Experiments in Market Design

Alvin E. Roth
June 9, 2014

Design: **Noun**: the arrangement of elements or details

Verb: to create or construct

1. Introduction

The phrase “market design” has come to include the design not only of marketplaces but also of other economic environments, institutions and allocation rules. And it includes not only the design of new institutions ("design" as a verb) but also renewed attention to how the design of economic institutions ("design" as a noun) influences their performance. It is both one of the oldest and one of the newest areas of experimental economics.

It is one of the oldest because every economic experiment involves the design of an economic environment, and many experiments compare the effects of different designs. And it is one of the newest because only since the 1990’s have economists become regularly involved in the detailed design of marketplaces and other economic institutions in ways that have led from the initial conception all the way to the adoption and implementation of practical new designs (and to the beginnings of a new scientific literature of market design). This new usefulness of market design has brought new uses for experiments.

To see that market design has always played a role in experimental economics, note that when Chamberlain (1948) sought to investigate competitive equilibrium, he designed not only the kind of marketplace (pairwise negotiation) that he wished to investigate; he also developed the technique that has since been widely used to induce particular supply and demand conditions, by giving each buyer and seller the prices and quantities at which they could in effect sell to or buy from the experimenter to fulfill any trades they made.[1]

[1] Thus e.g. a participant in an experimental market might be told that he could sell one unit of some good in the market, which would cost him $20 (to be subtracted from his sale price), while another might be told

That is, he not only designed marketplace rules, he designed a whole experimental market, complete with preferences of buyers and sellers. When Dresher and Flood proposed in 1950 to test Nash equilibrium in a challenging environment, they designed the underlying game (the Prisoner's dilemma), the environment in which subjects would encounter it (repeated play against a fixed other player), and the payoffs that would motivate the players. When John Nash subsequently proposed that a different (non-repeated) environment might produce different behavior, he was making a conjecture about how the design of the economic environment (repeated or non-repeated, see Flood, 1952, 1958) would influence the behavior of the participants. Similarly, when Vernon Smith (1962) proposed that competitive equilibrium would be reached more easily in a repeated double auction than in Chamberlain's non-repeated pairwise negotiations, he was investigating how elements of a market's design influenced its performance. Many subsequent experiments in these and other lines of investigation have since reported careful within-experiment comparisons focused on such issues of design, and on the more precise hypotheses that arose from series of experiments that built on one another. The first volume of this Handbook reported many series of experiments that resulted from this kind of conversation among experimenters, and between experimenters and theorists.[2]

In the practical market design efforts that will be the main focus of this chapter, there are still conversations among experimenters, and between theorists and experimenters, but those are only parts of a larger conversation. Often the need for a newly designed marketplace is sparked by a market failure, or a new law or regulation, and a new design will require coordination among many parties. So the conversation is conducted among economists, market participants, regulators, policy makers, and their constituents. Progress still emerges cumulatively from series of investigations, conversations, and debates, but in general not from series of experiments only.

---

that he would be paid $50 for the first unit he bought (minus his purchase price). So if those two participants happened to transact at a price of $30, the seller would earn $10 while the buyer would earn $20.

[2] See particularly Roth (1995a), Holt (1995), Kagel (1995), and Ledyard (1995).

Among the experiments covered in Roth (1995) but worth mentioning again in a discussion of market design are Hong and Plott (1982) and Grether and Plott (1984). Both experiments investigated how the regulation of pricing practices could influence prices, in connections with investigations by the Interstate Commerce Commission and the Federal Trade Commission, respectively. The ICC case concerned whether barge operators should be required to post prices and announce price changes in advance. The FTC case also involved advance notification of price changes, and other contracting policies of four chemical companies that supplied oil refiners with the additives for "leaded" gasoline, as that product was being gradually phased out due to environmental concerns. In both experiments, a simple laboratory environment was created that captured some of the important aspects of the situation, and in both experiments changes in the regulations concerning price announcements and contracts influenced prices. Hong and Plott note that it is difficult to draw general conclusions from such an experiment, but that the results "shift the burden of proof" and put the burden on those who would argue that in the markets of interest in the field the results would be different.

Experiments seem to have been most useful in practical design when they are used as complements to other empirical and theoretical work. Used together with other tools, experiments have played multiple roles, not only in designing new marketplaces and institutions, but in helping diagnose and understand market failures and successes, and in communicating results to policy makers. This will make the discussion of experiments in market design somewhat different from the other chapters in this volume and from the discussions of the older strands of market design in the previous volume (Kagel and Roth, 1995). To put market design experiments in context, it will be necessary to describe at least briefly the problems that a new market design was called on to solve, the technical and political and other obstacles that were faced, and how the experiments were used as complements to other, non-experimental work to bring the effort to a conclusion. The most complete account of the role of experiments in market design can therefore be given in connection with designs that have been adopted and implemented.

But there are lots of barriers to new market designs, and so of course experiments have also played a role in practically motivated design efforts that did not end with the adoption of a new marketplace. Much can nevertheless still be learned from some of the earliest market design experiments, which fall into this category. In Section 2 I'll discuss early experiments aimed at improving the allocation of airport takeoff and landing slots. This is a subject that experimenters have contributed to for over three decades now, without yet seeing the adoption of an efficient allocation scheme. Some of the issues that arise in connection with airport slots also arose in designing auctions for radio spectrum, discussed in Section 3, and here economists, including experimental economists, were more successful in contributing to design decisions that were ultimately implemented. Section 4 discusses how experiments have played a role in another area of auction design, concerning eBay's auctions and reputation system, and also discusses an experiment that has contributed to the so far unsuccessful attempt to replace a flawed Medicare procurement auction. Section 5 discusses labor market clearinghouses, a domain in which economists, and experiments, have played a large role in the design and implementation of new marketplaces, and also discusses signaling in decentralized labor markets and matching processes, and how these have been illuminated by an experiment in online dating. Section 6 discusses an experiment that was instrumental in the adoption of a new course allocation system at Wharton, and Section 7 concludes.

It's useful to think of economics experiments (and a good deal of economics in general) as being part of three big conversations, which I spoke about in Roth (1995a) as "speaking to theorists," "searching for facts," and "whispering in the ears of princes." Market design is clearly aimed at princes, and their modern day incarnations as businessmen, bureaucrats, politicians, and policy makers of all sorts. It turns out that some of the things that sway princes are the same things that persuade scientists, so that testing theory and discovering and documenting behavioral regularities play a role in bringing new designs from conception to implementation. But policymaking also involves a rich palette of demonstration, persuasion, and communication in addition to purely scientific concerns, and we'll see that experiments are also useful for these purposes (see e.g. Bolton and Ockenfels 2012 also).

One long anticipated use of experiments, that came to fruition particularly as a generation of computerized auctions came to be deployed, was as *test beds* used to test that the proposed auction designs were usable by bidders, much as wind tunnels are used to try out scale models of new aircraft before full size aircraft are actually built (see e.g. Plott 1987 on this). Often this involves using the laboratory infrastructure to test a system after it has been at least partly designed, but before it has been deployed. Testing is a use of experiments to which I'll return when speaking of the design of complex auctions, and again when describing how the new market for MBA course allocation was implemented at the Wharton School of the University of Pennsylvania.

Another use of experiments that grew in importance as economists became involved in designing auction markets in particular was as *demonstrations* of underlying economic principles. One of the clearest examples of a demonstration experiment that has been widely used to illustrate an economic principle is the auction of a jar of coins to illustrate the *winner's curse* in a common value auction. The idea is to concretely model an auction in which bidders do not know with certainty the value of the object being auctioned, such as an auction for the right to drill for oil or to harvest timber at a certain site. Bob Wilson invented a demonstration that has proved very durable, for advising both bidders and auction designers.

In one variation, the demonstrator circulates a closed jar filled with coins among the audience, allowing everyone to examine the jar (but not to open it). After everyone has examined it, a first-price, sealed bid auction is conducted for the value of the coins in the jar. That is, each member of the audience is invited to write down a bid; these bids are collected, and the highest bidder pays his bid and receives in return the value of the coins in the jar (which is typically paid in paper money, for the convenience of the winning bidder, and of the demonstrator who keeps the jar for another day). Sometimes the audience members are additionally asked to write down their estimate of the value of the coins in the jar, for discussion afterwards.

A very usual outcome is that the estimates of the value of the coins in the jar vary widely, but are distributed roughly around the actual value. That is, collectively, the estimates are not too bad. But the auction is often won by the bidder who had the highest estimate, and although his bid is typically lower than his estimate, it is almost always above the actual value of the coins. Thus the winning bidder loses money; he suffers from the "winner's curse."

The theory of common value auctions was initially explored in Wilson (1967, 1969) and Rothkopf (1969). They initiated the study of models in which n bidders each receive a noisy signal of the common value of the object being sold (e.g. estimates from their geologists of the amount of oil that can be recovered). This signal gives the bidder an estimate of the value, but not as good an estimate as if he could also see the other bidders' signals. A simple way to appreciate the problem facing such bidders is to suppose that all bidders adopt the same bidding strategy, and that the higher their own signal (i.e. the more valuable their information indicates the object is), the higher they bid. In this case, the bidder with the highest signal will win the auction. But even if all the signals are unbiased, i.e. if they are drawn from a distribution whose mean is the true value, the highest of n such signals (the nth order statistic) will be higher than the true value (and much higher if n is large). If bidders understand this, then their bidding strategy should take into account that they must substantially discount the naïve estimate based on their own signal alone, i.e. the naïve estimate that ignores the fact that, if they win the auction, their signal is the highest of n such signals.

A bidder who fails to reduce his bid sufficiently below his signal is likely to find that the value of the object he has won is less than his bid. In auctions in which such mistakes are widespread or persistent, it could be that winning bidders will regularly lose money. This possibility was brought to the attention of the oil industry by Capen, Clapp, and Campbell (1971). But an article in a petroleum journal urging oil companies to bid substantially less than their geologists advised them might be just an attempt to foster collusion, or to gain a bidding advantage over competitors, rather than a description of a

widespread mistake. The jar of coins demonstration was designed to show that this is an easy mistake to make.

In reply to my query about the origin of this demonstration, Wilson wrote:

> "I recall using it in a series of three lectures at Weyerhauser in Tacoma in about 1970-71 or so … and occasionally thereafter … such as at the Dept of Interior roughly 1973-4 and then late 70s with oil companies, and of course in classes at the GSB where invariably the students overbid" (Wilson, 4/18/2008 email)

When the demonstration is used to show that the auction design matters, the first-price sealed bid auction is sometimes followed by an ascending oral auction, in which the auctioneer calls out ascending prices and bidders indicate with raised hands whether they wish to continue bidding at that price. When I have used the demonstration for this purpose, I announce at the outset that I will auction the jar in two different ways, and that we will afterwards toss a coin to determine which of the two auctions will determine the winner and the winning price. The value of the jar of coins is revealed only after both auctions have been conducted.

The difference between the sealed bid and the oral auction is that the bidders in the oral auction can see when other bidders drop out, and so a bidder with a high estimate quickly learns that most other bidders had lower estimates. This allows bidders in the oral auction to update their estimates, in a way they cannot in the sealed bid auction.

Preston McAfee conducted such demonstrations during the discussion of the design of the FCC auctions of radio spectrum (to be discussed in Section 3). McAfee used a jar containing 200 M&Ms, and told the bidders that the (unknown number of) M&Ms in the jar would be worth $0.10 each, and he displayed a closed envelope that contained the (unknown) value. He wrote:

> "I sold the envelope, although I did pass the M&Ms around afterward… I sold a $20 bill for $140; this was the extreme. … I ran first a sealed bid, then before revealing the results, an oral auction. The winner's curse was invariably less in the oral auction, but [the] winner in the oral auction also lost money.
>
> "I did it six or seven times to different telecom audiences. The largest - with the $140 winner - was CTIA. I was on the cover of their magazine later.

> "One guy in the audience said "it doesn't matter what you bid, later it will be worth three times as much.". I said "you should bid a trillion dollars." Anyone wondering about the telecom meltdown only needs to know that participant's mindset." [McAfee, 6/11/08 email][3]

There are a number of differences between experiments primarily intended as demonstrations and experiments conducted to carefully test hypotheses (not that the jar of coins demonstrations don't test, and reject, the hypothesis that the winner's curse is just a hypothetical mistake that cannot readily be observed). Among these differences are how much attention is given to controlling the environment (e.g. to making sure that bidders don't directly communicate their private estimates to one another), how much effort is spent investigating relevant parameters (such as the number of bidders) by systematically varying them, how much care is taken with the experimental design (e.g. if we want to compare oral auctions with sealed bid auctions, it would be better to run them under identical conditions, instead of running the oral auction with bidders who had just participated in a sealed bid auction). Last but not least, in formal experiments care is taken to collect, analyze and report the data. So, despite the rapid spread of the jar of coins demonstration, particularly as a teaching tool in classes that covered auctions, it was a welcome development when the winner's curse started to be examined in the laboratory.

The first paper I know of reporting an experimental examination of the winner's curse is Bazerman and Samuelson (1983), who literally studied the jar of coins demonstration in the laboratory. They varied the number of bidders and the contents of the jar, and solicited confidence intervals about the value of the jar along with sealed bids. They report that the size of the winner's curse increases in the number of bidders and the uncertainty about the contents of the jar. Subsequent experimenters have implemented common value auctions in ways that give them more flexibility and control over the signals that bidders receive, and how these are related to the true common value. In a striking series of experiments by Kagel and Levin, the common value (that the winning bidder will receive) is drawn from a distribution known to the bidders, and then each bidder receives a signal independently drawn from a known distribution around this

[3] CTIA is a wireless industry association originally called the Cellular Telecommunications & Internet Association.

common value (see Kagel 1995 and Kagel and Levin 2002 for surveys of this literature). These experiments have documented new regularities (e.g. concerning the effect of providing an additional public signal in auctions in which the winner's curse is present), and have in turn inspired some new directions for both empirical investigation (see e.g. Kagel and Levin 1986) and theory (see e.g. Eyster and Rabin 2005).

In short, an experimental demonstration that grew out of theoretical issues related to auction design led in turn to a formal program of experimentation that has raised new issues, some of which are also of concern in market design. But even as a demonstration, the jar of coins experiment helped bring the winner's curse to the attention not only of bidders but also of auction designers, at the Department of the Interior for offshore oil leases, and to various parties with interests in the design of the FCC's radio spectrum auctions, in a way that would have been difficult to do from the theoretical literature alone.

In the case of the winner's curse, experiments were used to demonstrate and then to study a phenomenon that was originally controversial among economists, since it is an out of equilibrium phenomenon: at equilibrium, bidders all discount correctly and no one is a predictable victim of the winner's curse. In the next section we will see that experiments can also be used to demonstrate points that are relatively uncontroversial among economists, but may still need to be communicated to policy makers in an effective way.

2. Some early design experiments: allocation of airport slots

An early attempt to use experiments for practical market design comes from airline deregulation. Many aspects of the airline business that were once regulated have long since been deregulated and opened to decision making by individual airlines and competition among airlines. These include how ticket prices are set, how it is determined which airlines will fly between which cities, and even how new airlines may enter the market. But, at least since 1969 when the Federal Aviation Agency limited the number of

takeoff and landing slots at several of the busiest airports, the allocation of these slots by administrative means has been a source of potential inefficiencies. There has been a longstanding interest (which continues today) in replacing these administrative procedures with some kind of market both for allocating slots and for allowing airlines to trade already allocated slots. This might call for a market of considerable complexity and flexibility, as slots at a given airport are complements both to other slots at the same airport and to slots at the other airports at which the corresponding flights will begin or end.

The Civil Aeronautics Board (CAB) commissioned a study by Grether, Isaac, and Plott (1979) to compare the existing system of slot allocation by committee to a simple market. Grether, Isaac, and Plott (GIP) report that one of the first things they did was sit in on some of the committee meetings (transcripts of these meetings are included in their report). These committees had been formed after a 1969 Federal Aviation Agency (FAA) ruling limiting the number of takeoff and landing slots per hour at the most congested airports (at that time La Guardia, JFK, Washington National, and O'Hare). Each airport had a separate committee with members representing the airlines operating out of that airport. The committee discussions were restricted to slot allocations at a single airport, preventing discussion even of what other airports were involved in a flight that caused demand for a particular slot. The original task of these committees had been to coordinate the trade of slots among the incumbent carriers. But, following the Airline Deregulation Act of 1978, these committees would also have to allocate slots to new entrants. Plott (1985) writes

> "The CAB staff became concerned that the committees could be used as a barrier to new competition. I was contacted to study the committees because of my previous [experimental] work on committee behavior [Fiorina and Plott, 1978]."

GIP observed that the committees operated by consensus, that is, unanimity appeared necessary to make a change in existing allocations. However, at the time of their observations, since no committee had failed to reach consensus, it was not completely clear what the FAA would do in case of such a failure. The first laboratory experiments reported by GIP were therefore designed to simply *demonstrate* that the

outcome of a unanimity-rule committee could be highly sensitive to the "default" allocation that would result in the absence of agreement.[4]

Twenty-three laboratory committees (consisting of either 9 or 14 members) were observed under a variety of conditions. Committee members were given initial endowments of "cards" and "flags" of various colors, with instructions on how much combinations of these would be worth to them at the end of the session. Committees met twice, first to allocate "cards" and then to allocate "flags," with the value of a flag depending on how many cards had been obtained in the first meeting (think of cards and flags as being slots at different airports.) In each committee meeting, one of the following three default rules was used to determine what would happen in case the committee defaulted, i.e. in the absence of a unanimous agreement: 1. each committee member would receive his/her initial endowment; 2. each committee member would receive a random allocation unrelated to initial endowments; 3. each committee member would receive an endowment created by taking items at random only from those with large initial endowments and giving those items to those with small or zero initial endowments. (This latter condition was motivated by a belief that the FAA might mandate such adjustments to facilitate entry of new carriers.)

The results were that, in each condition, the final allocations were close to the expected value of the default allocations. When these were the initial endowments, final allocations were near the initial allocations or somewhat below them for those with large endowments, and when default would result in a random allocation, final allocations were essentially equal, independent of initial allocations. In particular, GIP noted that while the final allocation was sensitive to the default rule, it was not sensitive to the underlying distributions of values that determined the profitability of different combinations and

[4] Grether, Isaac, and Plott (1989) write in their introduction to the reprinted report (p.xi) "The experiments...are described in the report as "demonstrations." Because we used experiments, there was no need to explain the details of a game-theoretic model or a solution concept such as a core. The reader could look at the rules, use some intuition about what might happen, and then look at the data. In this way a level of understanding about the argument could be achieved without resort to complicated theory." In a similar spirit, GIP (1981) speak of the role of these experiments in communicating with policy makers as follows (p166): “This type of evidence will probably be of little value to economists who already have considerable experience with the behavioral properties of a variety of allocation processes.... [S]ome decision makers may have no experience with game-theoretic models, and rely on instincts and general theories of a completely different sort."

hence the efficient allocation. And the committees were also not able to coordinate efficiently between first and second meetings; in each meeting the outcome was primarily determined by the default rule and the initial endowment in that committee session.

