Turnaround Time and Bottlenecks in Market Clearing: Decentralized Matching in the Market for Clinical Psychologists

Turnaround Time and Bottlenecks in Market Clearing: Decentralized Matching in the Market for Clinical Psychologists

Alvin E. Roth

University of Pittsburgh

Xiaolin Xing

National University of Singapore

In the context of entry-level labor markets, we consider the *potential* transactions that have to be evaluated before equilibrium transactions can be identified. These potential transactions involve offers that are rejected. After an initial phase in which many offers can be proffered in parallel, subsequent potential transactions must be processed *serially*, since a new offer cannot be made until an outstanding offer is rejected. In this phase even a small time required to process offers and rejections may cause bottlenecks. In many, perhaps most, decentralized labor markets, this means that transactions have to be finalized before there is time for the market to clear, that is, before all the potential transactions that would need to be evaluated in order to reach a stable outcome can in fact be evaluated. This has implications for the strategic behavior of firms and workers. In particular, in deciding to whom to offer a position, a firm may have strong incentives to consider not only its preferences over workers but also the likelihood that its offer will be accepted, since if its offer is rejected it may find that many other potential employees have become unavailable in the interim. The analysis is carried out in connection with the decentralized

This work was partially supported by National Science Foundation grant SES-9121968. We are indebted to many people for information about the clinical psychology market, including Michael Carifio, Sue Campbell, Marsha Marcus, Ivan Mensh, Bob Perloff, Paul Pilkonis, Carl Zimet, and a number of participants in the market who wished to remain anonymous. We have also received useful advice from Vince Crawford, Jean Francois Richard, and Uri Rothblum.

[*Journal of Political Economy,* 1997, vol. 105, no. 2

market for clinical psychologists. The implications for other kinds of markets are considered.

This paper studies the procedures used to control and coordinate the timing of transactions in the (American) entry-level market for clinical psychologists. Transactions in this market are supposed to all be made by telephone on ''selection day,'' which is presently the second Monday in February, from 9:00 A.M. to 4:00 P.M. central standard time. The rules require that no offers be made before 9:00 A.M. and that all offers made during the course of the market and not yet rejected must remain open until 4:00 P.M. That is, both early offers and ''exploding offers'' (which require a decision before the end of the market) are not allowed. (The detailed rules, given in the Appendix, will be discussed later.)

Subject to many modifications of its rules, this kind of decentralized but uniform timing regime has been in use in this market since 1973. One kind of modification has concerned the length of the market, which is now only 7 hours. In the early 1970s the market lasted 5 days and was subsequently shortened to 3 days, and for most of the 1980s the rules specified that the market would take place from 8:00 A.M. Monday until noon the following day. This concern with the amount of time the market (and individual offers) should remain open is one that has been observed in many markets, and we shall consider why this is important.

We shall also compare the organization of this market with the more centralized organization of entry-level markets for American physicians (Roth 1984, 1986, 1996*a*, 1996*b*). One reason this comparison arises is that over the last 20 years, clinical psychologists have considered whether their own market might be better organized if they adopted the procedures used by physicians (see Roth and Xing 1994). The comparison is natural since there is a sense in which the procedures employed in these markets are very similar. However, we shall see that these apparently similar procedures operate very differently.

In both of these analyses—of the time the market remains open and of the comparison between the organization of the markets for psychologists and for physicians—it will turn out that a critical variable is the length of time it takes for an offer to be rejected and a new offer made. In the clinical psychology market, all interviewing is completed well before selection day, and because all participants come prepared to spend the day by the telephone, this time is extraordinarily short. On the basis of our (limited) site observations,

we have roughly approximated it as requiring only 6 minutes: only 1 minute for an offer to be rejected and only 5 minutes for a new offer to be conveyed to another candidate. We shall see that even this quick turnaround time creates bottlenecks because, during much of the time the market operates, offers must be processed *serially* rather than in parallel.

We shall argue that many markets, like the psychology market, go through an initial phase in which many potential transactions can be proffered and considered in parallel, but that eventually a high proportion of the potential transactions that remain have to be proffered and considered serially. It is in this latter phase that the turnaround time becomes the rate-determining factor.[1] It will also become apparent from our analysis that the amount of time a market remains open cannot be evaluated independently of the turnaround time required for an offer to be rejected and a new one issued. A market that is open for, say, 100 times as long as the psychology market but has a turnaround time longer than 10 hours (100 × 6 minutes) would in a critical sense be open for a shorter *effective* time.[2]

I. Institutions Related to Timing

In Roth and Xing (1994) we argued that one function of a market is to establish a time at which large numbers of buyers and sellers can plan to make transactions, but that establishing such a time can be difficult. We considered several dozen markets and submarkets that had experienced considerable difficulty in establishing a time at which transactions would take place. Most of these markets were

[1] As far as we know, turnaround time has not attracted prior attention as an aspect of market clearing. But it plays a large role in strategic models of bargaining, which have been incorporated in market models (cf. Rubinstein 1982; Osborne and Rubinstein 1990). Time plays an indirect role in these bargaining models, since attention is primarily given to equilibria in which transactions are made without delay (but in which the threat of delay influences transaction terms). In the markets we consider the delays are actually experienced.

[2] Economists at American universities may want to think of the market for new assistant professors of economics, in which a high proportion of transactions are made in the first few months of each calendar year. At many universities, each offer a department wishes to make requires separate approval by the dean, so the turnaround time at some universities is better measured in days than in hours. The effective length of this market, in terms of how many times a new offer can be made following a rejection by a previous candidate, is further shortened by the fact that there is no uniform time until which offers must be left open, so candidates who do not receive their most preferred offers in the first ''round'' may have already accepted less preferred offers before better offers might be forthcoming. This of course engenders strategic behavior on both sides of the market. We shall see that, even with a uniform time until which offers must be left open, related incentives for strategic behavior arise.

annual entry-level professional labor markets that had gone through a period in which, year after year, transactions were made earlier and earlier, often in a way that made the market at any moment very thin. In some cases this unraveling of transaction times proceeded to the point at which employers were hiring new employees up to 2 years before they would complete their professional qualifications and begin work. However, many markets that experienced such difficulties subsequently developed institutions to alleviate them. A number of markets around the world—for physicians, dentists, lawyers, and (recently) osteopaths—have adopted centralized market-clearing institutions that are organized roughly along the lines of the deferred acceptance algorithm, described next.[3]

The centralized part of such a market begins after applicants and employers have contacted and interviewed one another in the usual (decentralized) way. Each applicant then submits to a centralized clearinghouse a rank ordering, in order of preference, of each employer with which he or she has interviewed.[4] Similarly, each employer submits a rank ordering of all the applicants they have interviewed. (Leaving an applicant or employer off the preference list means that the worker or job is unacceptable.) These preference lists are then processed by an algorithm to produce a matching of applicants to positions.

