Social Preferences tell us about the Real World,” *Journal of Economic Perspectives*, 21 (2): 153-174.

Levitt, Steven D. and John A. List (2007b), “Viewpoint: On the generalizability of lab behaviour to the field,” Canadian Journal of Economics, 40(2), pp. 347-370.

Loewenstein, George (1999) "Experimental Economics from the Vantage-Point of Behavioural Economics," *Economic Journal*, 109, 453, Features, February, F25-F34.

Mäki, Uskali (2005), “Models are experiments, experiments are models,” *Journal of Economic Methodology*, 12, 2, June, 303-315.

May, Kenneth O. 1954. "Intransitivity, Utility, and the Aggregation of Preference Patterns," *Econometrica*, 22, 1- 13.

McKinney, C. Nicholas, Muriel Niederle, and Alvin E. Roth, “The collapse of a medical labor clearinghouse (and why such failures are rare),” *American Economic Review*, 95, 3, June, 2005, 878-889.

Milgrom, Paul (2007), “Package Auctions and Exchanges,” Fisher-Schulz Lecture, *Econometrica*, 75, 4 (July), 935-965

Nash, John (1950), “The bargaining problem,” *Econometrica* 28, 155-62.

Neelin, Janet, Hugo Sonnenschein, and Matthew Spiegel [1988], "A Further Test of Noncooperative Bargaining Theory: Comment," *American Economic Review*, 78, pp824-836.

Niederle, Muriel, Deborah D. Proctor and Alvin E. Roth. 2006. "What will be needed for the new GI fellowship match to succeed?" *Gastroenterology*, 130, 218-224.

Niederle, Muriel and Alvin E. Roth, “Unraveling reduces mobility in a labor market: Gastroenterology with and without a centralized match,” *Journal of Political Economy*, 111, 6, December 2003, 1342-1352.

Niederle, Muriel and Alvin E. Roth, “The Gastroenterology Fellowship Match: How it failed, and why it could succeed once again,” *Gastroenterology*, 127, 2 August 2004, 658-666.

Niederle, Muriel, and Alvin E. Roth, "Market Culture: How Rules Governing Exploding Offers Affect Market Performance," *American Economic Journal: Microeconomics*, 1, 2, August 2009, 199-219.

Niederle, Muriel and Alvin E. Roth (2010), "The Effects of a Central Clearinghouse on Job placement, Wages, and Hiring Practices", in *Labor Market Intermediation*, David Autor, Editor, The University of Chicago Press, forthcoming.

Niederle, Muriel, and Lise Vesterlund, "Do Women Shy away from Competition? Do Men Compete too Much?," *Quarterly Journal of Economics*, August 2007, Vol. 122, No. 3: 1067-1101.

Ochs, J. and Roth, A.E., "An Experimental Study of Sequential Bargaining," *American Economic Review*, 79, 1989, 355-384.

Ostrom, Elinor (1998), "A behavioral approach to the rational choice theory of collective action," *American Political Science Review*, 92, 1 (March), 1-22

Pearson, Matthew and Burkhard C. Schipper (2009), "The Visible Hand: Finger Ratio (2D:4D) and Competitive Behavior," working paper UC Davis.

Plott, Charles R. (1987), "Dimensions of parallelism: some policy applications of experimental methods," in *Laboratory Experimentation in Economics: Six Points of View*, A.E. Roth, editor, Cambridge University Press, 193-219.

Plott, Charles R. (1997), "Laboratory Experimental Testbeds: Application to the PCS Auction," *Journal of Economics & Management Strategy*, 6,3 (Fall), 605-638.

Rapoport, Anatol, O. Frenkel, and J. Perner (1977), "Experiments with cooperative 2x2 games," *Theory and Decision*, 8, 67-92.

Rassenti, Stephen J., Vernon L. Smith, and Robert L. Bulfin (1982), "A combinatorial Auction Mechanism for Airport Time Slot Allocation," *Bell Journal of Economics*, 13, 2, Autumn, 402-417.

Rosser, J. Barkley Jr., and Catherine Eckel (2010), "Introduction to JEBO special issue on 'Issues in the Methodology of Experimental Economics'," *Journal of Economic Behavior and Organization*, Special Issue On the Methodology, 73, 1 (January), 1-2.

Roth, A.E. *Axiomatic Models of Bargaining*, Lecture Notes in Economics and Mathematical Systems #170, Springer Verlag, 1979. (http://kuznets.fas.harvard.edu/~aroth/Axiomatic_Models_of_Bargaining.pdf)

Roth, A.E. "Laboratory Experimentation in Economics," *Advances in Economic Theory, Fifth World Congress*, Truman Bewley, editor, Cambridge University Press, 269-299, 1987. (Preprinted in *Economics and Philosophy*, Vol. 2, 1986, 245-273.)

Roth, A.E. "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory", *Journal of Political Economy,* Vol. 92, 1984, 991-1016.

Roth, A.E. "On the Allocation of Residents to Rural Hospitals: A General Property of Two-Sided Matching Markets," *Econometrica*, Vol. 54, 1986, 425-427.