GIP suggested that a more efficient way of allocating slots would be to auction them in each airport in a multi-unit sealed bid auction in which all bidders would pay the lowest accepted bid[5], and then to allow an aftermarket in which units bought in these individual markets could be traded. They conducted several demonstration sessions of laboratory auctions to initially allocate goods, and of oral outcry double auctions to trade them. At least one experimental session paralleled the "cards and flags" condition of their committee experiments, and they noted the increased efficiency of the resulting allocations.[6]

They also noted that it would be desirable for an actual market for trading takeoff and landing slots at different airports to allow some sort of package bidding for "blocks" of slots, since airlines have economies of scale at a given airport, and since takeoff and landing slots are related to business plans involving which routes an airline will serve. They wrote (GIP, 1979, pVI-8 [square brackets were footnotes in the original])

> "Each carrier would register in a central computer the maximum (minimum) price it would pay for (sell) a particular slot. Contingencies such as block provisions [A carrier may want to buy (sell) only if it can acquire (sell) a certain set of slots.] should also be listed. Such contingencies allow carriers to take advantage of interdependencies of operations which occur because of time and size (nonconvexities). By simply asking for a 'print out' each carrier can see the full pattern of offerings at any given time and can activate a transaction through the computer (an 'open book' feature).[The identity of the carrier making an offer (bid) to sell (buy) would <u>not</u> be available to the potential buyers (sellers).]

---

[5]GIP suggested that such an auction was roughly incentive compatible. (VI-3 "This particular market organization has the feature that the optimum bidding strategy is for each buyer to bid the maximum that he/she is willing to pay (except possibly for the marginal bids where the strategy is sensitive to the information state of the bidder).

[6] GIP draw a contrast between their double auction experiments and previous experiments conducted to demonstrate that double auctions could quickly converge precisely to competitive equilibrium, in which the market was run repeatedly with precisely the same conditions. In the preface to the reprint, GIP (1989, page x) they note "The relationship between initial endowments (randomly determined) and markets is studied experimentally for the first time in Chapter VI. As can be seen (Figure 26), the market has much more variance than is the case when the distribution of supply is constant over time." This refers to a double oral auction run for 18 periods (with a shift in supply and demand at period 7). Starting in period 14, the initial endowments were randomly permuted between players, so that from period to period players no longer had the same endowment (although aggregate supply and demand were kept constant). Most transaction prices become much further from the equilibrium prices (although the final transactions in each period are close to the equilibrium price.)

> Many techniques exist for summarizing information and allowing participants to be fully aware of the state of the market. [Those desiring further details about such a computerized market should contact the authors.]"

GIP did not, however, report any experiments with a market that allowed package bidding for blocks of slots. Their proposal for auctions of slots followed by trading of slots with package bidding was not adopted.

In 1982, a proposal for directly allocating slots via a package bidding auction was made, and an experiment was reported, by Rassenti, Smith and Bulfin (1982). Rassenti, Smith and Bulfin (RSB) proposed that all slots should be simultaneously allocated in a "combinatorial" auction that would allow airlines to place bids for packages of slots, with a bidding language that would allow an airline to make bids on multiple packages while specifying e.g. that it wanted one or the other of two packages but not both. The winning bids would then be determined by finding the revenue maximizing set of non-intersecting packages. They noted that this involved solving an integer programming problem.

They further noted that an auction of this kind would not be incentive compatible, that is, "the door is open to the possibility of strategically underbidding the true value of certain packages." But they conjectured that strategic bidding in this environment is complex, and that this might deter it. In particular, they noted that there had been a good deal of demand revelation in the Grether, Isaac and Plott experiments, and they proposed that the complexity of a package bidding auction would promote straightforward bidding:

> "Since the combinatorial auction we suggest for the airport slot problem is far more complex than any of the single commodity auctions that have been studied, we would expect to observe at least as much demand-revealing behavior in our auction as in the others. Since this is both an open and a behavioral question, we devised a laboratory experimental design ... (p407)

RSB considered an environment in which six participants played the roles of airlines needing to bid for packages of up to six slots, with each participant being given a redemption value for up to six packages of slots (chosen by the experimenters from the 63 possible nonempty packages), which he would be paid if he completed the session in possession of one of those packages.

RSB report they employed a 2x2x2 experimental design, in which the variables were (1) the GIP auction and secondary market versus the RSB package bidding auction followed by a secondary market; (2) experienced versus inexperienced subjects (with experience meaning prior participation in one of the cells of this experiment); and (3) an "easy" versus a "difficult" arrangement of values from the point of view of finding an efficient solution.

They only reported one market in each cell, for a total of 8 laboratory markets, including 4 with their package auction (one with inexperienced subjects in the "easy" environment, and one in the "difficult" environment, and the same with experienced subjects).[7] They noted that this precluded standard statistical tests by treatment, however they concluded that selling slots via the package bidding auction was more efficient than selling them individually.

Regarding the next steps prior to adoption of this mechanism they wrote:

> "We think the RSB mechanism, or some variant that might be developed from it, has potential for ultimate application to the time slot problem. But as we view it, before such an application can or should be attempted, at least two further developments are necessary. First, at least two additional series of experiments need to be completed. Another series of laboratory experiments should be designed, using larger numbers of participants, resources, and possible package combinations. The subjects in these new experiments should be the appropriate operating personnel of a group of cooperating airlines. Depending on the results of such experiments, the next step might be to design a limited scale field experiment with only a few airports and airlines."
>
> "Second, there should be extensive discussion and debate within the government, academic and airline communities concerning alternative means of implementing the combinatorial auction..." (p412)

Regarding possible computational difficulties of scaling up the integer programming solution from the six slot experiment to one of practical scale, they wrote:

> "The [integer programming] problem which results is recognized as a variant of the set packing problem with general right-hand sides. It can be solved,

[7] In those days some (but by no means all) experimental economists referred to each observation of a market as an experiment. So Smith et al. report they did "eight experiments". Roth 1994 argued against this convention, and in favor of reporting full experimental designs and all observations made in each cell of such a design as a single experiment.

> as was done for the experiments...with a specialized algorithm developed by Rassenti (1981 [unpublished]). A problem of the enormous dimensions dictated by even a four-city application (perhaps 15,000 constraints and 100,000 variables) will present a significant challenge for the finest configuration of hardware and software available. Fortunately, a practicable solution within 1 or 2% of the linear optimum, and very often the optimum itself in the discrete solution set, is almost assuredly achievable in a reasonable amount of time." (p404).

It is difficult to know precisely what to make of this. Set packing was one of Karp's (1972) original 21 equivalent NP complete problems, which essentially means that there is no guarantee that large problems can be solved in a practical way. RSB give no indication that they are aware that computational complexity may be an issue, but in 1982 computational complexity was starting to be widely recognized (following e.g. the well known book of Garey and Johnson, 1979). It may well have appeared to experienced observers that too much optimism was being expressed on the basis of an experiment and computations of modest size.[8]

However the failure to adopt airport slot auctions in the 1980's can hardly be attributed to the limitations of these pioneering experiments. A quarter century later, real progress has been made on the technical obstacles to practical auctions that would allow airlines to express preferences over at least some packages of landing slots (see e.g.

[8] Indeed, as I'll discuss in connection with the FCC spectrum auctions, when the FCC used a combinatorial auction in 2007, they felt that these complexity issues had still not been surmounted, and elected to employ an auction in which only certain predefined packages could be bid on. The modern experience with large NP complete integer programming problems suggests that NP completeness itself is sometimes less of a practical obstacle than might have been supposed, since average problems can be much more tractable than the worst case examples that determine if a problem is NP complete (cf. Leyton-Brown, Nudelman, and Shoham, 2006). But other kinds of computational difficulties had not yet been surmounted in 1982 and still require considerable attention today. For example, large integer programs of the kind needed have so many constraints that they cannot be easily written down, but must be generated automatically, and can easily surpass the memory capacity of even 21st century integer programming solvers (at least as of 2014). As a result, techniques have been developed to solve such problems by breaking them into subproblems that can be solved sequentially. My own experience of this comes in the design of kidney exchanges that involve solving NP complete integer programs to find maximal sets of exchanges (and chains) of bounded size (Roth et al. 2006, 2007). My colleague Utku Unver wrote a program that used the commercial integer programming solver CPLEX, which worked well for the regional kidney exchanges that used it (the New England Program for Kidney Exchange and the Alliance for Paired Donation), but which could not handle more than about 900 donor-recipient pairs. When legislation was passed in 2007 that removed the legal obstacles to a National Kidney Exchange Program, we appealed to computer scientists to find a way to deal with a larger number of pairs, and Abraham, Blum, and Sandholm (2007) used column generating techniques to feed subproblems to CPLEX in a way that allows up to 10,000 pairs to be handled. New computational problems related to memory appeared as large parts of kidney exchange shifted to chains, and workarounds for the computational complexities of these problems continue to be developed (see e.g. Anderson et al, forthcoming]).

Cramton, Shoham, and Steinberg 2006). But a 2007 NY Times headline summarized the political situation nicely:

> "**Airlines at La Guardia Fight Bush Administration Proposal to Auction Off Landing Rights**" (Wald, 2007)[9]

And, in the face of stiff political opposition, planned auctions were delayed and, in 2009, eventually abandoned (at least for the time being).[10]

The contemporary discussion still uses experiments, but in a supporting role to show the feasibility of various auction possibilities in something like a "war gaming" or simulation setting, intended to show participants and observers that the system could work under realistic conditions (sees e.g. Ball et al. 2007). It looks like airport slots at congested airports *eventually* may be allocated by auction, but it is not clear when.

Package bidding auctions and exchanges of various kinds have been proposed and sometimes adopted in commercial applications. Cassady (1967), for example, speaks of auctions that allow "entirety bidding" as well as ordinary bidding. In such an auction, bids are accepted both for individual lots and for all the lots together (as when a manufacturing plant is being auctioned off either piecemeal or in its entirety). The high bid for the whole package is compared with the sum of the winning individual bids to determine whether the highest package bid is the winner, or if the lots will be sold

---

[9] The Federal Aviation Administration managed congestion at LaGuardia under the High Density Rule from 1969 through 2006. During this time (starting in 1985) although slots were not regarded as the property of the airlines, airlines were permitted to buy, sell, or lease slots on a secondary market. Since January 2007, the regulatory authority has been extended while new rules are considered. Proposed new rules would have grandfathered most slots, but would have auctioned a small number of slots each year, perhaps by an ascending clock package bidding auction (see e.g. FAA 2008a). The situation is largely the same for Kennedy and Newark airports, see Wald (2008), FAA (2008b).

[10] The departing Bush administration had scheduled an auction for some takeoff and landing slots at JFK, LaGuardia, and Newark airports for early January of 2009, literally days before the Obama administration took office. It is hard to know how serious that plan was, and how much was just political posturing. In the event, the airline industry obtained an injunction, and Obama's Transportation secretary Ray LaHood later cancelled the planned auctions. (Voorhees, May 13, 2009)

separately. For example, in 2012 the State of Washington auctioned off 167 State liquor stores by an entirety auction, which was won by the individual bidders.[11]

More comprehensive kinds of package bidding auctions have been proposed, sometimes with the support of experiments. Ledyard, Olson, Porter, Swanson, and Torma (2002) describe how in 1993 they supported the adoption of a package bidding auction that was used five times in 1995 and 1996 by Sears Logistics Services (which organized truck transport for Sears). They helped design the auction, which they called a Combined Value Auction (CVA), motivated in part by the work of Banks, Ledyard and Porter (1989). Experiments were used to demonstrate that the auction was usable. They write

> "Once we had designed an acceptable CVA, we had to explain it to the SLS team and get its approval. The experimental test bed was an important demonstration tool. We took the test bed to SLS so the team members could participate in a CVA. The goal of the demonstration was to show them that trucking firms could understand the auction procedures and that SLS would incur savings. The demonstration convinced the SLS team that a CVA was workable." (p8)

Banks et al. write (p9) "The key measure of success for any new auction design is whether it is used. The CVA implemented by SLS has been a success."

Of course progress can be made by design proposals even if they do not always directly result in success by this measure. For some other proposals for package bidding auctions or exchanges put forward by the pioneering Cal Tech group of market design experimenters, see e.g. Brewer and Plott (1996, 2002), Ledyard, Hanson, and Ishikida (2005), Ledyard, Noussair and Porter (1996), Ledyard, Porter and Rangel (1997), and Plott and Porter (1996). Plott, Lee and Maron (2014) report some recent implementations.

Goeree and Lindsay (2012) look at a double auction market that allows for package bidding, and find efficiency gains compared to a simple double auction in a market in which everyone owns a "house" and wants one, but doesn't value having two. In this market trades may be complements, as an agent may only wish to buy a house if

[11] See http://marketdesigner.blogspot.com/2012/09/washington-state-liquor-stores-followup.html

he can sell his own, and be deterred from buying if it exposes him to the probability of exiting the market with two houses. They write of the comparison: "The results show that in a standard double auction market only a small fraction of the total gains from trade are realized…This poor performance is due to the exposure risk that arises when going from the initial allocation to the optimal one requires someone to temporarily make a loss. The solution presented in this paper is a simple package market."

Much of the discussion of package bidding has been in connection with the auction of radio spectrum licenses, and has concerned the exposure to risks associated with the need to assemble packages.

3. FCC spectrum auctions

Prior to the Omnibus Budget Reconciliation Act of 1993, licenses for radio spectrum had been given away for free (see e.g. McMillan 1994 and McAfee and McMillan 1996). The 1993 Act gave the FCC one year to design and run an auction. Each of the potential bidders proceeded to hire consultants, including both auction theorists and experimenters.[12] Plott (1997) notes:

> "By the fall of 1993 the business world was fully aware of the rulemaking process and had engaged many groups of consultants to help them position themselves. Businesses understood that the rules and form of the auction could influence who acquired what and how much was paid. The rules of the auction could be used to provide advantages to themselves or to their competitors. Thus, a mixture of self-interest and fear motivated many different and competing architectures for the auctions as different businesses promoted different rules. The position of the FCC was that the efficient allocation of the licenses was to be the primary criterion for deciding among the competing options." (p606)

[12] McMillan (1994) gives the following partial list: "Paul Milgrom, Robert Wilson, and Charles Plott were hired by Pacific Bell, Jeremy Bulow and Barry Nalebuff by Bell Atlantic, Preston McAfee by Airtouch Communications, Robert Weber by Telephone and Data Systems, Mark Isaac by the Cellular Telecommunications Industry Association, Robert Harris and Michael Katz by Nynex, Daniel Vincent by American Personal Communications, Peter Cramton by MCI, John Ledyard and David Porter by the National Telecommunications and Information Administration, and the author of this article by the Federal Communications Commission (FCC)."

Because of the deadline, every aspect of that initial design process, which culminated in auctions run in 1994, was done in a hurry, in a process that involved public comment from theoretical and experimental economists as well as from communications companies and other interested parties. Under the circumstances, McAfee and McMillan (1996) suggest that

> "A lesson from this experience of theorists in policy-making is that the real value of the theory is in developing intuition." (p. 172).

Plott (1997) argues that experiments initially played a similar role. For example, he notes that there was some concern about whether computerized auctions of any sort would be easy enough for bidders to use. He writes that the fact that computerized auctions had long been used in laboratory experiments helped to quell these concerns, especially when experimental auctions were demonstrated at a meeting held at Cal Tech in January 1994, at which presentations were also made by representatives of "the Pacific Stock Exchange and other parties familiar with the operations of electronic and computerized market processes."

While a number of experiments were conducted as parts of the discussion, and often focused on the merits of combinatorial auctions (see e.g. Ledyard, Porter, and Rangel 1997), the FCC eventually settled on a simultaneous ascending auction, proceeding in rounds, in which each license was auctioned separately but at the same time, and no auction could end until all auctions ended. To assure that bidding proceeded in an orderly way, there were activity rules that required bidders in later rounds to have been active in earlier rounds. To deal with "exposure" problems there were rules permitting bid withdrawals with potential penalties. The idea behind allowing withdrawals was to partially protect bidders who might otherwise be exposed to too much risk of winning only parts of a package of licenses that they were trying to assemble, and find it unprofitable to purchase only the licenses that they had won at the prices they had bid in anticipation of winning the whole package. While many of the experimental papers had proposed package bidding to address the exposure problem, the final simultaneous auction design seemed to mostly reflect the contributions of auction theorists (particularly Milgrom and Wilson, and McAfee and McMillan, cf. Milgrom, 2004).

(Plott (1997, p627) writes: "*A lack of confidence in technology, as well as a lack of theory, seemed to dampen enthusiasm for the implementation of a "smart market" that would be capable of dealing with complex bids for packages of licenses.*")

Nevertheless, Plott reports that experimenters played two unusual roles in the implementation of the first FCC spectrum auction in July 1994. First, Plott and his Caltech colleagues Ledyard and Porter were retained to test the software supplied by FCC contractors to run the auction.[13] They were also asked to attend the first auction, which began July 25, 1994, in the Omni Shoreham Hotel in Washington DC, to stand in as a "backup team" of auctioneers, in case one should be required because of a software or other failure. (In the event, no such failure occurred.)

Following the initial 1994 auction, the design discussion has continued unabated, with a number of FCC calls for further comment. Although there have been almost constant proposals for the FCC to explore more ambitious, combinatorial auction designs, often supported by new experiments, the initial simultaneous ascending auction design has proved surprisingly robust, and has accounted for the vast majority of FCC spectrum auctions to date, with few exceptions.

The simultaneous ascending auction design has also been used elsewhere. Binmore and Klemperer (2002) describe how the simultaneous ascending auction was chosen for the British third-generation mobile-phone license auction that concluded in April 2000, and how that debate was also a contentious one in which experiments played a role. They conclude in part (p. C95) that "*The value of computer simulations as an educational tool, and the persuasive power of laboratory experiments, was also brought home to us.*" (See also the experiments of Abbink et al. 2002, 2005.)

Before describing two exceptions to the FCC's use of simultaneous ascending auctions, and the experiments that supported them, it may be helpful to think about why

[13] "However, testing was made very difficult by FCC policies. The FCC adopted a policy of not letting the Caltech team have access to the final software, study (or see) the code, or even talk directly to a software developer. Thus, the auction process was a 'black box' from the point of view of the Caltech testers." Plott, 1997, p630.

the steady stream of experiments pressing the case for package bidding auctions faced so much resistance.

Vernon Smith (2008), in a chapter on the FCC auctions, attributes what he felt was a lack of success by experimenters at influencing policy to mistaken positions taken by policymakers and the other economists involved in the policy making process. He attributes the resistance to experiments 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)

It is easy to sympathize with Smith's frustration, and no one who has been engaged in complicated changes in policies with many constituencies can doubt that the process can be difficult on political as well as scientific grounds. (Indeed, the actual *adoption* of a new market design that will affect many interested parties in different ways is almost *by definition* political, and the FCC auctions took place on a national political/regulatory stage, with the whole apparatus of a formal process of public comments.) But 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.[14] However he is certainly correct that none of the particular design proposals advanced by experimenters were adopted, until the proposal by Goeree and Holt (2010) that I'll describe shortly. The issue seems to be the continued lack of confidence that the most ambitious proposals could in fact be implemented.

[14] The FCC's Evan Kwerel (2004) writes "The National Telecommunications and Information Administration (NTIA)…had proposed [a] combinatorial auction mechanism…The design, based on the work of Banks, Ledyard, and Porter (1989)…seemed far too complex for the FCC to implement in the time available….Though the FCC did not adopt the NTIA proposal, the fact that the NTIA proposed a simultaneous auction design was helpful in building support for the Milgrom-Wilson design. It made that mechanism look like a reasonable middle ground between sequential ascending bid auctions and simultaneous ascending auctions with package bidding." And Connolly and Kwerel (2007, p117) write that some of the experimental work presented at the Cal Tech conference "helped convince the FCC that auctioning licenses simultaneously" would be feasible and preferable to sequential auctions.