Roth (1984, 1991) showed that the algorithms used in a number of the successful centralized market-clearing mechanisms are approximately the same as the deferred acceptance procedure first formally studied by Gale and Shapley (1962). Their procedure produces a matching of job seekers to jobs that is *stable* in terms of the submitted preferences in the sense that no student and hospital that are not matched to each other would prefer to be so matched.[5] In

[3] These markets have been analyzed as two-sided matching markets. See Roth and Sotomayor (1990) for an overview of the theory and Crawford (1991) for a paper that makes clear why such models are particularly suited to the analysis of labor markets. See Bergstrom and Bagnoli (1993) and Pollak (1994) for the use of matching models to study marriage markets, and Collins and Krishna (1993) for an analysis of matching procedures for Harvard dormitory rooms.

[4] The positions offered by each employer in such a market are divided into categories, if necessary, in which each position is identical (e.g., first-year general internal medicine). The salary is part of the job description, fixed in advance. So applicants do not have to negotiate once they are matched and can therefore determine their preferences in a noncontingent way.

[5] Roth (1984) showed that the algorithm adopted in the early 1950s to organize the American entry-level market for physicians (the National Resident Matching Program [NRMP]) was equivalent to the deferred acceptance procedure. By the early 1980s, the presence of married couples in the market had prompted changes in the algorithm that made this equivalence only approximate. Substantial further changes in this market in the late 1980s and early 1990s make the approximation rougher still, and they are discussed in Roth (1996) as part of a design effort commissioned by the NRMP.

the deferred acceptance algorithm, each employer begins by offering each of its positions to the candidates at the top of its preference list; that is, if it has k identical positions, it offers them to its top k candidates. Each candidate rejects any unacceptable offers, and any candidate who has received more than one offer rejects all but the most preferred (highest-ranked) of them, which is held without commitment. Following any rejections, each firm offers the position to its next-highest-ranked candidate who has not yet rejected it, as long as acceptable candidates remain. Each candidate who gets new offers compares them with any offer he may be holding and again rejects all but the most preferred. The procedure stops when no firm wishes to make any further offers, at which point each candidate accepts (and is matched to) the position (if any) that he is holding. Gale and Shapley (1962) showed that the matching produced by this procedure is a stable matching and that when all agents have strict preferences, it is *firm-optimal* among the stable matchings in the sense that no firm prefers any other stable matching.

In a centralized clearinghouse, the matching is performed by computer, using the preference lists submitted by the participants. We shall now turn to the clinical psychology market and see that it uses a very similar procedure, carried out not in a centralized way by computer, but in a decentralized way, over the telephone network. And the procedure used in the psychology market terminates in a very different way and at a matching different from that of the deferred acceptance procedure. This in turn has implications for the strategies and incentives facing the participants.

A. The Market for Clinical Psychologists

1. Market Rules and Their Evolution over Time

Clinical psychologists are employed as interns just prior to completing their doctoral training or as postdocs just after completing it. In recent years the market for these positions has involved just over 2,000 positions a year, offered at about 500 sites (see Roth and Xing [1994] and the references there). As noted earlier, uniform timing regimes have been mandated in this market since 1973, with a gradual shortening of the time the market is supposed to remain open, to its present length of 7 hours. The organization created to administer the market is the Association for Psychology Postdoctoral and Internship Centers (APPIC). The APPIC rules for the 1993 market are given in the Appendix.

The basic market structure is given by rules 3–6. Rules 3 and 4 control the timing of the market. Rules 5 and 6 specify that, while

the market is open, offers will be made and rejected according to a decentralized version of the deferred acceptance procedure. As in the case of the market for physicians, the salaries and general job descriptions associated with positions are specified in advance and are not variable parts of the offers.

Many of the rules contain additional clauses, added over time in response to complaints about how the rules are stretched or broken. (Examining the rules, and how they change over time, is one of the best ways to gain insight into the operation of a market.) One complaint is that applicants are subjected to a great deal of informal pressure to indicate in advance whether they will accept an offer, that is, to indicate in advance whether a particular employer is their first choice (see rule 3*c*). This is also a common complaint in centralized matching procedures (see Roth 1984, 1991), but in this decentralized market it has additional force. In the centralized markets, where matching is done by computer, it is not uncommon for students who feel unfairly pressured to say that they will rank some program first to say so, but then not do so. However, in the psychology market, to say that you will rank some program first is tantamount to a promise that when they call you on selection day you will accept their offer immediately. The virtually face-to-face nature of the telephone interaction, coupled with the fact that many psychology submarkets are small worlds, make this a difficult promise to renege on.[6]

2. A Site Visit

To put various kinds of behavior into perspective, it may help to recount the situation at an internship program we visited on selection day in 1993. This program had five positions and received 200 inquiries that turned into 71 completed applications. Invitations were issued to 30 candidates to come for interviews, and 29 accepted. On the morning of selection day, the two program codirectors, who would make the calls, came equipped with a rank-ordered list of 20 acceptable candidates from among those interviewed. The rank ordering was obtained from polling the psychologists on the staff, and it was understood that the codirectors had discretion about how to use it.[7] Prior to selection day, about half a dozen of the candidates

[6] As one program director said to us, ''you see these people again.''

[7] They could also use their judgment to modify the preferences and indicated that they would move the candidate ranked 12 ahead of numbers 10 and 11.

had indicated that the program in question was their first choice and that they would accept an offer immediately if one were made.[8]

On selection day the codirectors said that their general strategy was ''don't tie up offers with people who will hold them all day.'' They therefore decided to make their first offers (for their five positions) to numbers 1, 2, 3, 5, and 12 on their rank-order list, with the rationale being that numbers 3, 5, and 12 had indicated that they would accept immediately and that 1 and 2 were so attractive as to be worth taking chances on.[9] Two phones were used to make these calls, starting precisely at 9:00 A.M. central standard time. Candidates 3, 5, and 12 accepted immediately, as promised. Candidate 1 was reached at 9:05 (on the fourth attempt, after three busy signals) and held the offer until 9:13, when he called back to reject it. During this period, an incoming call (on a third phone whose number had been given to candidates) was received from the candidate ranked eighth, who now said that the program was her first choice. She was thanked and told she was still under consideration, and when candidate 1 called to reject the offer he was holding, the codirectors decided to make the next offer to candidate 8 (and not to number 4, as initially planned).[10] The offer to number 8 was then made and accepted immediately, and while that phone call was in progress, an incoming call from candidate 2 informed them that she had accepted another position. The decision was then made to offer the remaining position next to the highest-ranked remaining candidate who had indicated that he would accept immediately, number 10, and this offer was accepted at 9:21. After the briefest of celebrations, the codirectors called the remaining candidates to inform them that all positions were filled. These calls were completed by 9:35, 35 minutes after the opening of the market. The five positions were filled with the candidates initially ranked 3, 5, 8, 10, and 12.