Roth, A.E. "Laboratory Experimentation in Economics: A Methodological Overview," *Economic Journal*, Vol. 98, 1988, 974-1031.

Roth, A.E. "New Physicians: A Natural Experiment in Market Organization," *Science*, 250, 1990, 1524-1528.

Roth, A.E. "A Natural Experiment in the Organization of Entry Level Labor Markets: Regional Markets for New Physicians and Surgeons in the U.K.", *American Economic Review*, vol. 81, June 1991, 415-440.

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

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.

Roth, A.E. "Introduction to Experimental Economics," *Handbook of Experimental Economics*, John Kagel and Alvin E. Roth, editors, Princeton University Press, 1995a, 3-109.

Roth, A.E. "Bargaining Experiments," Handbook of Experimental Economics, John Kagel and Alvin E. Roth, editors, Princeton University Press, 1995b, 253-348.

Roth, Alvin E. “The Economist as Engineer: Game Theory, Experimental Economics and Computation as Tools of Design Economics,” *Econometrica*, 70, 4, July 2002, 1341-1378.

Roth, Alvin E. "Deferred Acceptance Algorithms: History, Theory, Practice, and Open Questions," *International Journal of Game Theory*, Special Issue in Honor of David Gale on his 85th birthday, 36, March, 2008a, 537-569.

Roth, Alvin E. "What have we learned from market design?" *Economic Journal*, 118 (March), 2008b, 285–310.

Roth, Alvin E. “Market Design,” Handbook of Experimental Economics, volume 2, John Kagel and Alvin E. Roth, editors, Princeton University Press, forthcoming.

Roth, A.E. and I. Erev "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," *Games and Economic Behavior*, Special Issue: Nobel Symposium, vol. 8, January 1995, 164-212.

Roth, A.E. and Malouf, M.K. "Game-Theoretic Models and the Role of Information in Bargaining", *Psychological Review*, Vol. 86, 1979, 574-594.

Roth, A.E. and Murnighan, J.K. "The Role of Information in Bargaining: An Experimental Study," *Econometrica*, Vol. 50, 1982, 1123-1142.

Roth, A.E. and E. Peranson, "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design,” *American Economic Review*, 89, 4, September, 1999, 748-780.

Roth, A.E., Prasnikar, V., Okuno-Fujiwara, M., and Zamir, S. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review*, vol. 81, December 1991, 1068-1095.

Roth, A.E. and Schoumaker, F. "Expectations and Reputations in Bargaining: An Experimental Study", *American Economic Review*, Vol. 73, 1983, 362-372.

Roth, A.E. and M. Sotomayor *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Econometric Society Monograph Series, Cambridge University Press, 1990.

Rubinstein, Ariel [1982], "Perfect Equilibrium in a Bargaining Model," *Econometrica*, vol. 50, pp97-109.

Sauermann, Heinz and Reinhard Selten. 1959. "Ein Oligolpolexperiment," *Zeitschrift fur die Gesamte Staatswissenschaft*, 115, 427-471.

Schelling, Thomas C. 1957. "Bargaining, Communication, and Limited War," Journal of Conflict Resolution, 1, 19-36.

Schelling, Thomas C. 1958. "The Strategy of Conflict: Prospectus for a Reorientation of Game Theory," Journal of Conflict Resolution, 2, 203-264.

Schelling, Thomas C. 1960. The Strategy of Conflict, Harvard University Press, Cambridge.

Selten, Reinhard and Thorsten Chmura (2008), “Stationary concepts for experimental 2x2 games,” *American Economic Review*, 98, 3 (June), 938-966.

Smith, Vernon L. (2008) *Rationality in Economics: Constructivist and Ecological Forms*, Cambridge University Press.

Smith, Vernon L. (2010), “Theory and Experiment: What are the questions?,” *Journal of Economic Behavior and Organization*, Special Issue On the Methodology, 73, 1 (January), 3-15.

Starmer, Chris (1999) “Experiments in economics: should we trust the dismal scientists in white coats?,” *Journal of Economic Methodology* 6: 1–30.

Sugden, Robert (2009), “Credible Worlds, Capacities and Mechanisms,” Erkenntnis, 70, 1, 3-27

Thurstone, L.L. 1931. "The Indifference Function," *Journal of Social Psychology*, 2, 139-167.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297-323.

Ünver, M.Utku (2005) "On the Survival of Some Unstable Two-Sided Matching Mechanisms." *International Journal of Game Theory* 33: 239-254.

Ünver, M.Utku (2001) "Backward Unraveling over Time: The Evolution of Strategic Behavior in the Entry-Level British Medical Labor Markets." *Journal of Economic Dynamics and Control* 25: 1039-1080.

Wallis, W. Allen and Milton Friedman. 1942. "The Empirical Derivation of Indifference Functions," in Studies in Mathematical Economics and Econometrics in memory of Henry Schultz, O. Lange, F. McIntyre, and T.O. Yntema, editors, Chicago, University of Chicago Press, 175-189.