Milgrom (2007) notes that part of the problem was that some of the experiments themselves were not so transparent. He writes as follows (p953) about one of the experiments, prepared as a consulting report (Cybernomics, 2000), whose authors concluded that a combinatorial auction should be adopted:

> "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."[15]

But apart from issues of confidence in the experiments themselves, a continuing obstacle to the adoption of auctions that allow package bidding was their inherent complexity. If there are k licenses for sale, there are 2^k-1 possible packages that someone could bid on, so even a modest number of licenses quickly lead to an auction that is complex for bidders to participate in, and could be computationally complex to determine the winners, i.e. the set of packages whose bids would maximize revenue. In an attempt to address this concern, Rothkopf, Pekec and Harstad (1998), proposed a class of auctions that could be made computationally simple for both bidders and auctioneers, by severely reducing the set of packages on which bids would be allowed. Building on this work, an experiment by Goeree and Holt (2010)[16] set the stage for the use of a simplified combinatorial auction by the FCC, to auction 62MHz of spectrum in the 700MHz band in a multi-round auction, FCC auction #73, running from January to March, 2008. The gross revenue from winning bids was $19.12 billion, the largest amount in any single FCC

[15] Milgrom NBER May 2009 reports that, similarly, no data could be recovered from the combinatorial auction experiment reported in Porter, Rassenti, Roopnarine, and Smith (2003). Of this experiment Kagel, Lein and Milgrom (2010) say (p161) "Porter et al. (2003) report surprisingly efficient outcomes from an experiment testing the CCA. In 25 auction trials, efficiencies of 99% are reported in two trials and 100% in the remaining 23 trials. Unfortunately, these results cannot be replicated because detailed information about the valuations used in their experiment is unavailable."

[16] See also the experimental papers Goeree and Holt (2005) and Goeree, Holt and Ledyard (2006, 2007) on the FCC website http://wireless.fcc.gov/auctions/default.htm?job=papers_studies .

auction up to that time. Net revenue, accounting for bidding credits, was $18.96 billion."[17]

Recall that the attraction of package bidding arises when goods may be complements, so that some bidders value packages more highly than the sum of the values of their components (as when airlines get more value from packages of takeoff and landing slots than they would from individual slots, or, in this case, if phone companies get more value from a package of licenses that allows them to offer service over a wide area than from the individual licenses). So the question was whether, in an environment in which goods could be complements, it might be more efficient to allow bidders to bid directly on packages, rather than forcing them to try to assemble the packages they wanted by winning the auction for each component. The discussion *in favor* of combinatorial auctions focused on the *exposure problem* facing bidders who wanted packages but had to bid on individual licenses. The argument was that, even if the efficient, highest value use of spectrum involved assembling packages of licenses, a simultaneous auction might not achieve this if those who valued packages of auctions were deterred from bidding the full combined value on the individual licenses that made up a package, out of fear of winning only a partial package, at individual license prices higher than could be recouped without the missing parts of the package. That is, by

---

[17] This was actually the second auction run by the FCC that allowed package bidding. However the first, Auction 51 in 2003, was for a small slice of spectrum (Narrowband PCS) of licenses that had been awarded in a previous auction and subsequently cancelled. Auction 51 attracted only a single bid. Martha Stancill of the FCC writes (email, June 4, 2010): "There were 2 qualified bidders but only one of them bid -- on a package of the 5 regional licenses in one of the channels (there was only a single regional license on the other channel). The winning bid was the sum of the minimum opening bids -- $179,000, but since the winning bidder had a 25% bidding credit as a small business, the net winning bid was $134,250."
While the auction was a tiny one, it potentially offered the opportunity for the FCC to try out a package bidding auction in which bidders could define their own packages. The package bidding rules were as follows: "The Bureau proposed that, in addition to bidding on individual licenses, bidders be permitted to create and bid on up to twelve different packages of their own choosing during the course of the auction. A bid on an individual license does not count as a bid on a package; packages consist of two or more licenses. Bidders will not be required to identify or create their packages before the start of the auction, but may create their packages as the auction progresses. A bidder may modify or delete a package it has created up until the point where it has bid on the package and the round has closed. If the bidder submits a bid on a package and subsequently removes the bid during the same round, the bidder has the option of also deleting or modifying the package. However, once a bidder bids on a package and the round closes, the package may not be modified or deleted and counts as one of the bidder's twelve allowable packages." (FCC Report No. AUC-03-51-B (Auction No. 51), p25, http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-03-1994A1.pdf). [FCC auction information can be found at their auctions website (which has a drop down box on upper right side for auctions by number) http://wireless.fcc.gov/auctions/default.htm?job=auctions_home .

bidding up the component parts of a package, they would be exposed to the risk of winning only part of the package and paying too much, and this risk might deter them from bidding aggressively.

The arguments *against* package bidding auctions focused not only on their complexity for bidders and auctioneers, but also on the possibility that package bidding might allow bidders interested in big packages to win them even when the parts of the package might be more valuable separately. The idea is that if one big bidder wants a package of spectrum licenses with national coverage, but many small bidders collectively have a higher value for the individual licenses (or small packages), then the small bidders might face a *threshold* (or *free riding*) problem, and fail to outbid the package bidder. That is, only if the sum of the bids on disjoint small packages is greater than the bid on the whole package would the small bidders win, and since each small bidder has little effect on that, the temptation to bid low (and thus earn a high profit if the small bids win) might let the package bidder win cheaply.

One can see how some of the contending companies would worry more about the potential exposure problem that might confront large bidders seeking national coverage, while other companies would be more concerned about the threshold problem that might stymie regional companies seeking to maintain their local dominance (think e.g. Pacific Bell). The politics and economics of design were thus fully joined.[18]

Other competing interests were in play in the design of auction 73. While I have been focusing on the design of auction rules, the design of an auction also involves the question of what is being auctioned. In spectrum auctions, licenses are being auctioned, and licenses are contracts with rights and obligations.[19] Contracts also need to be designed[20], and some licenses in auction 73 carried an unusual obligation to provide an

---

[18] A subtler historian than I can figure out how the business interests involved in the spectrum auctions influenced which economists were hired as consultants by which firms, and how those interests were related to the designs they championed. See footnote 12.

[19] At various points Evan Kweral of the FCC played a critical role in defining property rights for spectrum licenses. Presently, these rights are understood to be attached to a certain bandwidth and type of spectrum, not to a particular frequency, allowing the FCC to occasionally rationalize the assignment of frequencies to accommodate new uses, as will be most apparent in the upcoming "incentive auction" to buy back spectrum from television broadcasters and re-sell it for new uses (see remarks by Paul Milgrom at end of post at http://marketdesigner.blogspot.com/2014/04/the-fccs-upcoming-incentive-auction-and.html)

[20] For experiments concerned specifically with contract design, see Grosskopf and Roth (2009), and Kessler and Leider (2012).

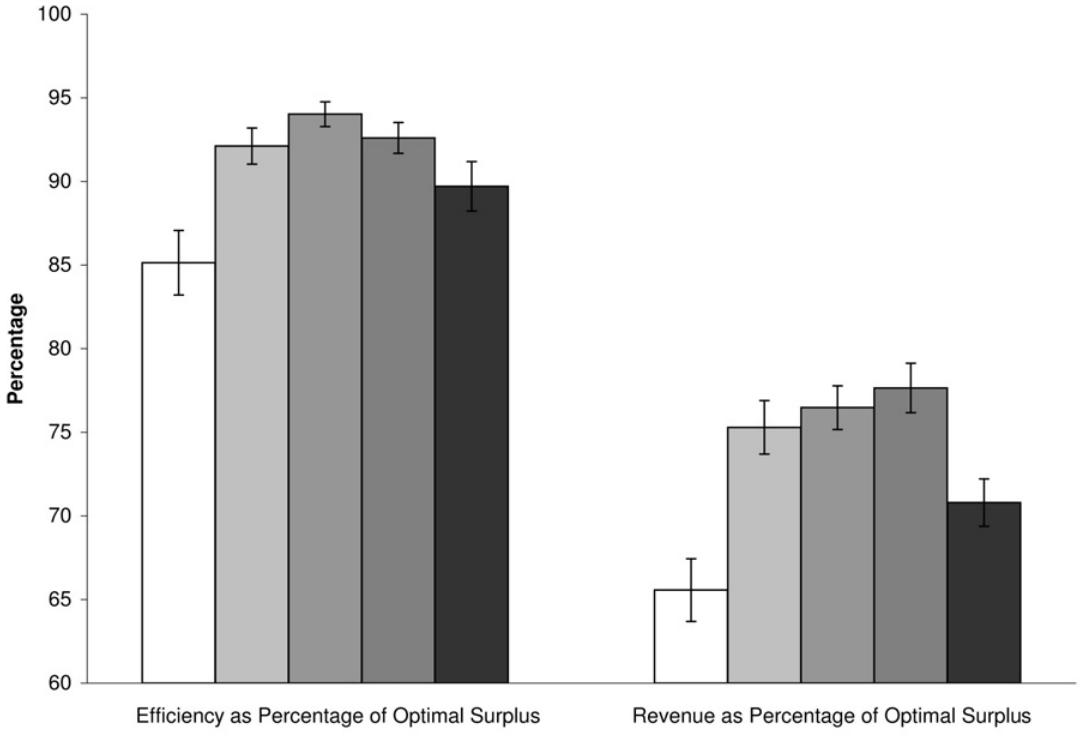
"open platform" that would be open to third party hardware and software providers (e.g. alternative providers of smart phones and their operating systems). The inclusion of this license provision was seen as a political victory for Google, a company with a lot to gain from this, in that it didn't operate a phone network, but had an active and growing interest in mobile computing and communication, which would be served by having a package of open access licenses with national scope. The FCC was persuaded that this was in the public interest, but was concerned that licenses encumbered with an open platform obligation might be unattractive to providers of phone service. They therefore set reservation prices for these licenses, and announced that if they were not met, then licenses for the unsold spectrum would be offered in a subsequent auction, without the open platform requirement. Against this background, Google announced its willingness to acquire a national package of licenses itself if necessary. [21]

Goeree and Holt's experiment looked at a "tiered" or "hierarchical" package bidding auction of the kind explored theoretically by Rothkopf , Pekec and Harstad (1998). At each stage the computation of the set of winning bidders is computationally simple (since only non-overlapping packages are considered), and in addition to determining the winning bid configuration, the auction produces prices for each individual license, designed to give bidders information on how much they might have to increase their bids to become part of the winning set of packages, and to help small bidders coordinate.

The experimental sessions each involved the auction of 18 licenses (denoted A through R), organized into either two or three tiers of packages. In the lowest tier were individual licenses, on a higher tier (in some sessions of the experiment) were a set of three predefined, non-intersecting "regional" packages of four licenses each, and on the highest tier was a "national" package consisting of the 12 licenses A through L. In each experimental session there were six "regional" bidders (1-6) and one "national" bidder (7). Each regional bidder had an interest in four adjacent licenses, not necessarily those in one of the pre-specified regional packages (and was permitted to acquire at most four licenses), and the national bidder had an interest in the twelve licenses in the national

[21] The primary sources for this story are the FCC Public Notice (DA-07-3415) and FCC Second Report and Order (FCC 07-132). A clear description, and an analysis of how the auction played out can be found in Brusco, Lopomo, and Marx (2009).

package, separately and as a package. Bidders' values for individual licenses were drawn from a distribution, and their value for the package they were eligible to bid on were scaled up to make the package more valuable as a whole. All bidders were eligible to bid on individual licenses, the national bidder could also bid on the national license, and in the sessions with regional packages (i.e. in the sessions in which there were three levels of the bidding hierarchy), either the even numbered or the odd numbered regional bidders were eligible to bid on a regional package. There were thus three conditions for the Hierarchical Package Bidding (HPB) mechanism—one with two levels and two with three levels. The other two conditions of the experiment used the same sets of values, and auctioned the licenses either via a simultaneous multi round auction (SMR) modeled on those the FCC has used for most of the spectrum auctions, or a Modified Package Bidding auction developed by the FCC, which allowed bidders to formulate their own packages (based on the Resource Allocation Design auction proposed and studied experimentally by Kwasnica et al. 2005). Each condition of the experiment was conducted in 5 sessions, with new values and different participants in each session.



SMR (white), HPB2 (light gray), HPB3-odd (medium gray), HPB3-even (dark gray), MPB (black). From Goeree and Holt (2010, p11)

Figure 1

The results of the experiment lent support to the use of Hierarchical Package Bidding not only from the point of view of simplicity, but also by showing that it might perform relatively well in terms of both efficiency and revenue (see Figure 1). And the detailed observation of particular transactions that is possible in the lab also gives some insight into why different auctions performed differently in the experiment. Goeree and Holt look at several of the auctions in which the Modified Package Bidding auction that allowed bidders to form their own packages did not select the efficient outcome, because of the threshold problem. They note:

"In each of these three rounds, the regional bidders were not able to coordinate a very strong response in the sense that their provisional winning bids left numerous provisionally unsold licenses… With fully flexible bidding, the regional bidders were bidding on "home-made" overlapping packages that did not 'fit' in the sense that the revenue maximizing allocation left unsold licenses, which made it easier for the national bidder to regain provisional winner status in the subsequent rounds."

Notice that this is a somewhat different manifestation of the "threshold problem" than had been the focus of much of the earlier discussion, that had mostly focused on free riding by small bidders. The issue observed in this experiment had more to do with coordination. Since the pre-specified packages are non-overlapping, they ease the coordination problem among small bidders, provided that the packages as specified are not too different from those that bidders want. The problem of specifying appropriate packages thus becomes part of the auction design.

Finally, note that aside from what we usually think of as the "results" of an experiment, the experiment also pointed up some design and implementation issues, some of them as simple as what features made it into early software, and some more complex.

> Holt writes ((5/29/2010 email) "an initial software implementation did not retain old bids in the database, which could create price

> cycles with declining bids, a potential disaster. For example, if people are bidding on license A and see a low price for B and switch to B, the the price for A might fall if prior bids on A were not retained, providing an incentive for bidders to switch back to A, etc. This cycling possibility was revealed by Jacob's software tests, and then it showed up in a classroom experiment that I did in a small honors class at Virginia. At that point, the commercial programmers changed the database structure to allow retention of old bids."[22]

As already mentioned, the FCC ran a package bidding auction in 2008. It involved only two levels of hierarchy, i.e. bidders could bid either on individual licenses, or on three pre-specified packages. Crawford, Kwerel and Levy (2008, pp190-91) report on the implementation and outcome, which, as it turned out, involved few package bids:

> "The 700MHz auction provided for "package bidding" on three pre-defined packages of licenses in the C Block: (1) the "50 States" package containing the eight Regional Economic Area Grouping (REAG) licenses comprising the continental U.S., Alaska, and Hawaii; (2) the "Atlantic" package containing the 2 REAG licenses comprising Puerto Rico, the U.S. Virgin Islands and the Gulf of Mexico; and (3) the "Pacific" package containing the 2 REAG licenses comprising U.S. Pacific territories. Under package bidding, bidders could place bids for individual spectrum "parcels" or for a package of parcels. If the highest individual parcel bids aggregated to more than the highest package bid, then the individual parcel bids would win; if not, then the highest package bid would win.
>
> As it turned out, there were few package bids during the auction. Ultimately, only a single package was won, the Pacific package. Google was the only party bidding on the 50 States package, and it stopped bidding when the reserve price was reached. Verizon bid individually on all 8 REAGs in the 50 States package and won all of them except the Alaska REAG.

---

[22] Holt further writes: "Another serious problem that the experimental testing of HPB identified was the need to weight licenses by some measure of size (e.g. population) in the calculation of prices. for example, suppose there is provisionally winning bid of of 100 on a package that includes NY, CT, and RI, with high bids of 60 on NY, 5 on RI, and 5 on CT. The "overhang" of 100 - 60 - 5 - 5 = 30 needs to be allocated to the three individual high bids so that they sum to 100. If the allocation is equal (ignoring the large population of NY), which was the way the FCC programming was going to do it at first, the new prices would be go to from 60 to 60 + 10 = 70 for NY and from 5 to 5 + 10=15 for CT and RI. These huge increases for the small licenses might cause small bidders who were only interested in those small licenses to drop out early in the auction, leaving the bidder with an interest in NY with no "friends" to help bid against the package bid. This equal allocation rule, therefore, could result in too big of an advantage for the large national bidders. This potential oversight did not come from the experiments, since for simplicity (the auctions were already complex enough) we had used licenses with equal weights. But thinking about the experiments and observing them helped us see past the simplicity of the experiment to avoid a disaster."

> While package bidding seems to have made little difference in auction 73 *ex post*, this could not have been known with certainty *ex ante*. For Google, with no existing spectrum holdings and seeking nationwide coverage, the availability of a nationwide package may have been important to their participation in the auction. By contrast, Verizon, as an incumbent with spectrum holdings nationwide, was not subject to the same "exposure" risk of failing to get complete nationwide coverage in the 700MHz band. Thus the availability of the 50 States package may have reduced exposure risk, facilitating new entry. *Further experience with package bidding is required for an understanding of its effects*." (emphasis added)

Looking back from our present vantage, the story of package bidding, and of the role of experiments in promoting it, is a complicated one. It is hard to evaluate the role that package bidding played in the 2008 FCC auction #73. None of the subsequent FCC auctions have allowed package bidding (at this writing, the most recent auction, #96 ran in January and February 2014).[23] But it is clear that package bidding auctions have been an important part of the *discussion* of the auction of complex goods, probably at least since the discussion of their possible use for airport takeoff and landing slots. And experiments played a very large role in this discussion.

The experiments themselves do not help us evaluate how big the effect of package bidding might be in allocating the various kinds of goods for which package bidding has been advocated. Rather, the experiments have been deployed as *demonstrations* that package bidding *could* make a difference, under some conditions.

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. The potential role for experiments is changing, as practical experience and new theory are developed.

Kagel, Lien, and Milgrom (2010) report an experiment comparing combinatorial auctions 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. They argue

[23] See http://wireless.fcc.gov/auctions/default.htm?job=auctions_all. In England, The British Office of Communications (Ofcom) has run spectrum auctions that allowed some package bids, in ways that appear to use some of the design proposals and theory contained in Ausubel, Cramton and Milgrom (2005) and Day and Milgrom (2007): see e.g. Ausubel and Baranov (2014).

that the space of potential combinations and valuations created by even the simplest experimental environment is much bigger than can be meaningfully explored without some theoretical guidance about where to look. They note that, to reach allocations that are in the core, or are as efficient as many that have been reported in experiments, it is necessary that bidders bid sufficiently aggressively on an appropriate set of packages. One way that bidders might do this is to bid on *every* profitable package, but the proliferation of packages quickly makes this impossible. They write

> "In problems of realistic scale, bidders cannot place bids on every package at every round, even if the rules permit that. Even in auctions with a limited number of items for sale, it is likely that bidders place bids only on a few packages. In such cases, for good outcomes to emerge from an experiment, the bidders must somehow ***identify*** the relevant packages and, in addition, must decide to bid aggressively on those packages."