Three things to note about this episode, which does not seem to be atypical, are the directors' concern not to make offers that ran the risk of being rejected late in the day, the consequent attention they paid to candidates who had indicated that they would immediately accept an offer from the program, and the willingness of candi-

[8] These directors indicated that they were careful not to pressure students to reveal their preferences, but that ''the savviest candidates always do'' reveal if a program is their first choice. We also heard of directors telling students that, while they would not ask if their program was the student's first choice, it would be most helpful to know should the student wish to say.

[9] Also, the candidate ranked number 1 was a minority candidate who, it was thought, would have many offers from top places and so would decide quickly.

[10] Note that this change of plans is something that can happen in a decentralized market in which firms make decisions sequentially, but not in the centralized markets in which firms submit an entire preference list at one time.

dates to convey such information. Note also that the time required to process offers is very short. In what follows, we shall analyze some computer simulations of the market, which show that even this short time has large consequences. But first we present a formal analysis that will provide a framework for comparison of these centralized and decentralized market institutions.

II. Comparison of Centralized and Decentralized Markets of Different Lengths

Several kinds of comparisons will be made in this section. Subsection *A* begins with some theoretical results comparing decentralized and centralized matching in markets that may differ in length. We can make welfare comparisons among different institutions, but these theoretical results do not allow us to predict the *magnitudes* of the welfare effects, which depend on the effective length of the market.

To consider magnitudes, we turn to simulation. The analytical results reveal that the welfare comparisons will come into play only when markets are too short to fully clear. Since this depends both on the duration of the market and on the length of time it takes to make offers and accept and reject them (and hence on the number of transactions that can be considered in the course of the market), when transaction times are very short (as in the psychology market), there might be no difference between centralized and decentralized market clearing. However, the simulations will demonstrate clearly that even short transaction times have very large effects.

The APPIC rules provide an unambiguous institutional structure for the simulations. The relatively arbitrary assumptions that must be made will concern the joint distribution of preferences of firms and workers. Subsection *B* will compare markets with and without fixed termination times, and different behavioral assumptions about the participants, under the assumption that the preferences of different firms and workers are uncorrelated. Subsection *C* will then consider the sensitivity of the results obtained to different assumptions about preferences (as well as to different concentrations of positions among firms) and show that the principal results of subsection *B* are robust.

It will help to keep in mind several differences between the centralized deferred acceptance algorithm and the decentralized procedure outlined in the APPIC rules. In a centralized market, participants must decide what preference lists to submit, after which offers, acceptances, and rejections are carried out automatically. But in the psychology market, participants do not submit preference lists; instead they can decide after each phone call what to do next, and

random events can determine the order in which offers are made, as when more than one firm attempts to telephone the same candidate at the same time and only one can get through. Also, when offers expire at 4:00 P.M., workers essentially must accept whatever offer they are holding (see rule 4*d*).[11] And (in contrast to the centralized market) when the organized part of the market ends, there may still be unmatched firms and workers who already know that they would prefer to be matched to one another, so the aftermarket in the decentralized case is very much a continuation of the original market, except with exploding offers that must be accepted (or rejected) immediately.[12]

A. *Random Matchings, Termination Times, and Aftermarkets: Some Formal Analysis*

1. An Analytical Framework: Definitions and Notation

For the static elements of our model, in which a firm may employ several workers but a worker may work for no more than one firm, we use the ''college admissions'' model as reformulated in Roth (1985) and Roth and Sotomayor (1990, chap. 5). The first elements of this model are two finite and disjoint sets, $\mathbf{F} = \{F_1, \ldots, F_n\}$ and $W = \{w_1, \ldots, w_m\}$, of firms and workers. For each firm F, there is a positive integer q_F, which indicates the number of (identical) positions F has to offer, that is, the maximum number of positions it may fill. (When we denote a particular firm by F_i, its quota of positions will be denoted q_i.)

An outcome is a matching of workers to firms, such that each worker is matched to at most one firm, and each firm is matched to at most its quota of workers. It will be convenient to denote a firm that has some number of unfilled positions as matched to itself in each of those positions, and similarly an unmatched worker will be matched to herself. To give a formal definition, we first define for any set X an *unordered family of elements* of X to be a collection of elements, not necessarily distinct, in which the order is immaterial.

We can now define a matching μ to be a function from the set $\mathbf{F} \cup W$ into the set of unordered families of elements of $\mathbf{F} \cup W$ such

[11] It is an equilibrium for job candidates to behave in this way because if all others do, then any candidate who allowed his offer to expire would face a market in which virtually all positions had been taken.

[12] The APPIC rules for the aftermarket say only that offers may have ''short but reasonable deadlines'' (rule 9*b*). This is a bit coy, in that offers made, say, 5 minutes *before* the deadline need remain open for only 5 minutes.

that (1) $|\mu(w)| = 1$ for every worker w and $\mu(w) = w$ if $\mu(w) \notin \mathbf{F}$; (2) $|\mu(F)| = q_F$ for every firm F; if the number of workers in $\mu(F)$, say r, is less than q_F, then $\mu(F)$ contains $q_F - r$ copies of F; and (3) $\mu(w) = F$ if and only if w is in $\mu(F)$. So $\mu(w_1) = F$ denotes that worker w_1 is employed at firm F at the matching μ, and $\mu(F) = \{w_1, w_3, F, F\}$ denotes that firm F, with quota $q_F = 4$, employs workers w_1 and w_3 and has two positions unfilled.

Each worker has preferences over the firms (and the possibility of remaining unmatched in the market), and each firm has preferences over the workers (and the possibility of leaving a position unfilled). All preferences are transitive. We shall write $F_i >_w F_j$ to indicate that worker w prefers F_i to F_j and $F_i \geq_w F_j$ to indicate that w likes F_i at least as well as F_j. Similarly, $w_i >_F w_j$ and $w_i \geq_F w_j$ represent firm F's preferences $P(F)$ over individual workers. Firm F is *acceptable* to worker w if $F \geq_w w$, and worker w is acceptable to firm F if $w \geq_F F$; that is, an acceptable firm is one that the worker prefers to being unmatched, and an acceptable worker is one that the firm prefers to leaving a position unfilled.