To choose valuations on which to experiment, and to predict in advance the outcome of these experiments, they conduct simulations of very simple bidders, and adopt the hypothesis that

> "*Simulations in which automated bidders bid only for the currently most profitable package will lead to (near) core or efficient outcomes in the same environments where experimental outcomes lead to approximate core or efficient outcomes."*[24]

I will not review their experiment in detail, but in conclusion they write

> "In principle, one way that bidders might bid aggressively enough on relevant packages is to bid equally aggressively on ***all*** packages, but that is not what we find. Bidders in our experiment typically bid on just the one or two most profitable packages and those packages often remained unchanged for many rounds during an auction. In our data, consistent with our theory, standard package auctions yield efficient allocations and core-level revenues most frequently when the packages that are selected by this sort of behavior are the relevant ones…
>
> Our finding that price-guided auctions can fail to direct bidders to relevant packages early enough in the auction suggests possible improvements to the auction design. One possible refinement is to make relevant bids more likely by making it easier to bid on more sets of

[24] Recall Rassenti, Smith, and Bulfin's conjecture about complexity leading to full demand revelation. The Kagel et al. conjecture is somewhat similar in spirit, but boundedly rational, in the face of the complexity inherent in the proliferation of possible packages.

> licenses. That might be accomplished by implementing a richer bidding language than the XOR language of our experiment."

Thus this paper and a subsequent one (Kagel, Lien and Milgrom 2014) point to the interaction between the auction design and bidder behavior in connection with how much help the bidders get in directing their attention to particularly relevant packages. This help could of course come from a number of sources.

In summary, if I had written this section on FCC auctions in early 2008, it would have been tempting to conclude on a triumphant note: after years of experiments promoting package bidding, the FCC had finally implemented a limited version of it. In view of the FCC's subsequent return to auctions without package bidding, a more sober assessment may be called for. But today new combinatorial auctions are in use in Europe, and so experience continues to accumulate. In any event, there's a lot we can learn from the very important role that experiments played so far, and the greater scope that they may have in the future as advances in the theory of auctions, and experience with existing auctions permit more focused experiments.

We turn next to the design of auctions by eBay and by Medicare, and after that to the design of labor market clearinghouses. For at least some questions in each of these areas there was already a good deal of theory and empirical evidence available when economists were called to aid in design, and so experiments could play a more targeted role that involved hypothesis testing as well as demonstration.

4. Other auctions

4.1 eBay auctions

Since shortly after eBay opened for business in 1995 (or at least since its annual merchandise volume first exceeded a hundred million dollars and it went public in 1998), it has been a source of data for the study of auctions, and of ecommerce generally. Because its auctions are available to anyone with an internet connection, eBay makes it

possible to gather data to investigate a wide range of hypotheses. Early investigators gathered data from individual auctions, but lately eBay has made vast amounts of data available to researchers.[25] I will recount here, however, two lines of investigation for which even the copious data now available from actual transactions are insufficient to clearly determine what was being observed, and in which laboratory experiments were able to supplement field data in critical ways (see also the survey of Ockenfels, Reiley, and Sadrieh (2007).

The first experiment I'll describe arose from an investigation into how eBay's rule for ending auctions influences the observed distribution of bids over time. The second experiment I'll discuss arose from an effort to understand and then redesign eBay's system of recording feedback after transactions are completed, which allows sellers and buyers to establish reputations. That experiment was needed not only to understand better what was being seen in the field data, but also to get a first look at the proposed new reputation system, for which no field data yet existed. (The first of these experiments thus falls into the "design as a noun" tradition of experiments that help us understand how particular aspects of a market's design affect its performance, while the second falls clearly into the emerging tradition of "design as a verb.")

4.1.1 eBay auction rules

In eBay's early days, Amazon.com ran an auction site very similar to eBay (in addition to Amazon's main fixed price sales site, which continues to thrive). Both sites ran second-price auctions, typically for a week, with a clearly announced end times.[26]

[25] One notable combination of these very large data sets with experiments is the search by Einav et al. (2013) for "seller experiments" found in the eBay data, which they interpret as experiments conducted by eBay sellers, who may sell identical items with different reserve prices, or required minimum bids, or shipping costs, auction duration, use of the"buy it now" option, and so forth. They identify hundreds of thousands of such experiments, and generally find significant effects on prices and the probability of sales that are in many cases smaller than the similar effects found in field experiments reported elsewhere in the literature.

[26] Bidders could bid in real time, or submit a reservation price (called a proxy bid) early in the auction and have the resulting bid register as the minimum increment above the previous high bid. As subsequent reservation prices are submitted, the bid rises by the minimum increment until the second-highest submitted reservation price is exceeded. Hence, an early bid with a reservation price higher than any other submitted

The one respect in which their auction rules were different concerned how an auction ended. eBay auctions ended precisely at the announced time, a "hard close." Any bids that arrived afterwards were not accepted. Amazon auctions, however, employed a "soft close," and ended at the initially announced time only if no bids had arrived in the ten minutes prior to that close. Otherwise, Amazon auctions were extended beyond the initially announced closing time, and ended only when ten minutes had passed since the arrival of the last bid.

Roth and Ockenfels (2002) observed that, despite the fact that most eBay auctions lasted for a week, the last bids in many auctions arrived in the final minutes or seconds before the end time. (Placing bids in the last moments of an auction is colloquially referred to as "sniping.") They also observed that, for a variety of reasons, eBay's hard close might make sniping a rational behavior, i.e. in a variety of circumstances it could be a best response to the behavior of other bidders, despite the fact that late bids sometimes fail to go through.[27] They further noted that, if strategic behavior caused by eBay's hard close was the cause of so many late bids, then there should be a big difference in the timing of bids on eBay and Amazon auctions. That is, if many late bids in eBay were motivated by the fact that other bidders would not have time to respond, then late bids should be less common in Amazon auctions in which they would cause the auction to be extended so that other bidders could respond.[28]

Examining transaction data from eBay and Amazon auctions, they showed that late bidding was much more common on eBay. For example, more than two-thirds of the

---

during the auction will win the auction and pay only the minimum increment above the second-highest submitted reservation price. eBay has since added additional selling formats.

[27] Roth and Ockenfels noted that, among the reasons that bidders might bid late are desire to conceal information in common value auctions, and, even in private value auctions, desire to avoid price wars either with rational bidders or with bidders who bid incrementally for any reason. Surveys of late bidders showed that there are two sources of risk involved in late bidding. One was that bidders who plan to bid late sometimes find that they are unavailable at the end of the auction. The other involves bidders who are attempting to bid at the last moment but who do not succeed due to, e.g., erratic internet traffic or connection times. See Ockenfels and Roth (2006) for some formal modeling.

[28] An alternative hypothesis is that late bids are largely due to nonstrategic causes like simple procrastination, or to a desire to retain the ability to bid on other items for sale, or to search tools that present first those auctions that end soonest, in which case late bids might be equally prevalent in both eBay and Amazon auctions.

eBay auctions in their sample had bids submitted less than an hour before the scheduled end time, in contrast to less than a quarter of the Amazon auctions. In the last 10 minutes, only 11 percent of the Amazon auctions received bids (i.e. only 11 percent of the Amazon auctions were extended past the scheduled deadline), while more than half the eBay auctions received bids in the last ten minutes (and over 10 percent of the eBay auctions received bids in the last ten seconds). And not only were late bids vastly more common on eBay than on Amazon, but more experienced bidders (as measured by their feedback scores) tended to bid late more often on eBay, but less often on Amazon.

Of course eBay and Amazon auctions differed in other ways than just their rules: eBay had many more items for sale than Amazon, and many more bidders. Furthermore, buyers and sellers themselves decide in which auctions to participate, so there might be important differences among the buyers and sellers and objects offered for sale on eBay and Amazon. Some combination of these uncontrolled differences between eBay and Amazon might be the cause of the observed difference in bidding behavior, instead of the differences in auction rules. A laboratory experiment therefore offered the chance to look at differences in ending rules under controlled conditions.

Ariely, Ockenfels and Roth (2005) reported an experiment on second-price auctions that differ only in the rule for how the auctions end. Subjects were randomly assigned to each auction type, so there were no systematic differences in bidder characteristics across auctions, and the number of bidders per auction was kept constant. Each bidder in the experiment participated in a sequence of auctions, allowing learning to be observed, as bidders gained experience with the auction environment.[29] The goods offered were artificial, independent private-value commodities (each bidder was given a redemption value he would be paid in cash if he won the auction, and these values were

[29] Another benefit of a laboratory experiment is better control of the effect of experience, because the proxies for experience in the field data ("feedback ratings") used by Ockenfels and Roth (2006) are imperfect. For example, feedback ratings only reflect the number of completed transactions, but not auctions in which the bidder was not the high bidder. In addition, more experienced buyers on eBay may not only have more experience with the strategic aspects of the auction, they may have other differences from new bidders, e.g., they may also have more expertise concerning the goods for sale, they may have lower opportunity cost of time and thus can spend the time to bid late, or they may be more willing to pay the fixed cost of purchasing and learning to use a sniping program.

drawn independently of the values of other bidders), so that bidding behavior in the experiment could be compared with reservation prices in a way not available in field data. (Private value goods were chosen to avoid the additional strategic issues involved in auctions in which bids reveal information about the value of the object, and to investigate the effects of the hard close rule in the simplest case.)

The treatments included four auction types (described in detail below): sealed bid, Amazon, eBay.8, and eBay1; the latter two treatments differed only in the probability that a "last minute" bid would be transmitted (80 percent in eBay.8 and 100 percent in eBay1). There were exactly two competing bidders in each auction. Each bidder in each auction was assigned a private value independently drawn from a uniform distribution between $6 and $10. The winner of an auction received his private value minus the final price, and a loser received nothing for that auction. The final price was determined by the second price rule that the bidder who submitted the highest bid won and paid (at most) a small increment ($0.25) above the next highest bid. If only one of the bidders bid, the price was the minimum bid of $1.[30] All auctions were run in discrete time, so that "bidding late" would be well defined without complications of continuous time decision making, such as individual differences in typing speed, which might differentially influence how late some bidders can bid.

About these experimental design choices, Ariely et al. note the following:

"*Because eBay and Amazon are online auctions, it would have been possible to conduct the auction using precisely the eBay and Amazon interfaces, had that been desirable, by conducting an experiment in which the auctions were on the internet auction sites (for a classroom demonstration experiment of this sort, in a common value environment, see Asker et al., 2004, and for a private value auction study along these lines see Ockenfels, 2004). This would not have served our present purpose as well as*

[30] As in internet auctions, the price never exceeds the highest submitted bid: If the difference between the highest and the second highest submitted bid is smaller than the minimum increment, the price paid is equal to the highest bid. If both bidders submitted the highest bid, the bidder who submitted his bid first is the high bidder at a price equal to the tie bid. If identical bids are submitted simultaneously, one bidder is randomly chosen to be the high bidder. Also, a bidder can bid against himself without penalty if he is the current high bidder, because it raises his bid without raising the price.

the discrete version described here. In this respect it is worth noting that what makes an experimental design desirable is often what makes it different from some field environment, as well as what makes it similar."

The experimental conditions can all be thought of as variants of the eBay.8 condition, which consisted of two kinds of bidding stages, stage 1 (early) and stage 2 (late). Stage 1 was divided into discrete periods. In each period, each trader simultaneously had an opportunity to make a bid. At the end of each period, the high bidder and current price (typically the minimum increment over second highest bid) were displayed to all. Stage 1 ended only after a period during which no player made a bid. This design feature ensured that there was always time to respond to a bid submitted 'early' in the auction.

Stage 2 of the eBay.8 auctions consisted of a single period. The bidders had the opportunity to submit one last bid with a probability $p = 0.8$ of being successfully transmitted.

In the eBay1 condition, the probability that a bid made in stage 2 would be transmitted successfully was $p = 1$, i.e. stage-2 bids were transmitted with certainty. Everything else was as in eBay.8.

Similar to the eBay.8 condition, in the Amazon condition stage 1 was followed by stage 2, and the probability that a stage-2 bid would be successfully transmitted was $p = 0.8$. However, a successfully submitted stage-2 bid restarted stage-1 bidding (followed by stage 2 again, etc.). Thus, in the Amazon condition, the risk of bidding late was the same as in the eBay.8 condition, but a successful stage-2 bid caused the auction to be extended.

In the sealed bid condition, the auction began with stage 2 (with $p = 1$), and ended immediately after, so that each bidder had the opportunity to submit only a single bid, without knowing the bids of the other bidder. While the sealed bid auction obviously

could not yield any data on the timing of bids, it provided a benchmark against which behavior in the different auctions could be assessed.

As in the internet counterparts, bidders in the eBay and Amazon conditions were always informed about current prices as the auction progressed, but the magnitude of the high bidder's current bid was not revealed to the low bidder.

These experimental games were intended to reproduce the pricing and information policies employed by Amazon and eBay on the internet, and capture the essential differences in ending rules. There was sufficient time to submit bids and respond to others' bids early in the experimental conditions (in stage 1), and the hard close in the eBay treatments did not allow bidders to respond to very late (stage 2) bids. The risk involved in submitting late bids in the eBay.8 condition reflects the fact that late bids run the risk of being lost in internet auctions. Successfully submitted late bids in the experimental Amazon condition automatically extended the auction (that is, moved the auction back to stage 1), giving other bidders sufficient time to respond to all bids. However, late bidding on Amazon faced the same risk as late bidding on eBay.8. Finally, as in eBay and Amazon auctions on the internet, the second price rule allowed a bidder in the experiments to bid by proxy.

Figure 2 graphs the percentage of bidders who placed a bid in stage 2, over time, as bidders experienced more auctions. These numbers can also be interpreted as the probability that a bidder would make a stage 2 bid (in the Amazon treatment, only one stage 2 bid is counted per bidder).

The experimental results reproduced the main observation from the field data: There was more late bidding in the fixed-deadline (eBay) conditions than in the automatic extension (Amazon) condition, and, as bidders gained experience, they were less likely to bid late in the Amazon condition, and more likely to bid late in the eBay conditions (and this is even clearer in eBay1, where there was no risk that late bids would be lost, so there was no cost to bidding late). In all the conditions, the number of bids in the early periods also declined; and in the eBay conditions, not only did the percentage of

late bids go up, but so did the magnitude of the price changes due to late bidding, as bidders learned to hold their fire until their final bid. So, as bidders gained experience with each auction type, and the eBay bidders learned to bid late, while the Amazon bidders learned to bid early, the prices at the end of the first stage became an excellent predictor for final prices in the Amazon auctions, and a very poor predictor of final prices in the eBay auctions.

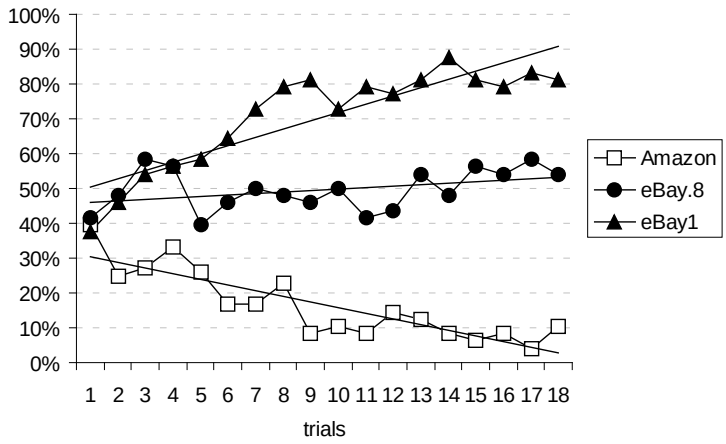


Figure 2: the percentage of bidders who placed a stage-two bid

Because these observations are experimental, we can also see things not available in field data, such as the relationship between the bids and bidders' valuation of the object for sale, (which in the experiment can be taken to be the amount they will be paid if they win the auction). And in the laboratory, unlike in the field data, we can see not only the bids that were successfully placed, but also attempted late bids that failed to go through.

Figure 3 shows the median bids, by round, as a percentage of each bidder's valuation. There are two outliers: in the early rounds, the median bidders in the sealed bid auctions bid much less than their valuations, and in the late rounds the final bids in the Amazon auction don't converge to 100% of bidders' valuations.

The Amazon results reflect the fact that, as bidders learned not to bid in stage two, stage one became an ascending value English auction which determined the final price. So, the

bidder with the higher valuation could stop bidding as soon as his bid exceeded the valuation of the second highest bidder. (Since there was no chance of bids being lost, the Amazon auctions returned the highest revenue in this experiment.) That is, a bidder on Amazon who was currently the high bidder had no incentive to increase his bid unless he was outbid, at which point he always had the opportunity to raise his bid. So once his bid exceeded the other bidder's value, he had no incentive to increase his bid to his own value.

The second price auction results are more at odds with conventional theory, since in that auction bidders have (essentially) a dominant strategy to bid their valuation (up to a very small adjustment having to do with the minimum increment by which the final price exceeds the second highest bid, only if the highest bid is high enough). But, unlike the dynamic auction conditions, a bidder in the sealed bid auction who mistakenly believed he could win the auction by submitting a low bid did not learn his mistake until the auction was over. In contrast, in the eBay and Amazon conditions, a bidder who started with a low bid learned that his bid was too low in time to raise his bid before the auction ended. So, in all the auctions, bidders could learn from experience to bid up to their valuations, but in the sealed bid auctions this experience came only after some auctions had been lost, while in the other auction formats learning could go on while the auction was still in progress.

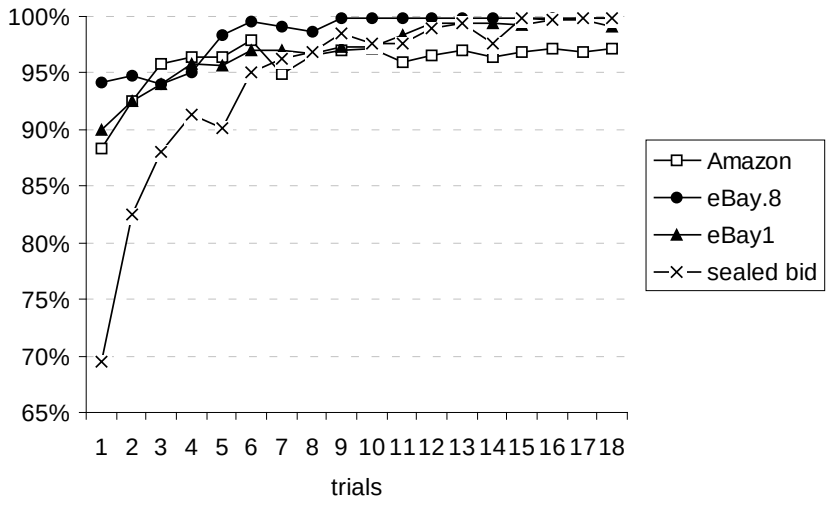


Figure 3: Median of final bids (including lost stage-2 bids) as a percentage of value[31]

The fact that participants have to learn how markets work is an important feature of market design, since, particularly when a novel design is introduced, no participants will have had experience with it. Experiments have therefore helped make clear that alternative designs cannot simply be compared by comparing their equilibrium properties.

To go back to the relationship between the experimental results and the field observations, it is worthwhile to consider both which results of this experiment help us better understand what we have seen in the field, and which results of this experiment are less likely to generalize to the field environments we are interested in. To put it another way, the "external validity" of particular results may vary, and has to do with (among other things) decisions made in the design of the experiment.

The experiment was designed to see if the difference in late bidding behavior already observed in the field data could be caused by the differences in the eBay and Amazon auction rules, as opposed to being entirely determined by other differences e.g. in the number and identity of those auctions' participants. By observing the same bidding differences in the laboratory, in a fixed subject pool, when the only differences between experimental treatments are the auction rules, the experiment allows us to conclude that the hard close rule for ending the auction does indeed cause more sniping than the soft close rule, particularly as bidders gain experience.