Each worker's preferences over alternative matchings correspond exactly to her preferences over her own assignments at the two matchings. Things are not quite so simple for firms, because even though we have described firms' preferences over workers, each firm with a quota greater than one must be able to compare groups of workers in order to compare alternative matchings. It will be sufficient for our purposes to assume merely that a firm's preferences over groups of employees it could be matched with (i.e., over groups of not more than q_F workers) are such that, for any two assignments that differ in only one worker, it prefers the assignment containing the more preferred worker (and is indifferent between them if it is indifferent between the workers). Any preferences of this sort are called *responsive* to the firm's preferences over individual workers (Roth 1985).

A matching μ is individually irrational if $\mu(w) = F$ for some worker w and firm F such that either the worker is unacceptable to the firm or the firm is unacceptable to the worker. Such a matching will also be said to be *blocked* by the unhappy agent. This reflects that the rules of the market allow every agent to withhold consent from such a match. Similarly, a firm F and worker w will be said together to block a matching μ if they are not matched to one another at μ, but would both prefer to be matched to one another than to (one of) their present assignments. That is, μ is *blocked by the firm-worker pair* (F, w) if $\mu(w) \neq F$ and if $F >_w \mu(w)$ and $w >_F \sigma$ for some σ in $\mu(F)$. (Note that either σ may equal some worker w' in $\mu(F)$ or, if one of firm F's positions is unfilled at $\mu(F)$, σ may equal F.) Matchings

blocked in this way by an individual or by a pair of agents are unstable in the sense that there are agents with both the incentive (because preferences are responsive) and the power (under rules that allow any firm and worker to conclude an agreement with each other) to disrupt such matchings. So we can now define a matching μ to be stable if it is not blocked by any individual or any firm-worker pair.[13]

Gale and Shapley (1962) showed that the set of stable matchings is always nonempty. Furthermore, when no agent is indifferent between any two mates, there exists for each side of the market (**F** or W) a stable matching (μ_F or μ_W) that is optimal for that side, in the sense that no agent on that side of the market prefers any other stable matching.

To consider random matching processes, we extend this framework slightly.[14] Define a *random matching* to be a random variable whose range is the set of all matchings. For each random matching $\underline{\underline{\boldsymbol{\mu}}}$, we obtain random variables $\underline{\underline{\boldsymbol{\mu}}}(v)$ for each agent v in $\mathbf{F} \cup W$, where each $\underline{\underline{\boldsymbol{\mu}}}(v)$ is the (random) assignment of v under $\underline{\underline{\boldsymbol{\mu}}}$. (The range of $\underline{\underline{\boldsymbol{\mu}}}(v)$ is $v \cup \mathbf{F}$ if v is in W or $v \cup W$ if v is in **F**.)

Given two random matchings $\underline{\underline{\boldsymbol{\mu}}}^1$ and $\underline{\underline{\boldsymbol{\mu}}}^2$ and a worker w with preferences P_w over $\mathbf{F} \cup w$, we say that $\underline{\underline{\boldsymbol{\mu}}}^2(w)$ stochastically P_w-dominates $\underline{\underline{\boldsymbol{\mu}}}^1(w)$ (and write $\underline{\underline{\boldsymbol{\mu}}}^2 \gg_w \underline{\underline{\boldsymbol{\mu}}}^1$) if, for every v in $\mathbf{F} \cup \{w\}$, $\Pr\{\underline{\underline{\boldsymbol{\mu}}}^2(w) >_w v\} \geq \Pr\{\underline{\underline{\boldsymbol{\mu}}}^1(w) >_w v\}$; that is, for any level of satisfaction the probability that w's match exceeds that level of satisfaction is greater under the random matching $\underline{\underline{\boldsymbol{\mu}}}^2$ than under $\underline{\underline{\boldsymbol{\mu}}}^1$. So if $\underline{\underline{\boldsymbol{\mu}}}^2 \gg_w \underline{\underline{\boldsymbol{\mu}}}^1$, then any utility maximizer with ordinal preferences P_w prefers $\underline{\underline{\boldsymbol{\mu}}}^2(w)$ to $\underline{\underline{\boldsymbol{\mu}}}^1(w)$.

2. Decentralized Deferred Acceptance with Random Elements and Termination Time (with and without an Aftermarket)

Figure 1 presents, as a flowchart, a very general model of the deferred acceptance procedure, consistent with the APPIC rules. It may have random elements, and they may involve arbitrary probability distributions. There may be an arbitrary termination time t^* (beyond which the acceptance of offers may not automatically be deferred) or no fixed termination time at all (if $t^* = \infty$), in which case the deferred acceptance procedure continues until no firm wishes to make any more offers. If there is a fixed termination time, there

[13] This definition of stability appears to account only for coalitions of size 1 or 2 but in fact accounts for coalitions of any size; i.e., stable matchings are in the core (see Roth and Sotomayor 1990).

[14] For other uses of random models of matching, see Roth and Vande Vate (1990, 1991), Blum, Roth, and Rothblum (in press), and Roth and Rothblum (1996).

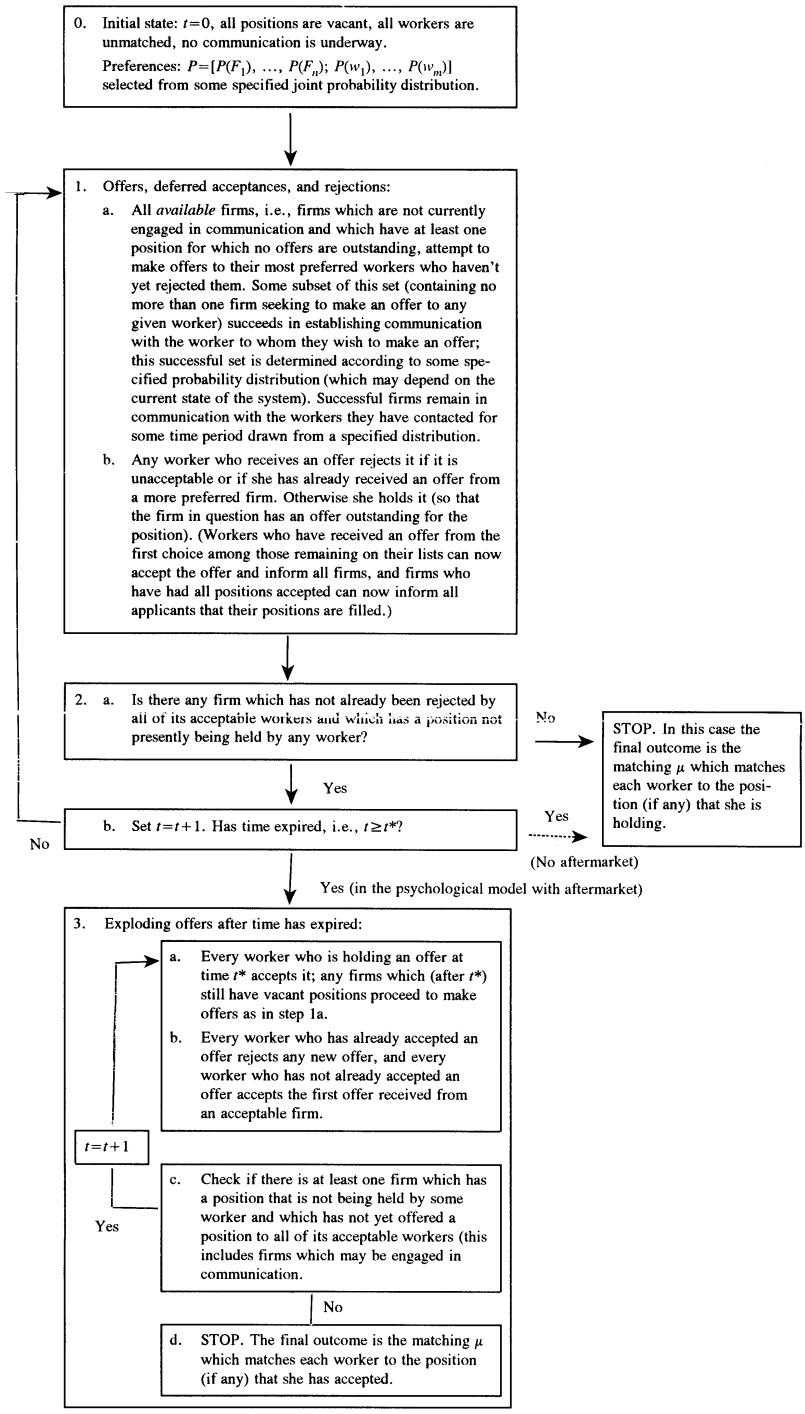

FIG. 1.—Decentralized deferred acceptance with random elements and termination time (with and without an aftermarket).

may be an aftermarket (as in the psychology market) or none. And, since we may not be able to observe the preferences in the market, they too may be regarded as random variables.

Box 0 of the flowchart begins the process with all positions vacant and all workers unmatched. The (random) preferences of the agents, chosen here, are fixed throughout the remainder of the process.

Box 1 models the deferred acceptance procedure, by telephone, in which a random element is introduced by the fact that only one firm may speak to a worker at a time. The possible reasons for termination are modeled in boxes $2a$ and $2b$ either because all offers have been exhausted (box $2a$) or because time has run out (box $2b$). Note that the centralized deferred acceptance algorithm terminates only when all offers are exhausted, so it would be modeled here with a termination time $t^* = \infty$ in box $2b$. If the psychology market terminates because all offers are exhausted, then there is no need for an aftermarket. However, if the deferred acceptance part of the psychology market terminates because time has run out, then the aftermarket opens, in which acceptances can no longer be deferred because offers are now exploding. Here too the telephone network introduces a random element, precisely as in box 1.

For comparison purposes only, it will be convenient to consider also an artificial market in which there is a termination time but no aftermarket (indicated by the dotted line from step $2b$).

We shall use the model given in figure 1 to compare centralized and decentralized matching and to consider the effects of changing the length of the market. It will help clarify matters to consider first how to compare the preferences agents submit in the centralized procedure with the choices we might observe them make in the decentralized procedure. A firm that makes offers to several workers in the decentralized procedure can be said to prefer them in the order in which the offers are made, and a worker who rejects an offer while holding another can be said to prefer the offer held. However, this will typically yield only a partial order, for two reasons. First, the procedure will typically terminate before full preference lists are revealed. Second, if, for example, a worker holds an offer from firm F while rejecting one from F' and then receives and rejects an offer from F'' while continuing to hold F, this ''reveals'' an ordering in which F is preferred to both F' and F'' but gives no information about comparisons between F' and F''. Thus there may be more than one preference relation consistent with the revealed preferences for each such agent.[15] But it is not hard to see that the outcome

[15] The revealed part of the preferences may also be ''overrevealed'' in the sense that the revealed preferences are always strict, but an agent who is in fact indifferent

of the deferred acceptance procedure is insensitive to those parts of the preferences that are not revealed, that is, that the outcome is the same for any preferences consistent with those that are partially revealed by the decentralized procedure. In addition, *if there is no fixed termination time,* we can state the following theorem. (Note that the *revealed* preferences are always strict, whether or not the underlying preferences are.)

THEOREM 1. If the decentralized deferred acceptance procedure is run without any fixed termination time (i.e., $t^* = \infty$), then the outcome would be the same stable matching as that produced by the centralized deferred acceptance procedure. In particular, both procedures produce the firm-optimal stable matching with respect to the revealed preferences, μ_F.

Sketch of the proof. We need to show that the outcome of the decentralized procedure is not influenced by the random elements of step 1*a*, but is always a fixed matching (as a function of the realized preferences). (When $t = \infty$, the random elements of the aftermarket [step 3*a*] are never reached.) Then we have to show that this fixed matching is always the firm-optimal stable matching with respect to the revealed preferences. Both elements of the proof are almost immediate from the standard proof that the centralized deferred acceptance procedure (with no random elements) produces the matching μ_F. First, the outcome of the decentralized procedure is stable with respect to the realized preferences because (regardless of the order in which firms have made offers) there can be no blocking pairs for the final matching. For if a firm prefers some worker to one of its matched employees, it must have already proposed to that worker and been rejected. Second, no firm is ever rejected by a worker to whom it could be matched at some stable matching. (This follows in the standard way by induction; see, e.g., Roth and Sotomayor [1990, p. 33].) Thus the final matching is always the firm-optimal stable matching, and so the outcome is independent of the random elements in step 1*a*. Q.E.D.

Describing the outcome in terms of the revealed preferences raises the question of whether the agents have incentives to behave strategically in ways that make their revealed preferences different from their true preferences (i.e., from the preferences they would use to choose an outcome if they could do so as a single-person decision rather than through a complex strategic interaction). The deferred acceptance algorithm has received a good deal of study from this point of view (cf. Roth and Sotomayor 1990). The simplest case to summarize is the one in which each firm seeks only one

---

between two choices may have chosen arbitrarily between them. However, this applies equally to the centralized and decentralized procedures.

worker. In this case it is a dominant strategy for firms, but not for workers, to reveal their true preferences; yet all equilibria in undominated strategies produce outcomes that are stable with respect to the true preferences (even though workers may not have revealed them). But as we shall see, a fixed time limit in the decentralized procedure (as in the APPIC rules) gives agents on both sides of the market reasons to behave strategically. The remaining theorems of this section, which are stated in terms of straightforward (truthful preference revealing) play, will set the stage for this observation by showing that straightforward play has different consequences for different termination times.