It is less clear that the learning behavior among inexperienced bidders would be as important in the field as in the lab, since in the field there might be other sources of advice that would enable bidders to quickly learn the properties of the second price sealed bid auction, for example. But the results I would have the *least* confidence in trying to generalize from this experiment to the field would be the revenue comparisons. It is natural that the hard close has slightly lower revenues in this experiment, since it

[31] Since we are looking at median bids, the graph doesn't show the persistent overbidding by some bidders discussed in Kagel and Levin's chapter.

encourages late bidding, and late bids have a positive probability of being lost. But the lost bids have such a clear relation to revenue in the experiment because there are always exactly the same numbers of bidders in each auction, regardless of the rules. This was a design decision that allowed us to control for the possibility that the greater number of bidders on eBay was the cause of sniping, not the auction rules. But, by the same token, this experiment doesn't allow us to investigate if the auction rules might influence the number of bidders, in ways that might increase revenue. For example, maybe on the internet sniping makes auctions more exciting, and attracts more bidders, or attracts more informed bidders and hence different kinds of sellers, in a way that makes up for the revenue lost by late bids that fail to go through.[32]

My point here is that the same features of the experimental design that allow us to draw strong general conclusions about some aspects of the experiment may make other parts of the experimental results difficult to generalize away from the specific experimental environment. The likelihood of being able to generalize conclusions to other environments depends on what the experiment controls for, and what it does not. In this case, controlling the environment so that everything remains constant except the auction rules allows us to determine the effect of those rules on bidding behavior, thanks to the kind of control that the laboratory offers. But it does not allow us to see how those rules might affect elements of the environment that were simply controlled by the experimental design.

Note that there's much more to eBay than the design of each individual auction. eBay has created a marketplace to which many buyers and sellers come, sometimes repeatedly. The same item may be offered by many sellers, while a given seller may offer a variety of

[32] For example, a field experiment on bidding conducted by Ely and Hossain (2009) involving newly released DVD's auctioned on eBay reported that sniping only produced small changes in final price. In general, the easy accessibility of internet auctions has made them an excellent place to do field experiments (as well as to simply collect field data). But established internet auctions don't as easily lend themselves to *market design* experiments, since most of the design decisions have been taken by the auction provider. However some aspects of the design of each particular auction are sometimes left as choices for the seller, and so there have been field experiments focused on these, in which the experimenters sell identical goods online under different auction rules. For example, Hossain and Morgan (2006) and Brown, Hossain and Morgan (2010) focus on the revenue effect of how much of the auction price is framed as a shipping cost, while Ariely and Simonson (2003) study the effect of varying the minimum allowable initial bid.

goods for sale. But unlike the case of sellers who operate physical stores, it is hard for buyers to keep track of sellers who are known only by their username. So it may be hard for buyers to distinguish trustworthy and reliable sellers from those who are less so, and the resulting lack of trust may be an obstacle to commerce. We turn now from eBay in the small to eBay in the large, and consider how eBay initially sought to solve this problem by designing a feedback mechanism, and how it recently redesigned that mechanism, with the aid of experiments and experimenters, led by Axel Ockenfels.[33]

4.1.2 eBay's Reputation Mechanism

The original eBay feedback system, set up before the introduction of convenient online payment mechanisms (and hence when there was an issue of trustworthiness for buyers as well as for sellers), was meant to allow both sides of a transaction, buyers and sellers, to leave feedback on each other that would be available to future potential transactors. The initial feedback rules, which involved leaving both a positive-neutral-negative rating and a text comment, underwent some modifications based on experience. It eventually settling down to a system in which feedback was identified by the username of the person leaving it, and only the winning bidder and the seller could leave feedback about one another, so ratings couldn't be easily influenced by multiple feedbacks from the same individual.

By the time eBay commissioned the study of their reputation system reported by Bolton, Greiner, and Ockenfels (2013), there was growing concern that the reputation system might not be providing reliable information about the quality of transactions. Paradoxically, the concern grew out of the fact that the overwhelming majority of feedback resulting from transactions was mutually positive feedback from both buyer and seller. This was despite the fact that chat groups and other channels of communication made clear that some non-negligible proportion of transactions experienced problems, and were the source of considerable dissatisfaction.

[33] In addition to the work described below, Ockenfels writes (personal communication, 7/3/10) that he engaged in some experimental work to help eBay evaluate and redesign its multi-unit "Dutch" auction format that it has since abandoned (see also Kittsteiner and Ockenfels (2008).

Bolton, Ockenfels, and Greiner (henceforth BGO) had at their disposal a great deal of feedback data from eBay transactions, and one of the things the data revealed was a strong reciprocal pattern of feedback. The majority of feedback was mutually positive, with the seller giving the buyer a positive review *after* having received a positive review from the buyer. Sellers relatively rarely left feedback before buyers, and in the very small percentage of cases in which buyers gave "problematic" (neutral or negative) feedback, it was quickly followed by problematic feedback about the buyer from the seller. Together with the fact that some feedback came only weeks or months after the transaction, it thus appeared that while the feedback system might possibly be playing a part in post-transaction dispute resolution (e.g. "I'll leave negative feedback unless you replace the broken part…"), the final, reciprocally positive nature of most feedback gave little information about the level of satisfaction with each transaction.

By this time, most buyers were paying in advance by credit card, so the need for sellers to be able to evaluate the trustworthiness of buyers had diminished.[34] So one way to try to make feedback more informative, by preventing it from being simply a reciprocal exchange of favors, would be to eliminate seller feedback about buyers, so that the only feedback would be about the seller in each transaction. Another proposal was to make reciprocation—and retaliation—harder by making feedback anonymous. (As a practical matter, feedback would be kept "blind" by making it both anonymous and essentially simultaneous—there would be a period of time after a transaction in which feedback could be left by both buyer and seller, and feedback would only be published after this period had ended.)

As it happens, there were some field data suggesting that both one-sided feedback and "blind" feedback system might result in a higher frequency of negative feedback. The German site of Amazon, Amazon.de effectively had a one-sided feedback system, and it showed more negative feedback than eBay's system. eBay itself had a blind feedback system in Brazil, where it had purchased the *MercadoLivre* marketplace and kept its

[34] This was not the opinion of at least some sellers in user discussion groups, who felt that being able to respond to buyer feedback was their only defense against extortion from unreasonable buyers who threatened to give them negative feedback.

feedback system in place, which produced substantially more negative feedback than eBay's other country sites. And the software marketplace RentACoder had changed from something like eBay's conventional feedback to a blind system, and the correlation between buyer and seller feedbacks had dropped after the change.

But, of course, there might be other differences between Amazon and eBay, and between Brazil and other countries that could account for negative feedback, and what happened in the RentACoder market for software might not be attributable only to their change in feedback system, or might depend on it in a way peculiar to the market for custom software. And the field data did not reveal how differences in the amount of negative feedback influenced the efficiency of the resulting transactions, whose critical details (such as the value of the transaction to the parties) were invisible

In addition, the question of how to modify the feedback system to make it more informative was constrained by the desire not to harm the thriving market that eBay had created with what had become its conventional feedback system. The discussion therefore focused on the potential effects of *adding* additional blind feedback to the existing conventional feedback. There weren't any field data available on how such a combined system might work.

For all of these reasons, there was room for a controlled experiment.

BGO report that, as part of their analysis, they conducted a laboratory experiment that consisted of a three-stage transaction, among cohorts of 3 potential buyers and 1 seller (whose roles rotated between rounds). Each buyer i had a private valuation of v_i Experimental Currency Units drawn independently from a uniform distribution on the integers in the interval [100, 300].

The first stage of each transaction was an eBay style second-price auction that determined the winning bidder and the price he or she would pay to the seller, and informed all parties of the final price p and all but the highest bid. In the second stage (once the price had been determined), the seller decided on the *quality* of the good, a number q between 0 and 1. This determined the winning buyer's payoff from the

transaction, $qv_i - p$, and the seller's payoff, p-100q. (Notice that since v_i is greater than 100, providing quality is always more valuable to the buyer than it is costly to the seller, and so it is efficient for the seller to provide the highest quality.)

Finally, the third stage of the transaction was the feedback stage, in which the experimental variable was that one of three feedback systems was used, in each of which the players could leave feedback for a small charge. There were three forms of feedback used as treatment variables. First, as a baseline against which to compare possible changes was CF – Conventional Feedback, in which buyer and seller could each rate the other as positive, negative or neutral, in (up to) two stages with a soft close (so that the buyer and seller each have the opportunity to reply with feedback if the other gives feedback in the first stage). Next was CF + DSR – Conventional Feedback plus Detailed Seller Ratings: in addition to conventional feedback, there was an opportunity for buyers to report a 5-point quality rating on sellers that was not revealed until after the CF ratings are closed. The final feedback treatment was Blind CF, which consisted of only one feedback stage, so that feedback was simultaneous, and each party had to choose whether to leave feedback, and what feedback to give, without seeing his counterparty's feedback.

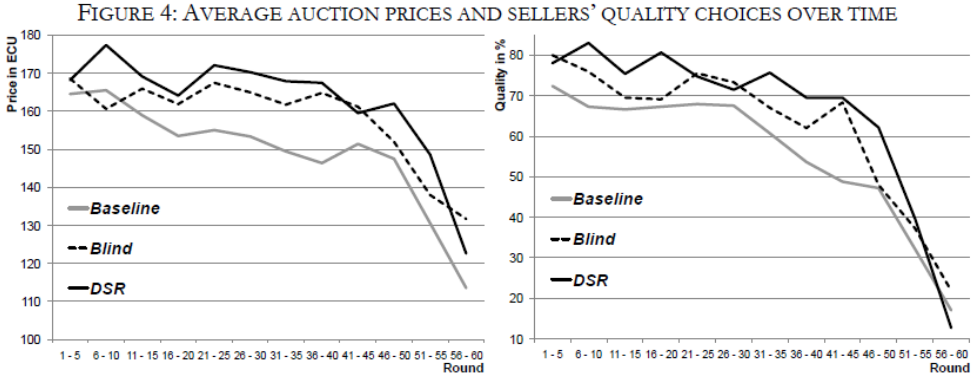
Groups of 8 subjects were randomly matched into groups of one seller and three potential buyers, with each subject playing the role of seller twice every eight rounds. Feedback was aggregated across an individual's roles as buyer or seller (as on eBay). So feedback provided a noisy signal about how much quality a seller had provided in previous transactions. The goal of the experiment was to see if the novel feedback systems would provide more accurate information about a seller's past behavior, and by doing so give sellers more incentive to provide high quality by encouraging buyers to bid more in auctions whose sellers had good feedback.

In terms of feedback, the results of the Conventional Feedback treatment mirror the reciprocal patterns observed on eBay. Moreover, the experiment showed that, when the seller provided very high or very low quality, feedback was uniformly positive or negative under all feedback treatments. But when the seller provided intermediate level quality (from .6 to .8), the feedback remained quite positive under the conventional

feedback treatment (in which the threat of retaliation was quite real), but not under the other two feedback treatments. (The correlation between buyer and seller feedback also dropped, as the opportunity for effective retaliation was removed.) And sellers responded to the more accurate feedback by providing higher quality, and were rewarded with higher prices, see Figure (4 in BGO)

Fig 4 about here,from BOG

FIGURE 4: AVERAGE AUCTION PRICES AND SELLERS' QUALITY CHOICES OVER TIME



Like the experimental study of eBay's auction rules, this experimental design also abstracts away from factors that may be important in the field. For instance, the experiment did not analyze how the transition from one feedback system to the other might influence the use of the new system; the underlying assumption being that transitional turbulence would wash out in the long-run. Also, the experiment gave buyers no opportunity for bad behavior, so that feedback given by sellers to buyers was not informative. Instead, the experimental design focused on isolating the effect of different feedback systems on the reciprocal relationship between traders when producing reputations - the phenomenon that was hypothesized to be a key driver for distortion of feedback. What the experiments demonstrate is that the design of the feedback system is sufficient to cause the empirical patterns observed in the field in a variety of different markets (Amazon, MercadoLivre, RentACoder), and that reducing the opportunities for reciprocation and retaliation could increase efficiency..

Based in part on these results, eBay decided to add Detailed Seller Ratings to its marketplace, which it did on a pilot basis in several of its country specific markets in early 2007 and worldwide later that year.[35] eBay data since then show no reduction in the frequency of feedback (and no changes in the nature of the conventional feedback), but the anonymous, one-sided detailed seller ratings now show more negative ratings.

Thus it appears that the change in eBay's feedback system has the effect on feedback that would have been predicted on the basis of the experiment, and roughly the effect that other internet marketplaces had experienced. Because the reputation system plays such an important role in eBay's marketplace, its design is likely to be subject to continued adjustment.[36]

4.2 **A poorly designed auction (for Medicare supplies)**

While many experiments are motivated by the desire to investigate why some auction or other market design works well, sometimes the job of market designers is to explain why some existing or proposed institution works poorly, and experiments can help here too. This was the case with a procurement process run by the U.S. Centers for Medicare and Medicaid Services (CMS) to allocate contracts for Medicare supplies.

Peter Cramton became a vocal critic of this process, which CMS refers to as an auction, and he and Ian Ayres summarized some of the criticisms in a Freakonomics column that ran as an op-ed in the New York Times under the title "Fix Medicare's Bizarre Auction Program" (Ayres and Cramton, 2010).[37] They pointed out that the procurement process

[35] eBay had this to say on its American site "Because detailed seller ratings are anonymous, sellers can't see which buyer gave them which rating. This means that buyers should feel free to be honest and open about their buying experience, and sellers can get a more complete picture of their performance." http://pages.ebay.com/help/feedback/detailed-seller-ratings.html. Starting in May 2008 eBay also removed sellers' ability to leave negative feedback: "Buyers can leave a positive, neutral, or negative rating, plus detailed seller ratings. Sellers can leave a short comment and positive ratings only for their buyers." http://pages.ebay.com/help/feedback/questions/leave.html.

[36] When I visited eBay in May 2014 I heard that they were contemplating doing away with the anonymity of the detailed seller ratings.

[37] See also the papers and congressional testimony compiled at http://www.cramton.umd.edu/papers/health-care/

adopted by CMS in response to a Congressional requirement to use auctions was not in fact an auction in any ordinary sense, and was unlikely to either reduce costs or promote efficiency.

Two features made the CMS procurement process hard to view as an auction. First, although bids from potential suppliers were used to set the price that Medicare would pay, these *bids were not binding commitments* on the part of the bidders who made them. Second, as Ayres and Cramton write,

> "*As is standard in multi-unit procurement auctions, bids are sorted from lowest to highest, and winners are selected, lowest bid first, until the cumulative supply quantity equals the estimated demand. Non-standard is that the current system sets reimbursement prices using the median of the winning bids rather than using the clearing price. Since most providers are small, they lack the resources to invest in information and strategy in preparing bids. For them an effective and easy strategy is the low-ball bid, as any one firm's impact on price is negligible*."

That is, since any single bid is likely to leave the median price little changed, and since bids are not binding, a bidder who bids a very low price makes himself eligible to be one of the auction winners (since he will have one of the lowest bids), without obligating himself to sell at the median price if that price is too low for him to make a profit, and without having to worry that he lowered the price from what it would have been if he had bid more (but still below the median).

Thus there is ample reason to believe that this procurement process will not achieve the efficiency goals for which an auction is normally used, since the bids may not be closely related to costs, nor will the price at which Medicare purchases supplies (the median bid) be closely related to the bids of the winning (low) bidders.

Merlob, Plott and Zhang (2012) reported an experiment that compared this "median-bid procurement auction with nonbinding bids," with a more standard "excluded bid" auction (with binding bids) in which the low bidders win the auction and are paid a price equal to

the lowest losing bid. In the case in which suppliers have only a single unit to sell, which was the case in this experiment, it is easy to see that the excluded bid auction makes it a dominant strategy for the bidders to bid their true costs, and that the winning bidders are therefore expected to be those with the lowest costs. Merlob et al. proposed to compare these two auctions. In their introduction they note that their experimental environment is considerably simpler than the environment in which suppliers bid to sell to Medicare, but that

"*Auction architectures performing poorly in simple cases studied experimentally provide a realistic warning about problems that can surface in complex cases.*" (p794).

That is, they anticipate that their experiment may serve to demonstrate some of the pitfalls of the Medicare procurement process, even in the absence of a fully developed theory for procurement processes of this design.

And indeed it does. They find that in the excluded bid auction suppliers tend to reveal their costs, and that the resulting prices are approximately competitive, and the outcomes are efficient, with the lowest cost suppliers winning. In contrast, in the Medicare procedure suppliers with high costs submit low bids, the resulting prices are lower than the competitive price, and the outcome is inefficient both in terms of who wins the auction and how much is supplied.

Given the early history of experiments in market design, we should perhaps not be too surprised that the effect of this experiment has been far from immediate. So far, Medicare continues to use this procurement process. But the experimental results make clear that there are good reasons to believe that this process does not serve the purposes intended by Congress when it mandated that Medicare develop an auction process to purchase medical supplies. And I anticipate that as the debate continues, experiments may make this point clearly to policy makers who may not find more theoretical arguments as accessible or as persuasive.

5. Labor Market Clearinghouses

Designing labor markets for doctors

Experiments have played an explicit role in the design of two medical labor markets in which I have been involved. 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).

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 similar 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). 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: Stable and Unstable Centralized Clearinghouses

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[38]	yes	yes
• Clinical psych[39]	yes	yes
• Lab experiments	yes	yes
• "	no	no

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

[38] First used in 1997-98.
[39] First used in 1999.

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 $15 (plus or minus an individual payment of not more than $1) if they matched to a high productivity partner, and $5 (plus or minus an individual payment of not more than $1) if they matched to a low productivity partner. So the high productivity agents on each side of the market were the top choices of everyone on the other side of the market, but no one knew how a particular individual would order them, because of the individual payments.

In each market 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.

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. Firms therefore 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 (i.e. ten rounds each consisting of a three-period market), a centralized matching technology was introduced for period 0 of markets 11 through 25 (periods -2 and -1 of those markets 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.

Figure 5 shows that the experimental results reproduce what we see in the field: after the market has unraveled, the introduction of the stable matching mechanism in market 11 reverses unraveling, the unstable one does not. (The figure shows the costs the players paid when they matched early, in periods -2 or -1, so costs are higher when matches are made earlier.)

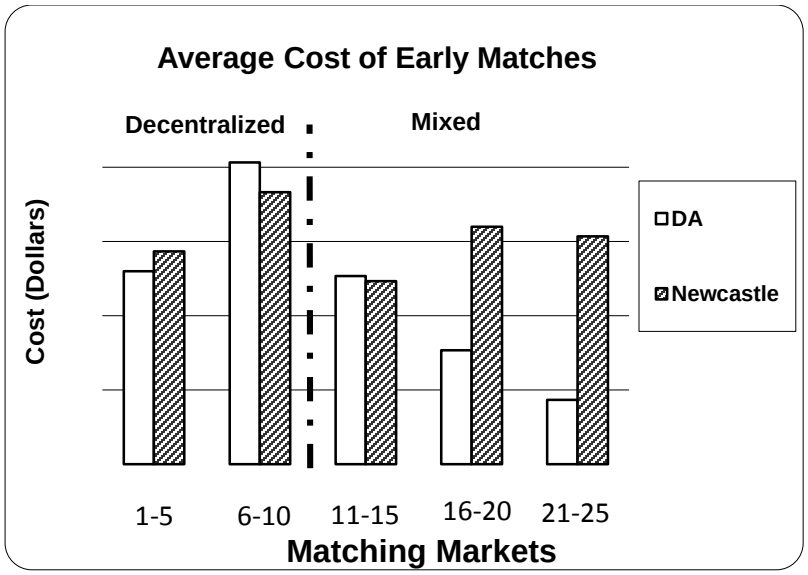


Figure 5: Average cost of early matching, by market number (round)

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 is particularly clear in Figure 6, which shows that the reason there came to be no matching of high productivity firms and workers in the earliest period, -2, is not because high productivity firms stopped making offers, but because high productivity workers stopped accepting them: by the time the players had experienced the stable centralized procedure 10 times (i.e. after market 20), the rate of very early offers remained high, but the acceptance rate had dropped to zero.