In the markets modeled by the flowchart in figure 1, when there is a finite termination time, the random elements matter, and we get random matchings, which depend on the termination time. It will be easiest to understand this by considering first a hypothetical market run without an aftermarket.

THEOREM 2. For markets in which there is *no* aftermarket, let $\tau < \sigma < \infty$, and let $\underline{\underline{\mu}}^{\tau}$, $\underline{\underline{\mu}}^{\sigma}$, and $\underline{\underline{\mu}}^{\infty}$ be the random matchings that result from straightforward play in otherwise identical decentralized deferred acceptance procedures having termination times τ, σ, and ∞, respectively. For any worker w with realized preferences P_w, $\underline{\underline{\mu}}^{\infty}(w) \gg_w \underline{\underline{\mu}}^{\sigma}(w) \gg_w \underline{\underline{\mu}}^{\tau}(w)$.

Proof. Because of the finiteness of the sets of firms and workers, there are only finitely many sample paths that can be realized even in the procedure with $t^* = \infty$. (Every sample path will terminate in finite time.) Because the three procedures that give rise to the random matchings $\underline{\underline{\mu}}^{\tau}$, $\underline{\underline{\mu}}^{\sigma}$, and $\underline{\underline{\mu}}^{\infty}$ are identical except for their termination times, we can consider each sample path that has a positive probability of occurring in the $t^* = \infty$ procedure and observe that the part of this path that is realized up to time τ has an equal chance of occurring in the procedures with $t^* = \tau$, σ, and ∞. Similarly, the part of this path that is realized up to time σ has an equal chance of occurring in the procedures with $t^* = \sigma$ and ∞. On such a path the set of offers that have been made to each worker, and rejected or held, is the same up to time τ in all three procedures and up to time σ in the $t^* = \sigma$ and $t^* = \infty$ procedures. To put it another way, in comparisons of two random procedures that are identical before the termination time of one of them, each sample path of the shorter procedure corresponds to a family of sample paths of the longer procedure that are all identical up to the earlier termination time.

Consider now a worker w in the $t^* = \tau$ procedure. If the sample path we are considering terminates before time τ, then w receives the same match in procedures with $t^* \geq \tau$. So in what follows we are

free to concentrate on those sample paths in which the procedure terminates at τ because of lack of time (i.e., terminates because of rule 2*b* in the flowchart of fig. 1). If w is holding an offer from firm F_i at time τ, then $\underline{\underline{\mu}}^{\tau}(w) = F_i$. The probabilities $\Pr\{\underline{\underline{\mu}}^{\sigma}(w) \geq_w F_i\} = \Pr\{\underline{\underline{\mu}}^{\infty}(w) \geq_w F_i\} = 1$ because in the procedures that continue after time τ, worker w holds the offer from F_i to the end unless she receives a preferable offer. And if w is holding no offer at time τ, then because there is no aftermarket, w is unmatched; that is, $\underline{\underline{\mu}}^{\tau}(w) = w$. So it follows trivially that $\Pr\{\underline{\underline{\mu}}^{\sigma}(w) \geq_w w\} = \Pr\{\underline{\underline{\mu}}^{\infty}(w) \geq_w w\} = 1$. We can compare $\underline{\underline{\mu}}^{\sigma}(w)$ and $\underline{\underline{\mu}}^{\infty}(w)$ in the same way. Since $\underline{\underline{\mu}}^{\infty}(w) \geq_w \underline{\underline{\mu}}^{\sigma}(w) \geq_w \underline{\underline{\mu}}^{\tau}(w)$ on each sample path, it follows that $\underline{\underline{\mu}}^{\infty}(w) \gg_w \underline{\underline{\mu}}^{\sigma}(w) \gg_w \underline{\underline{\mu}}^{\tau}(w)$. Q.E.D.

We note with the following counterexample, however, that the comparison between the market with $t^* = \infty$ and the psychology market is not so simple.

COUNTEREXAMPLE. For a market *with* an aftermarket, it is not the case that if $\tau < \sigma$, then $\underline{\underline{\mu}}^{\sigma}(w) \gg_w \underline{\underline{\mu}}^{\tau}(w)$.

Proof. Let $\mathbf{F} = \{F_1, F_2\}$, $W = \{w_1, w_2\}$, and the joint distribution of preferences be such that the two firms always have the preference $w_1 >_F w_2$ for F in $\mathbf{F}$ and the two workers always have the same preferences: either $F_1 >_w F_2$ for each w in W or the reverse. Then at the firm-optimal stable matching, w_1 is matched to the most preferred firm; so this is the outcome in the deferred acceptance process with $t^* = \infty$. Now consider $t^* = \tau < \infty$ with τ small enough that there is time for only one offer to reach w_1. So there is a positive probability that w_1 will have received an offer only from the less preferred firm at time τ, that is, a positive probability that $\underline{\underline{\mu}}^{\tau}(w_1)$ is the less preferred firm and $\underline{\underline{\mu}}^{\tau}(w_2)$ is the more preferred firm. Then $\underline{\underline{\mu}}^{\infty}$ does not stochastically dominate $\underline{\underline{\mu}}^{\tau}$ from w_2's point of view. Q.E.D.

Although the counterexample shows that theorem 2 will not apply directly to the comparison between the centralized and decentralized markets, the following theorem shows that a conditional version applies. To state the next theorem, define $w(t)$ to be the number of acceptable offers worker w has received up to time t in a given run of the decentralized procedure.

THEOREM 3. *Conditional* on having received at least one acceptable offer by time τ, the distribution of $\underline{\underline{\mu}}^{\infty}(w)$ stochastically dominates that of $\underline{\underline{\mu}}^{\sigma}(w)$, which in turn stochastically dominates that of $\underline{\underline{\mu}}^{\tau}(w)$ for $\tau < \sigma < \infty$. That is, for every v in $\mathbf{F} \cup \{w\}$,

$$\Pr\{\underline{\underline{\mu}}^{\infty}(w) >_w v | w(\tau) \neq 0\} \geq \Pr\{\underline{\underline{\mu}}^{\sigma}(w) >_w v | w(\tau) \neq 0\}$$

$$\geq \Pr\{\underline{\underline{\mu}}^{\tau}(w) >_w v | w(\tau) \neq 0\}.$$

Proof. The counterexample shows that the part of the proof of theorem 2 that does not go through when there is an aftermarket is the case in which w is holding no offer at time τ. But conditional on having received an offer by time τ, worker w will be holding an offer from some firm F_i at time τ, and the proof of theorem 2 for that case constitutes the proof of theorem 3. Q.E.D.

There are no welfare comparisons for firms similar to those for workers given by theorems 2 and 3 (in particular the opposite stochastic dominance relations do not hold for firms) because it can always happen that an offer is rejected just before the termination time, so the firm has an empty position when the deferred acceptance part of the market ends. If, instead of terminating at that moment, the market were to continue, such a firm could do better than if the market were to stop completely. But firms not in this situation always do better with shorter termination times. To state this formally, define $F(t)$ to be the number of workers holding offers from firm F at time t.

THEOREM 4. Let $\tau < \sigma < \infty$, and let $\underline{\underline{\boldsymbol{\mu}}}^{\tau}$, $\underline{\underline{\boldsymbol{\mu}}}^{\sigma}$, and $\underline{\underline{\boldsymbol{\mu}}}^{\infty}$ be the corresponding random matchings resulting from straightforward play. Then conditional on all its positions being held at time τ, the distribution of $\underline{\underline{\boldsymbol{\mu}}}^{\tau}(F)$ stochastically dominates that of $\underline{\underline{\boldsymbol{\mu}}}^{\sigma}(F)$, which stochastically dominates that of $\underline{\underline{\boldsymbol{\mu}}}^{\infty}(F)$, from the point of view of a firm F with realized preferences (over individuals) $P(F)$ and responsive preferences over groups of workers. That is, for any feasible assignment of workers $\mu(F)$,

$$\Pr\{\underline{\underline{\boldsymbol{\mu}}}^{\tau}(F) >_F \mu(F) \,|\, F(\tau) = q_F\} \geq \Pr\{\underline{\underline{\boldsymbol{\mu}}}^{\sigma}(F) >_F \mu(F) \,|\, F(\tau) = q_F\}$$

$$\geq \Pr\{\underline{\underline{\boldsymbol{\mu}}}^{\infty}(F) >_F \mu(F) \,|\, F(\tau) = q_F\}.$$

The essential element of the proof, after which the argument is the same as for theorems 2 and 3, is that on any sample path, since F has responsive preferences (and since $F(\tau) = q_F$), firm F prefers $\underline{\underline{\boldsymbol{\mu}}}^{\tau}(F)$ to any set of workers that hold its offers at some later time $t > \tau$. Q.E.D.

Theorem 4 is much more delicate than theorem 3. The comparison in theorem 3 is for workers who have received at least one offer *by* time τ, but theorem 4 concerns firms that have all their offers held *at* time τ, a status that could evaporate at any instant if the market were to continue.

Note also that theorems 3 and 4 present comparisons between the distributions of the matchings resulting from markets with different termination times, but do not tell us anything about the *magnitudes*

of these comparisons. If the transaction times are so fast that the market terminates because all offers are exhausted before the deadline τ, then theorem 1 tells us that the distributions compared in theorems 3 and 4 will in fact be equal. And if, instead, the market deadline was very short compared to the number of transactions required even to get offers to most workers, then the events on which the probability comparisons in the two theorems are conditioned would rarely occur.

To put it another way, theorems 3 and 4 could have been stated in terms of changes in the time needed to accept and reject offers rather than changes in the total time available for the deferred acceptance part of the market. What matters is the effective length of the market, that is, how many of the transactions needed for the market to fully clear can in fact be completed before the termination time. Simulations will allow us to see that with communication times like those observed in the psychology market, the market will generally not terminate before time runs out, but most workers will have received an offer and most firms will have all their positions held when the deferred acceptance part of the market ends. Therefore, the unconditional distributions will have essentially the same relationship as the conditional distributions described in the theorems.

B. *Some Market Simulations with Uncorrelated Preferences*

Both the psychology and medical markets contain partially overlapping specialized submarkets of different sizes (cf. Roth and Xing 1994). To cleanly compare the effect of the different *procedures* used in these markets, we begin by considering a representative ''generic'' market or submarket consisting of 200 potential employees and 50 employers, each with four positions to fill. Each worker has preferences over 20 randomly selected employers, such that he is equally likely to prefer the employers in any order. Each employer has preferences over all the workers who apply to it (i.e., over all workers such that the employer is included in the worker's preferences), and each employer is equally likely to prefer the workers who have applied to it in any order.[16] (For later comparisons we shall

[16] So an employer appears on a student's preference list if and only if the student appears on the employer's preference list. This is essentially an implementation of APPIC rule 2, which has the effect that a student is not put in the position of waiting for an offer that has no possibility of being made. The expected number of applicants to each firm in this simulation is $(2/5)200 = 80$. The parameters make the simulated market larger than a typical psychology specialty submarket and smaller than a typical medical specialty submarket.

also consider the case in which all workers have preferences over all firms, and vice versa.)

Actions in the simulation occur each minute. Each employer has two phones, one for outgoing and one for incoming calls. (The site we discussed earlier had two phones for outgoing calls and one for incoming calls. We shall see that our results are sensitive to the number of phones only in the opening hours of the market—which constitute its parallel processing phase—and this sharply limits the maximum potential effect of increasing the number of phones.) Students are modeled as having one telephone, used for both incoming and outgoing calls. Except when specifically indicated otherwise, calls initiated by employers for the purpose of making offers last 5 minutes, and all other calls last 1 minute.

In what follows we report the results of a number of simulated markets, which differ both in their rules for termination and in the assumptions we make about the behavior of participants on both sides of the market. It will be simplest to describe all these results by first considering the simplest model: the deferred acceptance procedure conducted by telephone, in which the market terminates only when all transactions are completed and in which employers and students decide which offers to make, accept, hold, and reject by straightforwardly consulting their preferences without delay. As indicated above, the outcome of the market in such a case will correspond to the outcome using the centralized deferred acceptance algorithm with the same preferences. For this reason we refer to this model as the medical model (although keep in mind that the modern American medical market has complications that require significant modifications of the algorithm and change the properties of its outcome; see Roth [1995, 1996*b*]). By following the timing of events in this deferred acceptance algorithm conducted by telephone, we shall provide a basis of comparison for the psychology market, with its fixed termination time.