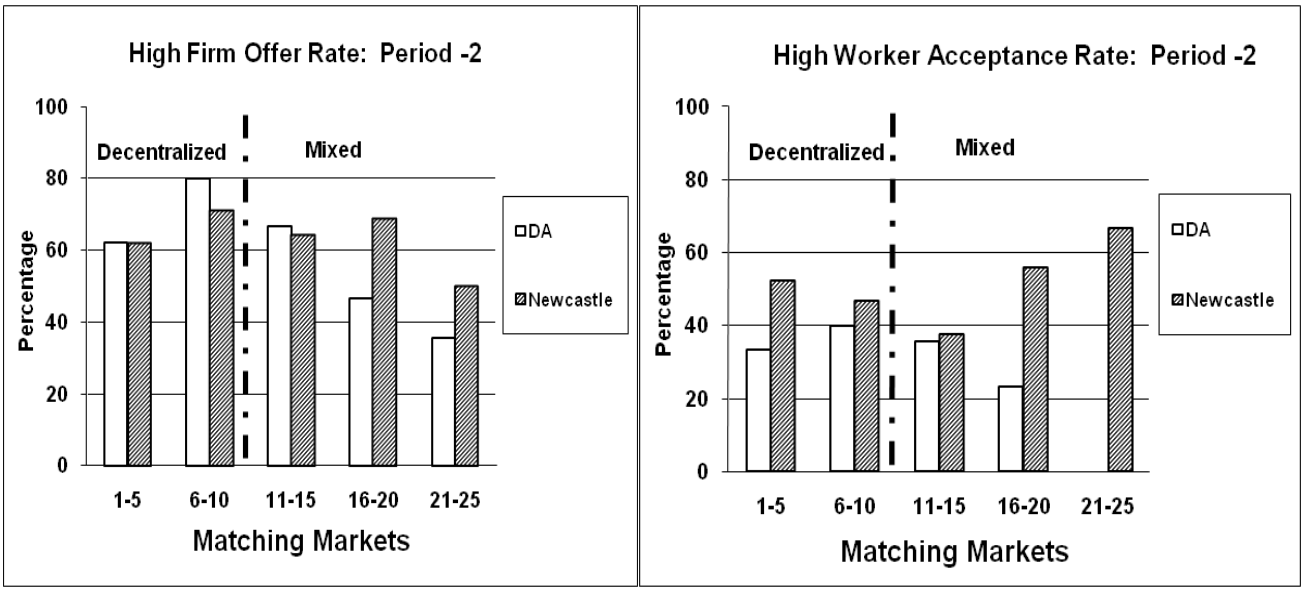


Figure 6: Early offers and acceptances by high productivity firms and workers

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.[40]

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.

[40] 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).

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, i.e. a clearinghouse, 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.

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 a less widely discussed 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 the scheduled clearinghouse.

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.[41] 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 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 (Niederle and Roth, 2003).

[41] 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.

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—fellowship program directors--were accustomed to making early exploding offers, and those who wished to participate in the match worried that if their competitors continued to make 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 American research 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. This experiment was designed to investigate how exploding offers were used strategically, what costs this imposed on participants, and how exploding offers could be deterred, in an uncongested environment in which there would be enough time to make offers (e.g. through a centralized clearinghouse), so that exploding offers were a strategic choice firms could make, but firms could still hire without making them. Also, unlike in the experiments described above, in which the costs of going early were simply imposed by the rules of the experiment, this experiment was designed so that the costs of going early would emerge naturally from the fact that early matches were more likely to be mismatches.

Each market involved 5 firms and 6 applicants, and consisted of 9 periods in which firms could make offers. (So this decentralized experimental market was not designed to be 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.

Two kinds of offers could be made by firms to applicants. An *exploding* offer is an offer that the applicant can only accept right away, i.e. in the same period in which it was made. If an exploding offer is not accepted immediately, it is rejected. An *open* offer is an offer the applicant can also hold (until period 9). That is, an applicant who receives an open offer may accept or reject it immediately, or may hold it, to accept or reject at a later period. An applicant must reject a held offer if he wishes to hold or accept another offer.

In a given period, first all firms decided what offers they would make. An unmatched firm that had no open offer being held by an applicant could decide to make at most one offer. Then each applicant learned of all offers he received in that period before having to decide how to respond to each of them. If an applicant accepted the offer of a firm, the applicant and the firm were matched, and all market participants were informed of this. Offers were made in private; i.e. until they were accepted they were not announced to the other firms and workers.

We considered three environments, characterized by different rules governing offers and acceptances.

Treatment 1: Exploding and Open offers

Each firm can decide whether to make each offer open or exploding. Once an applicant accepts an offer, the acceptance is binding, and firms cannot make subsequent offers to

an applicant who has already accepted an offer. (One can also think of the applicants' ability to make binding agreements as an agreement among firms to not make offers to applicants who accepted another firm's offer.)

Treatment 2: Open Offers Only

Firms can only make open offers. Once an applicant accepts an offer, the acceptance is binding, and firms cannot make subsequent offers to an applicant who has already accepted an offer.

Treatment 3: Renege

In this treatment, firms can again decide whether to make open or exploding offers. However, an applicant who accepted an offer may still receive further offers. An applicant can renege on initial acceptances and accept a new offer at a cost of 1 point (that is subtracted from his final payment).

When players were inexperienced, the average match was made early, after only two signals were available. But as the subjects gained experience, matches became later and more efficient in the two conditions in which binding exploding offers could not be made, but *not* when they could. Figure 7 shows that when the acceptance of exploding offers was binding, final matches resulted from offers that were made when only one or two signals were available as often as from offers made when all three signals had been observed so that the qualities of applicants had been revealed. But in the other two conditions, in which exploding offers could not be made, or in which an applicant who had accepted such an offer could later change her mind, the vast majority of matches resulted from offers made when quality was already known, i.e. after all three signals had been observed.

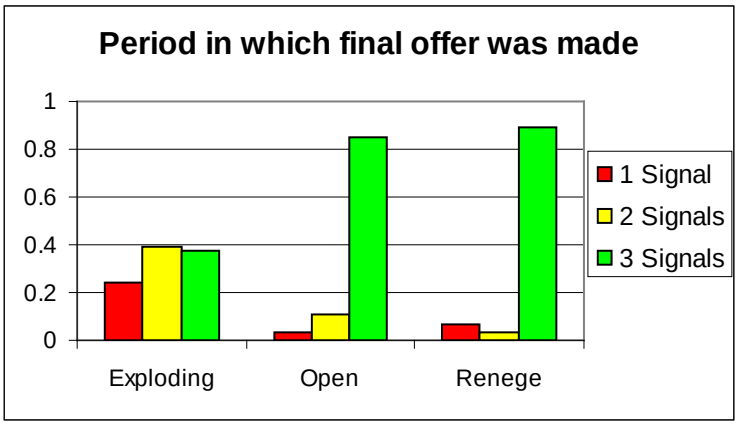


Figure 7. For each treatment, in the last five markets (markets 16-20), the proportion of final offers that were made when one, two or all three signals (and hence the final quality) about applicants' quality were available.

Figure 8 shows the effects this had on efficiency. At an efficient match (far right of the figure), the highest quality firm (F5) is matched to the highest quality applicant (6), and so forth. The open offer and renege treatments get very close to this efficient sorting of firms and applicants. But the efficient matching can only be determined once all three signals are available, and so in the exploding offer treatment, in which matches are made earlier, there is a good deal of compression in which firms are matched with which quality workers. This reduces efficiency, and it reduces the welfare of firms on average, but it can increase the welfare of some firms: in the figure we see that, in this experiment, Firm 2 profited from early matching by often being matched to a higher quality worker than he could get at an efficient matching. So, firm 2 in any of the conditions would have an incentive to match early. But in the renege condition, firm 2 can't capture early applicants who later turn out to be of high quality.

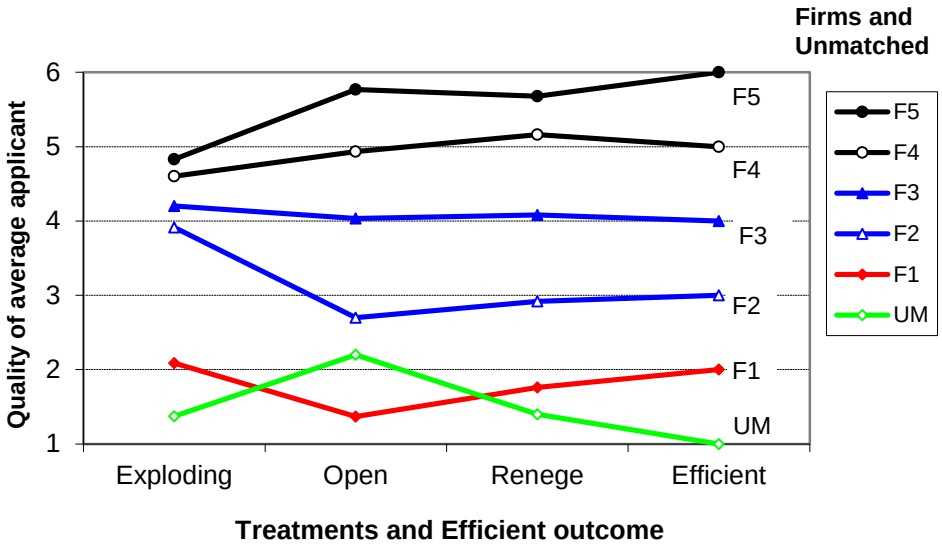


Figure 8: The average quality of the applicant each firm is matched to in the last five markets of each treatment. UM shows the quality of applicants who remain unmatched. Efficient shows for each firm the quality of the applicant in the unique stable and socially efficient match.

The rise in efficiency in the Renege condition came about not because early offers were made, 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 79% of the eligible fellowship programs (121 of 154). 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. And in subsequent years, the match seems to have established itself as a reliable marketplace; in the match for 2010 fellows, 153 certified fellowship programs

participated. This suggests that the policies adopted to decrease the frequency and effectiveness of exploding offers have been effective (cf. Niederle and Roth 2009a,b).

Matching without a clearinghouse: The Market for Economists, and Online Dating

Centralized clearinghouses like those discussed above, which have become common in medical labor markets, help those markets solve a number of problems. The clearinghouses provide a uniform time at which many participants can rely on each other to provide a thick market, and clearinghouses that use a computer to process preferences submitted in advance can process many offers and acceptances and rejections quickly, which avoids congestion. Congestion is a problem that often accompanies success in making a market thick, because when there are many participants on the market at the same time, there may not be enough time to make and consider all the offers that employers and candidates would wish.

One thick market that experiences congestion is the North American market in which new Ph.D. economists are hired as assistant professors. Each year at the end of the first week in January there is a three-day meeting sponsored by the American Economic Association at which departments of economics (and many other potential employers) send their recruiting committees to interview potential candidates. In the months prior to that meeting, available positions are advertised, and students who plan to be "on the market" apply to those for which they think they might be eligible and interested. So each student sends out many applications and each department receives many: it is not unusual for a department with a single position to receive hundreds of applications. Consequently employers cannot interview every applicant in the time available at the meetings, but must select perhaps twenty or thirty to interview. (Typically, interviews at the meetings last for about 45 minutes, and some subset of the candidates interviewed will be later invited to fly out for a day-long visit with the whole department, after which the department will hope to make an offer to and hire as many candidates as it has positions.)

So departments cannot interview as many candidates as they might like. Because it is easy for candidates to send out many applications, receiving an application is not a strong signal of how interested the candidate is in the department, and in any event the candidates send their applications and the departments mostly choose candidates to interview without knowing which (other) departments will interview those candidates. Except for a very few of the highest prestige departments, it would therefore be unwise for departments to simply devote their twenty or so interview slots to the candidates whose work they most admire, since these candidates are likely to be in high demand and hence difficult to hire. Instead, most departments need to choose a portfolio of candidates to interview, taking into account not only how promising the candidate looks to them, but also how likely they are to be able to hire that candidate. Failure to coordinate efficiently across the market at this stage leads to some candidates "falling through the cracks," and not getting offers from any of the departments that interviewed them, while other departments that might have wanted to hire them chose not to interview them.[42]

To facilitate coordination of interviews among departments and candidates, an AEA committee on the job market that I chaired developed and implemented a signaling mechanism. In December, after most job ads have appeared, candidates are invited to log on to an AEA website to send up to two signals of interest to departments to which they have applied. (Each position that is advertised in the AEA's Job Openings for Economists is required to indicate an email address to which signals should be sent.) Departments receive these signals, and know that candidates chose to use one of their (only two) signals to indicate that they would like to be interviewed for that job.

My committee was well aware that proposing any innovations in the main job market for new economists would be greeted with a certain wariness, which we felt would prevent us from incorporating an experiment into the new design, even though that would be the

[42] Departments can also fall through the cracks and fail to hire a candidate if all of the candidates who they interviewed and wanted to hire received preferable offers, and they failed to interview candidates they would have liked for whom they would have made the best offer. Along with the signaling mechanism I will describe, the AEA also constructed a "scramble" web page to help applicants still available and departments with positions still open late in the market to find one another.

best way of evaluating its impact. For example, one could imagine an experiment in which some signals were randomly chosen not to be transmitted, or in which some candidates were randomly chosen to be allowed to send a third signal, so that the influence of signals on subsequent interviews, offers and hires could be examined independently from the selection by candidates of where to direct their signals. However such experiments would give some candidates an unfair advantage over others on the job market, and also might impede the effectiveness of signals by introducing uncertainty into their transmission.[43] And so the proposal we made to the AEA's executive committee did not include any experimental variation in the new signaling mechanism that we proposed, and that was subsequently adopted and implemented.[44] And, while the early evidence from the first few years of operation looks quite positive, it is far from definitive (see Coles et al 2010)[45]. This evidence suggests that signals increase the chance of interviews, particularly when a candidate sends a signal to a department that is less prestigious than the one from which the candidate is receiving his Ph.D.[46] The early empirical evidence also suggested that the signals were helping coordination: with some exceptions among the top 100 ranked employers, signals were fairly uniformly distributed. (Among the top ranked employers, those at the very top—which are not expected to pay attention to signals of interest-- receive very few signals, while some mid-ranked universities in attractive cities receive many.)

---

[43] For example, if some signals were randomly chosen to be not transmitted, candidates might claim to have sent (untransmitted) signals to more than two employers.

[44] When I formally presented the proposal of the signaling mechanism to the executive committee, I mentioned that we would not be including such an experiment, and my remark was greeted with laughter, as if the very idea of including such an experiment were a joke. This reflects one of the considerable obstacles to field experiments in market design, which is that when the stakes are high for the participants, an important part of gaining support for market institutions is that they be seen to be equitable. Random or other arbitrary variations in access to the market that disadvantage some participants seldom meet this standard. Somewhat similar issues arise in clinical trials in medicine, in which patients must quickly be switched to the most effective treatment after sufficient evidence is collected to indicate which that is. But variations in treatment are simpler since patients are not competing with one another, and so no one suffers if an experimental treatment tested on some patients proves to be effective for them. But job applicants do compete with one another in a labor market, so that if an experimental treatment (like being given extra signals) benefited some candidates, this benefit might come at the expense of others.

[45] The authors of the paper were the members of the committee.

[46] For judgments of prestige we used one of the widely available research productivity rankings of academic and non-academic employers of economists, based on journal publications, which included over 500 employers.

An experiment on a similar environment could provide much more definitive evidence of the effect of signals, and just such an experiment was conducted by Lee and Niederle (2014), who introduced a signaling mechanism into a special event held for a limited time by a Korean online dating/marriage site. The event enrolled approximately equal numbers of men and women, 304 and 309, respectively[47], and gave them the ability to send a contact message—a "proposal"—to up to ten potential dates, over a period of five days. Each participant was also given two virtual "roses" each of which they could attach to one of their proposals as a signal of particular interest. And a randomly selected 20% of participants were given additional "roses" for a total of 8, so that this group had the advantage of being able to send a signal of interest to many more people.

After the contact period, participants could decide which if any of the proposals they had received to accept (without yet knowing which of their own proposals were accepted), and after this stage men and women who had accepted each other's proposal were given one another's contact information. So the main data of this experiment will be the effect of roses on the acceptance rate of proposers, and the relative success of participants who had many roses compared to those who had few.

This site also had regular subscribers, and as part of its algorithm for suggesting matches to them it rated everyone who enrolled on their desirability as a marriage partner, using a measure which took into account judgments of physical attractiveness as well as verified financial, employment, education and family data.[48] Participants were not told their desirability score. Participants in this special event were also rated. So Lee and Niederle were able to examine the effect of proposals, with and without roses, as a function of the relative rated desirability of both the sender and the receiver. For the purposes of some of their analyses they classified participants as being in the top, middle or bottom part of the distribution of rated attractiveness.

---

[47] In an effort to have a relatively homogeneous thick market instead of one with many distinct submarkets, enrolment in this event was restricted to participants who were Korean, college educated, never married, and between 26 and 38 years old for men, and between 22 and 34 for women.

[48] Lee (2009) verified from site data that this desirability index is a good predictor of whether a client is attractive as a dating partner.

Among the general patterns that Lee and Niederle observe is that participants are more likely to send proposals to recipients who are rated as more desirable, and that this effect is more pronounced among more desirable senders of proposals. However the decision to add a rose to the proposal is not additionally correlated with the overall desirability rating. For example while participants in the top desirability group received more proposals than others, the fraction of proposals with a rose attached was similar. So it appears that the roses may be allocated among proposals based on idiosyncratic personal preferences.

Looking at participants who received at least one proposal, Lee and Niederle observe that the overall probability of accepting a proposal was 16% for women, who received an average of 6 proposals, and 29% for men, who received an average of 4 proposals. The higher the desirability rating of the sender, the more likely the proposal was to be accepted. (No participant accepted over 8 proposals.)

Proposals with roses attached were 20% more likely to be accepted (3.3 percentage points increase in acceptance rate). Lee and Niederle note that "this positive effect of sending a rose is comparable to (and about three-quarters of) the benefit of being in the middle desirability group relative to being in the bottom group." However, unlike participants in the other two desirability groups, participants in the top group did not appear to pay attention to roses in deciding which proposals to respond to. Indeed, the effect of a rose was clearest (and always positive and significant) when the sender was in a higher desirability group than the receiver.

Lee and Niedlerle also assessed, more tentatively, whether roses merely reallocated acceptances, or increased the total number of proposals that were accepted, by examining a subset of participants in the middle desirability groups who received similar numbers of offers and who seemed otherwise similar, focusing on a group that seemed equally likely to have received zero or one rose. They conclude that the group that receives a rose accepts about 0.26 more offers on average.

Dating and courtship are important in their own right, and this experiment reveals how signaling might usefully be employed to convey information, relieve congestion, and improve coordination in online dating. It also provides some evidence of how signaling may be working in the economics job market. In this connection, the theory associated with the signaling of preferences is still in its infancy (in contrast with the theory associated with signals about costs and abilities, in the manner of the literatures started by Spence 1973 and Zahavi 1975). Early papers on preference signaling have provided some guidance on what to look for in experiments (see especially Coles et al. 2013), and experimental results such as these will provide some guidance to developing further theory.

6. Course allocation

Sometimes an experiment serves multiple roles in the process of designing and testing and demonstrating needed to bring a new market design from conception to implementation. That was the case with an experiment conducted by Eric Budish and Judd Kessler as part of the process by which a new course allocation mechanism, now called "Course Match," was adopted in academic year 2013-14 to assign MBA students to courses at the Wharton School of the University of Pennsylvania.[49]

Course allocation is both a prosaic task that colleges engage in regularly, and an allocation task that to accomplish well presents some of the most difficult problems in market design, since it involves assigning to each of many students a different package of indivisible goods, namely places in classes, and the preferences of the students over packages of classes may not be a simple function of preferences over individual classes.

[49] The online Course Match User Manual is here: https://spike.wharton.upenn.edu/mbaprogram/course_match/Course%20Match%20Manual%20-%20v1_4.pdf

Some classes may be substitutes while others may be complements, and this may be different for different students. Packages of classes also come with schedules, and the same set of courses may be achievable with more or less desirable schedules when there are multiple sections of the same course or of courses that are close substitutes (e.g. the same class taught by different instructors). The problem is additionally complicated by the fact that most universities would be reluctant to charge different tuition to students studying for the same degree based on which particular courses they took among those offered. This constraint removes from consideration the variety of package bidding auctions that we have discussed for allocating other kinds of goods, like licenses for electromagnetic spectrum. And, not least, the problem facing students competing for space in over-demanded classes is at least partly strategic, since which courses are over-demanded depends on the preferences and strategies of other students, which may in turn depend on their perception of others' preferences and strategies. And because students can learn from the experience of older students who have already gone through the process, allocation systems that are not strategy proof are likely to be subject to more or less well informed attempts to game the system.

The demand to allocate courses well has been particularly clear in MBA education. A typical MBA degree program lasts for two years, and in the first year students often take a prescribed list of required courses, while in the second year there is an opportunity to choose among a wide variety of very different courses that in turn lead to different kinds of post-graduation careers, e.g. in finance or operations or management consulting or entrepreneurship. Sometimes courses taught by particular professors may give students access to a valuable network of contacts. One approach to allocating courses which has long been popular at graduate schools of business is to try to achieve some of the efficiencies of an auction, without using money, by having students bid for courses in "artificial money," i.e. in a currency that is allocated to each student solely for the purpose of course allocation. Wharton started using such a system, in 1996.

The particular system that was in place at Wharton was a multi-stage auction that allowed students to both buy and sell places in courses, which in turn led to well-

publicized speculative behavior[50], in which some students would place bids in early stages for courses that they wished not to enroll in but to sell in later stages in order to turn a profit that would allow them to make successful high bids for popular courses that they did wish to attend. But, while this particular form of strategic behavior was particular to the Wharton bidding process, the whole class of auctions in artificial money suffered from an important design flaw that often left students frustrated and unsatisfied, and led to courses being allocated inefficiently (as described in Sönmez and Ünver, 2003, 2010).

Briefly, the way these auctions worked was that students would allocate their budget of artificial money among bids for each course, and each course would be filled up to its capacity by the highest bidders for that course, for example at a price equal to that of the highest rejected bid. Courses that had fewer bidders than their capacity would therefore have a price of zero, and students who devoted their bids to over-demanded courses and did not succeed in filling their schedules could enroll in under-demanded courses afterwards.

For example, consider a student who bids 100 on three over-demanded courses, each of which ultimately sells at a price of 110. He does not win a place in any of them, and eventually enrolls in a set of courses that include some on which he placed a winning bid, and others that are under-demanded and sell for 0, leaving him with an unused budget of 300 units of artificial money. Now, if the money were real, having it would be some consolation for not having his preferred courses, but even this consolation is absent when the unused 300 units are useless for any purpose other than the one that they just failed to perform, namely getting the student access to the courses he wants. In some kinds of multi-stage auctions this student would have some opportunity to redeploy his budget, e.g. by bidding more for two out of the three courses that he wished to get. But the

[50] See e.g. http://whartonjournal.com/2012/03/19/no-more-auctions-details-on-the-new-course-allocation-system/ or http://www.businessweek.com/printer/articles/153446-gaming-whartons-course-scheduling-system-just-got-tougher .

resulting allocation could be inefficient in the sense that post-auction trades could be arranged among students that would give them all preferred bundles of courses.

Because course assignment is inherently difficult, most ad hoc procedures have problems with efficiency and/or incentives for truthful revelation of preferences. Budish and Cantillon (2012) studied the very different course allocation procedure at the Harvard Business School and cataloged how it also encouraged strategic behavior and often failed to produce efficient outcomes. Budish (2011) proposed a new allocation mechanism, to address some of these deficiencies.

Specifically, Budish proposed to elicit from students their preferences over bundles of courses, and use these preferences to compute an approximate competitive equilibrium from equal incomes (CEEI). In this mechanism, students would be given (approximately) equal endowments of artificial money (with the inequalities being used to break ties), and their preferences over bundles of courses would be used to compute a price for each course that would form an (approximate) competitive equilibrium, at which each student would receive the most preferred bundle of courses that he could afford. (Note how this is very different from the previous example of a simple auction, in which the student who bid too low on each of three courses received none of them although he might well have been able to afford two of them even after prices adjusted.) Budish further shows that the amount of approximation involved would become small in large markets, and so would the potential to manipulate prices (and hence allocations) by mis-stating preferences..

Now, in a theory paper it is perfectly appropriate to have students report their preferences over bundles of courses, but in practice it would be difficult or impossible to elicit all possible preferences, since the number of packages of courses that can be constructed becomes astronomical for even a modest universe of possible courses from which packages can be chosen. So a practical mechanism must simplify the language in which preferences can be reported, and by doing so it will restrict which preferences can be reported. This will raise the empirical question of how well the restricted preferences that can be reported approximate the true preferences of the participants, and how

successful participants will be at using this language to report their approximate true preferences.

The proposed new Wharton course allocation system allowed students to assign a value between 0 and 100 to each course, and to enter "adjustments" for pairs of courses, e.g. to indicate that a pair were substitutes by indicating that the value of being assigned both courses would be less than the sum of the individual course values; or, complements by indicating that the pair would be worth more than the sum.[51] The assumption was that student preferences would be well represented by trying to maximize the sum of the (adjusted) cardinal values of the courses to which they were assigned.

An additional potential obstacle to using CEEI as a practical course assignment mechanism is that it is much less transparent than most mechanisms that are used in practice. Prices are computed by doing an approximate fixed point calculation, not by any transparent price formation process as in an auction. And so it was not clear how satisfied participants would be with such a mechanism, or how well it could be explained so that it would be easy to use.

Prior to the final design phase of the new choice mechanism, and prior to the decision by the Wharton deans to adopt it, Budish and Kessler therefore conducted an experiment whose participants were Wharton MBA students. The students were presented with a somewhat simplified version of both the new CEEI mechanism and the existing Wharton Auction, and a subset of Wharton MBA courses meant to represent those available for a single semester. The multiple goals of the experiment included comparing the performance of the two mechanisms from the point of view of producing course assignments that the students preferred, assessing the ease of use of the two systems, and revealing any unforeseen issues that might influence the final design decisions or the decision whether to adopt the new mechanism in place of the old one.

[51] The interface also provided the students with feedback on the preferences that they had submitted by computing their most preferred schedules under those preferences so that they could see if these corresponded to what they wanted.

To this end, one part of the experiment was designed to directly assess the effectiveness of the preference reporting language that the mechanism introduced. A second part of the experiment tested the relative performance of the new and old allocation mechanisms in producing schedules that were preferred by students. In addition, participants were surveyed following their use of the two mechanisms, for feedback about how easy they were to understand and use.

To test how accurately this compressed language could represent actual student preferences, students were presented with a sequence of choices between two possible schedules, so that their actual choices could be compared to the choices predicted on the basis of the course values and adjustments they had submitted. The assumption here is that students' choices between two schedules allows them to express their preferences accurately, in a way that reflects parts of their preferences that may have been suppressed by the limited language in which they could report them.[52] Budish and Kessler report that across all the data, the binary comparison response contradicted the preference reports just under 16% of the time, but that few of these cases involved large differences in the cardinal values reported. That is, when the compressed language reported large differences, the predicted choices overwhelmingly agreed with the observed binary choices between pairs of schedules. When they further analyzed the differences between predicted and observed choices they found that the unpredicted less popular choices often had "unbalanced" schedules that did not spread the classes evenly over Monday through Thursday, which was a preference that could not easily be expressed in the compressed preference language.

Comparisons of the two allocation mechanisms were made by having each participant participate in both mechanisms, with the order in which the two mechanisms were used randomly chosen in each experimental session. Afterwards, the resulting schedules were compared both by having the subjects make binary choices of which of the two resulting schedules they preferred, and using the approximate predicted preferences. It makes

[52] Note that this experiment does not induce artificial preferences for abstract commodities, but rather attempts to assess the preferences of Wharton MBA students over schedules of existing courses.

sense to compare schedules by experimental sessions, i.e. by groups of students who participated at the same time, since students are to some extent competing for popular courses. Budish and Kessler report that, of the 8 sessions they conducted, a majority of students preferred the schedules they received from the new CEEI mechanism in six sessions, and they were evenly divided in two sessions, so that in no session did more students prefer the schedule they received under the old mechanism. By other measures as well, e.g. by distribution of cardinal scores for schedules received, and by the (in)frequency with which students "envied" another student's schedule, the new mechanism also outperformed the old one.

The surveys conducted after students had used both mechanisms revealed considerably more satisfaction with the new mechanism than with the old one, particularly in connection with the ease of use (which didn't require extensive strategizing), as well as the outcomes achieved. One source of dissatisfaction was the difficulty in understanding how the mechanism produced the outcome that it did, e.g. in understanding why it produced a specific schedule and not one that would have been preferred.

This latter concern led to a change in how the results were presented. In the experiment, each student saw only the schedule he received. But in the mechanism as it was prepared for implementation, students could also see the prices that had been computed for each course, so they could tell that they could not afford schedules they might prefer.

Taking account of these results, Wharton went ahead and adopted the new course allocation system, which replaced the prior auction, and was used for the first time in the 2013-14 academic year. Preliminary reports from the first year look promising (Kessler and Budish report both increased student satisfaction and decreased inequality among students in connection with how many of the most popular courses they receive), and more evidence will accumulate in the coming years.

7. Conclusions

Experiments and market design have a shared history. It goes back considerably longer than the recent developments that have allowed experiments to play an important role in developing designs for new markets that have then been adopted and implemented largely according to plan.

Perhaps paradoxically, experiments have become more useful in practical market design as experimenters and designers have moderated their ambitions for how much weight can be placed on experimental evidence by itself. One of the lessons experimenters have started to learn is that simple experiments may not be effective by themselves, either scientifically or politically, to carry the day in debates about complex markets. (The same, not coincidentally, can be said of abstract economic theory.[53]) Experiments seem to have more often met with success when used in conjunction with other kinds of investigation. That is, in recent years experiments have begun to play a more modest but more effective role in helping market designs become implemented in functioning markets.

Although experiments haven't played an important role in every successful modern market design, in those in which experiments have played a role they have often played *multiple* roles, in the considerable amounts of discovery, demonstration, and persuasion that are necessary to coordinate market participants to move from a failed market design to a better one.[54] Whether experiments are called for seems to depend on which kinds of

[53] For example Ostrovsky and Schwarz (2009) report on their successful implementation of optimal reserve prices a la Myerson (1981) in ad auctions run by Yahoo! only in 2008, and after running a large scale randomized experiment. Varian and Harris (2014) report on the implementation of VCG (Vickrey 1961, Clark 1971, Groves 1973) auctions for a broad range of Google ad auctions in 2014, after years of running the simpler-to-describe Generalized Second Price auction (Varian 2007 and Edelman, Ostrovsky and Schwarz 2007) for ads based on search words. In each case, a very long time elapsed between the development of the theory and the implementation of some basic theoretical ideas, which happened only after lots of experience, including experiments, with online auctions for advertisements had accumulated.

[54] Although I am a committed experimenter, experiments have not (yet) played a critical role in all of the market designs in which I have participated. For example, kidney exchange has largely proceeded without laboratory experiments (although we have used computational experiments based on contemporary field data to good effect, as in Ashlagi et al 2011a,b). It may eventually turn out that experiments play a more important role in discussions related to transplantation from *deceased* donors: see e.g. Kessler and Roth 2012, 2014a,b. And experiments have not played a large role so far in implementing school choice designs in New York, Boston, Denver, New Orleans and elsewhere. However the experiment of Chen and Sönmez (2006) played a role in persuasion and communication early in our engagement with Boston Public

critical questions can't be answered well or demonstrated clearly on theoretical grounds or by empirical investigation of markets "in the wild." Thus for example, when a novel design is proposed that doesn't yet exist in the wild (as in the case of Wharton's novel course allocation system), experiments can provide empirical evidence that doesn't exist anywhere else. And even when empirical evidence from existing markets is available, it is difficult to draw conclusions about one complex market based on the experience of a different, differently complex market (e.g. by comparing the markets for doctors in the U.S. and the U.K. or the markets for gastroenterology fellows and graduate students). A simple experimental environment is also very different from any of these markets, but because it is simple, the treatment effects can be easily connected to the experimental treatment variables, as opposed to being attributed to conjectured interactions with the complexities of one of the markets. So there are advantages of simplicity. But simple experimental results are often met with skepticism by practitioners who are steeped in the important complexities of their market. It can therefore be important to have the combined weight of a compelling theoretical argument along with evidence from both complex natural environments and simple experimental environments that together provide a range of evidence needed to move a design proposal forward to adoption and implementation.

Notice that while market design experiments are vastly simpler than the target markets they are meant to illuminate, they are often more complex than some of the best known laboratory experiments used to illuminate aspects of human behavior.[55] This is because design experiments are meant to incorporate, albeit simply, some relevant institutional market features. So the goal of a design experiment is to illuminate behavior in a specific institutional environment, rather than behavior in possibly great generality. This parallels the development of the theoretical literature of market design. In contrast to

---

Schools, and other experiments have played a role in the academic discussion of school choice; see e.g. Featherstone and Niederle 2013, Calsamilgia, Haeringer and Klijn 2010 and in related contexts, e.g. Guillen and Kesten 2012, who compare versions of the deferred acceptance and top trading cycles algorithms in a related (housing) context. Much the same could be said for experiments concerned with other market design discussions, such as those concerning markets for pollution permits.

[55] Think for example of the prisoner's dilemma, and its vast experimental literature subsequent to the initial experiment in 1950.

traditional economic theory, which is valued for its simplicity, elegance and potential generality, theoretical developments in market design are often addressed to how specific institutional features (e.g. of medical labor markets, or kidney transplantation) present unique problems. That is, both theory and experiments in market design aim for less generality and more specificity than work that does not aspire to being implemented in particular markets.

In summary, practical market design takes on responsibility for complexity, and it is this complexity that both limits how much weight can be put on experimental evidence alone, but also opens the door for experiments to usefully complement other empirical, institutional and theoretical investigations, and to serve the multiple roles of discovery, communication, and persuasion, as well as testing new designs that can't yet be observed anywhere else but in the lab.

Bibliography

Abbink, Klaus, Bernd Irlenbusch, Bettina Rockenbach, Abdolkarim Sadrieh, and Reinhard Selten, (2002): The behavioural approach to the strategic analysis of spectrum auctions: The case of the German DCS-1800 auction. *Ifo Studien* 3, 457-480

Abbink, Klaus, Bernd Irlenbusch, Paul Pezanis-Christou, Bettina Rockenbach, Abdolkarim Sadrieh, and Reinhard Selten, (2005): An Experimental Test of Design Alternatives for the British 3G / UMTS Auction. *European Economic Review* 49(2), 505-530

Abraham, D., Blum, A., and Sandholm, T. 2007. Clearing Algorithms for Barter Exchange Markets: Enabling Nationwide Kidney Exchanges. In Proceedings of the ACM Conference on Electronic Commerce (EC).

Anderson, Ross, Itai Ashlagi, David Gamarnik, Michael Rees, Alvin E. Roth, Tayfun Sönmez and M. Utku Ünver, "Kidney Exchange and the Alliance for Paired Donation," *Interfaces,* forthcoming.

Ariely, Dan, Axel Ockenfels, and Alvin E. Roth (2005), "An Experimental Analysis of Ending Rules in Internet Auctions," *Rand Journal of Economics*, 36, 4, Winter, 891-908.

Ariely, Dan and Itamar Simonson (2003). Buying, Bidding, Playing or Competing?: Value Assessment and Decision Dynamics in Online Auctions. *Journal of Consumer Psychology* 13(1–2). 113–123.

Asker, John, Brit Grosskopf, C. Nicholas McKinney, Muriel Niederle, Alvin E. Roth and Georg Weizsäcker, "Teaching auction strategy using experiments administered via the Internet," *Journal of Economic Education*, 35, 4, Fall 2004, 330-342.

Ashlagi, Itai, Duncan S. Gilchrist, Alvin E. Roth, and Michael A. Rees (2011a), "Nonsimultaneous Chains and Dominos in Kidney Paired Donation – Revisited," American Journal of Transplantation, 11, 5, May, 984-994

Ashlagi, Itai, Duncan S. Gilchrist, Alvin E. Roth, and Michael A. Rees (2011b), "NEAD Chains in Transplantation," American Journal of Transplantation, December; 11: 2780–2781.

Ausubel, Lawrence M. and Oleg V. Baranov, "Market Design and the Evolution of the Combinatorial Clock Auction," *American Economic Review: Papers & Proceedings*, May 2014, 104(5): 446–451

Ausubel, Lawrence, Peter Cramton, and Paul Milgrom. 2005. "The Clock-Proxy Auction: A Practical Combinatorial Auction Design," in *Combinatorial Auctions*. Peter Cramton, Yoav Shoham and Richard Steinberg eds. Cambridge, MA: MIT Press.

Ayres, Ian and Peter Cramton (2010), "Fix Medicare's Bizarre Auction Program," New York Times, Sept. 30, 2010, http://www.freakonomics.com/2010/09/30/fix-medicares-bizarre-auction-program/

Budish, Eric and Eduardo M. Azevedo (2013), "Strategy-proofness in the Large," working paper, University of Chicago, March 2013.

Ball, Michael O., Lawrence M. Ausubel, Frank Berardino, Peter Cramton, George Donohue, Mark Hansen, and Karla Hoffman "Market-Based Alternatives for Managing Congestion at New York's LaGuardia Airport," in *Optimal Use of Scarce Airport Capacity, Proceedings of AirNeth Annual Conference*, The Hague, April 2007.

Banks,J.S., J.O. Ledyard, and D.P. Porter Allocating uncertain and unresponsive resources: An experimental approach - The Rand Journal of Economics, 1989, 20, 1 (Spring), 1-25.

Bazerman, M.H., & Samuelson, W.F. "I Won the Auction But Don't Want the Prize," *Journal of Conflict Resolution*, 1983, 27, 618-634.

Bolton, Gary, Ben Greiner, and Axel Ockenfels, "Engineering Trust: Reciprocity in the Production of Reputation Information," Management Science, 59, 2, February 2013, 265-285.

Bolton, Gary, and Axel Ockenfels (2012), "Behavioral economic engineering," Journal of Economic Psychology, 33, 3 (June), 665-676.

Brewer, Paul J. and Charles R. Plott (1996), "A binary conflict ascending price (BICAP) mechanism for the decentralized allocation of the right to use railroad tracks," *International Journal of Industrial Organization*, 14, 857-886.

Brewer, Paul J. and Charles R. Plott (2002), "A Decentralized, Smart Market Solution to a Class of Back-Haul Transportation Problems: Concept and Experimental Test Beds," *Interfaces*, 32, 5 (September-October), 13-36.

Brown, Jennifer, John Morgan and Tanjim Hossain "Shrouded Attributes and Information Suppression: Evidence from the Field," *Quarterly Journal of Economics*, May 2010, 125(2), 859-876

Brunner, Christoph, Jacob K. Goeree, Charles A. Holt, and John O. Ledyard (2010), "An Experimental Test of Flexible Combinatorial Spectrum Auction Formats" *AEJ: Microeconomics*, 2, 1, February, 39-57.

Brusco, Sandro, Giuseppe Lopomo, and Leslie M. Marx (2009), "The 'Google effect' in the FCC's 700 MHz auction," *Information Economics and Policy*, 21, 101-114.

Budish, Eric (2011), "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes," *Journal of Political Economy* Vol. 119(6), Dec 2011, pp 1061-1103

Budish, Eric and Estelle Cantillon (2012), "The Multi-unit Assignment Problem: Theory and Evidence from Course Allocation at Harvard," *American Economic Review* 102(5): 2237–2271.

Budish, Eric and Kessler, Judd (2014), "Changing the Course Allocation Mechanism at Wharton," working paper.

Calsamiglia, Caterina, Guillaume Haeringer, and Flip Klijn. 2010. "Constrained School Choice: An Experimental Study." *American Economic Review*, 100(4): 1860-74.

Capen, E.C., R.V. Clapp, and W.M. Campbell (1971), "Competitive Bidding in High-Risk Situations," *Journal of Petroleum Technology*, **23**, 641-653.

Chamberlin, Edward H. 1948. "An Experimental Imperfect Market," *Journal of Political Economy*, 56 no. 2, 95-108.

Chen, Yan and Tayfun Sonmez (2006), "School Choice: An Experimental Study," *Journal of Economic Theory*, 127, 202-231

Clarke, E. (1971). "Multipart Pricing of Public Goods". *Public Choice* 11 (1): 17–33
Coles, Peter, John H. Cawley, Phillip B. Levine, Muriel Niederle, Alvin E. Roth, and John J. Siegfried, "The Job Market for New Economists: A Market Design Perspective," *Journal of Economic Perspectives*, 24,4, Fall 2010, 187-206.

Coles, Peter, Alexey Kushnir and Muriel Niederle, "Signaling in Matching Markets", American Economic Journal, Microeconomics, 2013, 5(2): 99–134.
Connolly, Michelle and Evan Kwerel (2007), "Economics at the Federal Communications Commission: 2006-2007," *Review of Industrial Organization*, 31, 107-120.

Cramton, Peter, Yoav Shoham, and Richard Steinberg, editors (2006), *Combinatorial Auctions*, MIT Press, Cambridge, MA.

Crawford, Gregory S., Evan Kwerel, and Jonathan Levy (2008), "Economics at the FCC: 2007-2008," *Review of Industrial Organization*, 33, 187-210.

Cummings Ronald G., Charles A. Holt, and Susan K. Laury "Using laboratory experiments for policymaking: An example from the Georgia irrigation reduction

auction," Journal of Policy Analysis and Management, Volume 23 Issue 2, Pages 341 – 363, 2004

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

Day, Robert W. and Paul Milgrom. 2007. "Core-Selecting Package Auctions." *International Journal of Game Theory*, pp. 393-407.

Edelman, Benjamin, Michael Ostrovsky and Schwarz (2007), "Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords, *American Economic Review*, 97(1), March, 242-259

Einav, Liran, Theresa Kuchler, Jonathan Levin, and Neel Sundaresan, "Learning from Seller Experiments in Online Markets," December 2013, http://www.stanford.edu/~leinav/Seller_Experiments.pdf

Ely, Jeffrey C., and Tanjim Hossain, "Sniping and Squatting in Auction Markets," *American Economic Journal: Microeconomics*, August 2009, 1(2), 68-94

Eyster, Erik and Matthew Rabin (2005), "Cursed Equilibrium," *Econometrica* 73 (5) , 1623–1672

Featherstone, Clayton and Muriel Niederle, "Improving on Strategy-Proof School Choice Mechanisms: An Experimental Investigation", July, 2013, working paper.

FEDERAL AVIATION ADMINISTRATION, Supplemental Notice of Proposed Rulemaking for Congestion Management at LaGuardia Airport, New York, **Docket ID:** FAA-2006-25709, April 17, 2008, http://www.federalregister.gov/OFRUpload/OFRData/2008-08308_PI.pdf (accessed 6/18/08)

FEDERAL AVIATION ADMINISTRATION 05/15/2008, Congestion Management Rule for John F. Kennedy International Airport and Newport Liberty International Airport; Proposed Rule, FAA-2008-0517-0001.1 http://edocket.access.gpo.gov/2008/08-1271.htm (accessed 6/18/08)

Fiorina, M. and Charles R. Plott (1978), "Committee Decisions Under Majority Rule: An Experimental Study," *American Political Science Review*, 72, June, 575-98.

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

Flood, Merrill M. 1954a. "On Game-Learning Theory and some Decision-Making Experiments," Decision Processes, edited by R.M. Thrall, C.H. Coombs, and R.L. Davis, Wiley, New York, 139-158.

Garey, Michael R. and David S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman.

Goeree, Jacob K., Charles A. Holt, Comparing the FCC's Combinatorial and Non-Combinatorial Simultaneous Multiple Round Auctions: Experimental Design Report, Prepared for the Federal Communications Commission, April 27, 2005, http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-05-1267A2.pdf

Goeree, Jacob K., Charles A. Holt, and John O. Ledyard, "An Experimental Comparison of the FCC's Combinatorial and Non-Combinatorial Simultaneous Multiple Round Auctions," Prepared for the Wireless Telecommunications Bureau of the Federal Communications Commission, July 12, 2006, http://wireless.fcc.gov/auctions/data/papersAndStudies/fcc_final_report_071206.pdf

Goeree, Jacob K., Charles A. Holt, and John O. Ledyard, "An Experimental Comparison of Flexible and Tiered Package Bidding," Prepared for the Wireless Telecommunications Bureau of the Federal Communications Commission, Final Report, May 25, 2007, http://wireless.fcc.gov/auctions/data/papersAndStudies/fcc_report_052507_final.pdf

Goeree, Jacob K. and Charles A. Holt (2010), "Hierarchical package bidding: A paper & pencil combinatorial auction," *Games and Economic Behavior*, 70(1), September 2010, 146-169.

Goeree, Jacob K. and Luke Lindsay (2012), "Designing Package Markets to Eliminate Exposure Risk," Working Paper, April, Dept. of Economics, University of Zurich.

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.

Grether, David M., R. Mark Isaac, and Charles R. Plott (1981), "The Allocation of Landing Rights by Unanimity Among Competitors," *American Economic Review*, Papers and Proceedings, May, 1981, 166-171.

Grether, David M., R. Mark Isaac, and Charles R. Plott (1989), *The Allocation of Scarce Resources: Experimental Economics and the Problem of Allocating Airport Slots*, Westview Press, Boulder, CO.

Grether, David M. and Charles R. Plott (1984), "The effects of market practices in oligopolistic markets: An experimental examination of the Ethyl case," *Economic Inquiry*, 22, 479-507.

Grosskopf, Brit and Alvin E. Roth (2009), "If you are offered the Right of First Refusal, Should you accept? An Investigation of Contract Design," *Games and Economic Behavior*, Special Issue in Honor of Martin Shubik, 65 (January), 2009, 176–204.

Groves, T. (1973). "Incentives in Teams". *Econometrica* **41** (4): 617–631

Guillen, Pablo and Onur Kesten (2012), "Matching Markets with Mixed Ownership: The Case for a Real-life Mechanism," *International Economic Review* 53(3), 2012; 1027-1046.

Holt, Charles A. (1995), "Industrial Organization: A Survey of Laboratory Research," Chapter 5 in J.H. Kagel and A.E. Roth (eds.), *Handbook of Experimental Economics*, Princeton University Press, 349-443.

Hong, James T. and Charles R. Plott (1982), "Rate filing policies for inland water transportation: An experimental approach," *Bell Journal of Economics*, 13, 1-19.

Hossain, Tanjim and John Morgan, (2006) "...Plus Shipping and Handling: Revenue (Non) Equivalence in Field Experiments on eBay," *Advances in Economic Analysis & Policy*: Vol. 6 : Iss. 2, Article 3, http://www.bepress.com/bejeap/advances/vol6/iss2/art3

Kagel, John H. (1995), "Auctions: A Survey of Experimental Research," Chapter 7 in J.H. Kagel and A.E. Roth (eds.), *Handbook of Experimental Economics*, Princeton University Press, 501-585.

Kagel, John H. and Dan Levin (2002), *Common Value Auctions and the Winner's Curse*, Princeton University Press.

Kagel, John H., Yuanchuan Lien, and Paul Milgrom (2010), "Ascending Prices and Package Bidding: A Theoretical and Experimental Analysis," American Economic Journal: Microeconomics, 2, August, 160-185

Kagel, John H., Yuanchuan Lien, and Paul Milgrom "Ascending Prices and Package Bidding: Further Experimental Analysis," *Games and Economic Behavior* 85, May 2014, 210-231.

Kagel, John H. and A.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.

Karp, Richard M. (1972). "Reducibility Among Combinatorial Problems", in R. E. Miller and J. W. Thatcher (editors): *Complexity of Computer Computations*. New York: Plenum, 85–103.

Katok, Elena and Anthony M. Kwasnica, "Time is Money: The effect of clock speed on seller's revenue in Dutch auctions," *Experimental Economics*, 2008, 11: 344-357.

Kessler, Judd B. and Alvin E. Roth (2012), "Organ Allocation Policy and the Decision to Donate," *American Economic Review*, Vol. 102 No. 5 (August), 2018-47.

Kessler, Judd B. and Alvin E. Roth (2014a), "Getting More Organs for Transplantation," *American Economic Review*, Papers and Proceedings, May,104 (5): 425-30.

Kessler, Judd B. and Alvin E. Roth (2014b), "Loopholes Undermine Donation: An Experiment Motivated by an Organ Donation Priority Loophole in Israel, "*Journal of Public Economics*, 114, June, 19-28.

Kittsteiner, Thomas, and Axel Ockenfels (2008). "On the Design of Simple Multi-unit Online Auctions," In: *Negotiation, Auctions and Market Engineering*, eds. Nick Jennings, Gregory Kersten, Axel Ockenfels, and Cristof Weinhardt. Lecture Notes in Business Information Processing, Berlin, Heidelberg: Springer, 68-71.

Kwasnica, Anthony M. and Elena Katok "The Effect of Timing on Bid Increments in Ascending Auctions," *Production and Operations Management* , 2007,16(4), pp. 483-494.

Kwasnica, Anthony, John O. Ledyard, David P. Porter, and Christine DeMartini. 2005. "A New and Improved Design for Multi-Object Iterative Auctions." *Management Science*, 51, 3, 419-34.

Kwerel, Evan, (2004), Foreword to Paul Milgrom, *Putting Auction Theory to Work*, Cambridge University Press, xv-xxi.

Ledyard, John O. (1995), "Public Goods: A Survey of Experimental Research," Chapter 2 in J.H. Kagel and A.E. Roth (eds.), *Handbook of Experimental Economics*, Princeton University Press, 111-194.

Ledyard, John, Robin Hanson and Takashi Ishikida (2008), An Experimental Test of Combinatorial Information Markets," *Journal of Economic Behavior and Organization*, forthcoming.

Ledyard, John O. Charles Noussair, and David Porter (1996), "The allocation of a shared resource within an organization," *Economic Design*, 2, 163-192.

Ledyard, John O., Mark Olson, David Porter, Joseph A. Swanson, and David P. Torma (2002), "The First Use of a Combined-Value Auction for Transportation Services," *Interfaces*, 32, 5 (September-October), 4-12.

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

Lee, Soohyung (2009) "Marriage and Online Mate-Search Services: Evidence from South Korea," University of Maryland working paper.

Lee, Soohyung and Muriel Niederle, "Propose with a Rose? Signaling in Internet Dating Markets," working paper, February 2014.

Leyton-Brown, Kevin, Eugene Nudelman, and Yoav Shoham (2006), "Empirical Hardness Models for Combinatorial Auctions," Chapter 19, in Peter Cramton, Yoav Shoham, and Richard Steinberg, editors (2006), *Combinatorial Auctions*, MIT Press, Cambridge, MA, 479-504.

Lucking-Reilly, David (1999), "Using field experiments to test equivalence between auction formats: Magic on the internet," *American Economic Review*, 89, 5, 1063-1080.

McMillan, John (1994), "Selling Spectrum Rights," *Journal of Economic Perspectives*, 8, 3 (Summer), 145-162.

McAfee, R. Preston, and John McMillan (1996): "Analyzing the Airwaves Auction," *Journal of Economic Perspectives*, 10, 159–175.

C. Nicholas McKinney, 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.
Merlob, Brian, Charles R. Plott and Yuanjun Zhang (2012), "The CMS Auction: Experimental Studies of a Median-Bid Procurement Auction with Nonbinding Bids," *The Quarterly Journal of Economics*, 127 (May), 793-827.

Milgrom, Paul (2004), *Putting Auction Theory to Work*, Cambridge University Press.

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

Myerson, Roger B., (1981), "Optimal Auction Design," *Mathematics of Operations Research*, 6, 1 (Feb.), 58-73

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

Niederle, Muriel, Deborah D. Proctor, and Alvin E. Roth, “The Gastroenterology Fellowship Match – The First Two Years,” *Gastroenterology*, 135, 2 (August), 344-346, 2008.

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, “The Gastroenterology Fellowship Market: Should there be a Match?,” *American Economic Review*, Papers and Proceedings, 95,2, May, 2005, 372-375.

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.

Ockenfels, Axel, David Reiley, and Abdolkarim Sadrieh (2007). “Online Auctions,” In: *Handbooks in Information Systems*, ed. Terrence J. Hendershott, ,, Economics and Information Systems, Vol. I, North Holland: Elsevier, 571-628.

Ockenfels, Axel and Alvin E. Roth (2006), "Late and Multiple Bidding in Second-Price Internet Auctions: Theory and Evidence Concerning Different Rules for Ending an Auction," *Games and Economic Behavior*, 55, 297-320

Ostrovsky, Michael and Michael Schwarz (2009), “Reserve Prices in Internet Advertising Auctions: A Field Experiment,”Working Paper, December 2009 (http://faculty-gsb.stanford.edu/ostrovsky/papers/rp.pdf)

Plott, Charles R. (1987), "Dimensions of parallelism: Some Policy Applications of Experimental Methods," in Alvin E. Roth, editor, *Laboratory Experimentation in Economics: Six Points of View*, Cambridge University Press, Cambridge, England, 193-219.

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

Plott, Charles R., Hsing-Yang Lee and Travis Maron, “The Continuous Combinatorial Auction Architecture,” *American Economic Review: Papers & Proceedings*, May 2014, 104(5): 452–456

Plott, Charles R. and David P. Porter (1996), “Market architectures and institutional testbedding: An experiment with space station pricing policies,” *Journal of Economic Behavior & Organization*, 31, 237-272.

Porter, David, Stephen Rassenti, Anil Roopnarine, and Vernon Smith. 2003. “Combinatorial Auction Design." *Proceedings of the National Academy of Sciences*, 100, 11153-57.

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.

Reiley, David H. (2006) “Field experiments on the effects of reserve prices in auctions: more *Magic* on the Internet,” RAND Journal of Economics, 37, 1, Spring 2006, 195-211.

Roth, A.E. "Let's Keep the Con Out of Experimental Econ." Empirical Economics (Special Issue on Experimental Economics), 1994, 19, 279-289.

Roth, Alvin E. (1995), “Introduction to Experimental Economics,” in J.H. Kagel and A.E. Roth (eds.), *Handbook of Experimental Economics*, Princeton University Press, 3-109.

Roth, Alvin E. and Axel Ockenfels (2002), "Last-Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet," *American Economic Review*, 92 (4), September, 1093-1103.

Roth, Alvin E., Tayfun Sönmez, M. Utku Ünver, Francis L. Delmonico, and Susan L. Saidman, “Utilizing List Exchange and Undirected Good Samaritan Donation through “Chain” Paired Kidney Donations,” *American Journal of Transplantation*, 6, 11, November 2006, 2694-2705.

Roth, Alvin E., Tayfun Sönmez and M. Utku Ünver “Efficient Kidney Exchange: Coincidence of Wants in Markets with Compatibility-Based Preferences,” *American Economic Review*, 97, 3, June 2007, 828-851.

Rothkopf, Michael H. (1969), “A Model of Rational Competitive Bidding,” *Management Science*, 15, 7 (March), 362-373.

Rothkopf, Michael H., Aleksandar Pekec and Ronald M. Harstad 1998 “Computationally manageable combinatorial auctions,” *Management Science* 44, 1131-1147.

Smith, Vernon L. 1962. "An Experimental Study of Competitive Market Behavior," *Journal of Political Economy*, 70, 111- 137.

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

Sönmez, Tayfun, and Utku Ünver. 2003. “Course Bidding at Business Schools.” Working paper, http://ideas.repec.org/p/wpa/wuwpga/0306001.html.

Sönmez, Tayfun, and Utku Ünver 2010. “Course Bidding at Business Schools.” International Econonic Review. 51 (1): 99–123

Spence, Michael (1973), “Job Market Signaling,” *Quarterly Journal of Economics*, Vol. 87, No. 3. (August), pp. 355-374.

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*, (June) 25: 1039-1080.

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

Varian, Hal R. (2007), "Position Auctions," *International Journal of Industrial Organization* 25, 1163-1178.

Varian, Hal R., and Christopher Harris. 2014. "The VCG Auction in Theory and Practice." *American Economic Review* Papers and Proceedings, 104(5): 442-45.

Vickrey, William (1961). "Counterspeculation, Auctions, and Competitive Sealed Tenders". *The Journal of Finance* 16 (1): 8–37

Voorhees, Josh (2009) “DOT scraps auction plan for NYC airports,” New York Times, May 13, http://www.nytimes.com/gwire/2009/05/13/13greenwire-dot-scraps-auction-plan-for-nyc-airports-19116.html

Wald, Matthew L. (2007) "Airlines at La Guardia Fight Bush Administration Proposal to Auction Off Landing Rights" New York Times, February 18, http://www.nytimes.com/2007/02/18/nyregion/18laguardia.html

Wald, Matthew L. (2008a) “U.S. Plans Steps to Ease Congestion at Airports,” New York Times, May 17, http://www.nytimes.com/2008/05/17/washington/17delay.html

Wald, Matthew L. (2008b) “Court Order Delays Auction of Landing Slots at Airports,” New York Times, Dec. 8, http://www.nytimes.com/2008/12/09/nyregion/09slots.html?ref=nyregion

Wilson, Robert B. (1967), “Competitive Bidding with Asymmetric Information,” *Management Science*, 13, 11 (July), 816-820.

Wilson, Robert B. (1969), “Competitive Bidding with Disparate Information,” *Management Science* 15, 7 (March), 446-448.

Zahavi, Amotz (1975). "Mate selection—a selection for a handicap". Journal of Theoretical Biology, 53 (1): 205–214