1. The Simulated Medical Model Telephone Market

The simulation of the medical model is as follows. When the market opens, each employer places a call to its top-ranked candidate. (When multiple employers place simultaneous calls to the same student, one selected at random is connected and the others receive busy signals.) A phone conversation initiated by an employer to convey an offer takes 5 minutes to complete (so the phone is busy for 5 minutes). If the call comes from the student's first-choice employer (or from his first-choice remaining employer after more preferred

employers have called to inform the student that all their positions have been filled; see below), then the student accepts the offer in the course of the phone call. If the student has already received an offer from a preferred employer (whether this offer has already been accepted or is being held), then the student rejects the offer in the course of the phone call. Otherwise (i.e., if the offer is the best the student has so far received but there are more preferred employers that have not yet announced that all their positions are filled) the student ''holds'' the offer just received.

Employers who have just spoken to a candidate immediately call their next most preferred candidate if they have any positions remaining that neither have been accepted nor are being held. (Employers who have received a busy signal and who have only a single position available for which they have not already made an offer that is outstanding continue to try to place the call until it goes through. If they have more than one vacant position, they call the other candidates for those positions before returning to the busy candidate.) If, following a phone call, all an employer's positions have been accepted, the employer immediately calls all applicants with whom it has not yet communicated to inform them of this. (See rule 10*a*. These information calls take 1 minute.) If, following a call, an employer has some offers on hold (and no positions that have not been either accepted or held), then it waits and initiates no further calls until it receives one from a student holding one of its offers.

Students who receive an offer from an employer that they prefer to an offer they are already holding hold the new offer (or accept it if it comes from their first-choice employer that has not yet announced that its positions are filled) and immediately call the employer whose offer they were already holding to reject that offer (rule 6*a*). If they have accepted the offer, they also call all the employers on their preference list whose offers have not already been rejected to report that they have now taken an offer (rule 8). The employer whose offer was rejected immediately calls the highest-ranked student on its preference list who has not previously rejected it or called to announce that another offer has been accepted, and conveys an offer to that student.

In the medical model there is no fixed termination time. Instead, the process terminates whenever no student is holding two offers and no employer still has an offer to make. The resulting outcome matches each student who has accepted an offer or is holding one when the process terminates to the corresponding employer.

Results of the medical model telephone simulations.—The medical model simulation is an implementation of the deferred acceptance procedure, and its outcome is the employer-optimal stable match-

TABLE 1

MEDICAL MODEL TELEPHONE MARKET: RESULTS OF 100 SIMULATIONS FOR EACH OF THREE TURNAROUND TIMES

	NUMBER OF MINUTES REQUIRED TO MAKE AN OFFER (and Reject One)		
	5 1	10 2	25 5
	A. Preferences over 20 Firms; Uncorrelated Random Preferences		
Mean time to termination at a stable outcome	18:18 (8:10)	36:32 (16:20)	91:14 (40:52)
Median time to termination	16:24	32:39	81:19
Mean time by which 90% of students have received an offer	1:02	2:03	5:04
Mean time by which 99% of students have received an offer	5:19	10:35	26:22
Longest time to termination	39:25	78:25	196:22
Shortest time to termination	4:59	9:55	25:00
	B. Preferences over All 50 Firms; Uncorrelated Random Preferences		
Mean time to termination at a stable outcome	22:53 (12:03)	45:35 (24:04)	113:42 (60:12)
Median time to termination	18:57	37:44	94:09
Mean time by which 90% of students have received an offer	1:09	2:15	5:35
Mean time by which 99% of students have received an offer	7:02	13:55	34:39
Longest time to termination	55:15	110:03	274:48
Shortest time to termination	6:10	12:12	30:50

NOTE.—Standard deviations are in parentheses.

ing. The simulation allows us to observe how the offers are made over time. Column 1 of table 1 (panel A) shows the results of 100 simulations of this process when, as described above, offers take 5 minutes and rejections take 1 minute.

The first thing to notice is that it is time-consuming to run the deferred acceptance procedure by telephone: the mean time to achieve a stable outcome is over 18 hours (and the median time is over 16 hours). The time required in any particular simulation depends a great deal on the particular preferences, as shown by the fact that the standard deviation of these termination times is also high, at just over 8 hours. But even those simulations that are a full standard deviation faster than the mean require 10 hours to produce a stable match. What is taking all this time?

The next entries in column 1 of table 1 (panel A) indicate that the problem is not that it takes a long time to make initial contact with the bulk of the students. The average time required before 90 percent of the students have received at least one offer is barely more than 1 hour, and the average time before 99 percent of the students have at least one offer is less than 5½ hours. We are dealing with a population of 200 students per market, so when 99 percent of the students have received offers, only two students still do not have an offer. In these simulations there are exactly as many students as positions, so the process terminates as soon as all 200 students have received offers. What can we make of the almost 13 hours (18:18 − 5:19) that it apparently takes, on average, for the last two students to get offers?

To understand what is going on, we shall examine directly the hourly progress of the market in terms of offers made, rejected, held, and so forth. But note first that the rate-determining factor in these simulated markets is the turnaround time it takes for an offer to be rejected and a new offer to be made. To see this, look at columns 2 and 3 of table 1. They report the same 100 simulations (i.e., begun with the same random preferences) as in column 1, but with a doubling and quintupling, respectively, of the turnaround times. This is achieved by doubling or quintupling both the time required to make an offer and the time required to reject an offer that is being held. The times for all other events—information calls from students to employers or from employers to students, and the time needed to redial after busy signals—remain constant at 1 minute. So if the time required for information calls, for example, played an important role in determining the rate at which offers were made, the numbers in column 2 would be substantially less than twice those in column 1, and the numbers in column 3 would be substantially less than five times those in column 1. But this is not what we see. The figures in these two columns are very close to two and five times the column 1 figures. (And panel B shows that there are no important differences when students have preferences over all firms, and vice versa.)

Of course busy signals could still potentially play a large role because when we increase the length of calls, we increase the number and duration of busy signals. But it turns out that there is a strict limit on how much the market can be sped up by reducing busy signals (e.g., by having all employers and students represented by staffs of telephonists at multiple phones). To see why and to answer the questions raised above, we need to look at the simulations in more detail.

Looking first at the very last row of table 2, we see that the longest

TABLE 2

HOURLY PROGRESS OF THE MEDICAL MODEL TELEPHONE MARKET (100 Simulations)

Hour	Number of Students with at Least One Offer	Number of Students with an Offer from the Firm They Will Ultimately Match With	Number of Offers Made	Number of Offers Not Rejected Immediately
0	.00	.00	.00	.00
1	178.47	86.32	400.08	278.06
2	191.24	116.06	531.96	333.90
3	194.83	132.75	602.36	360.04
4	196.50	143.81	648.58	375.70
5	197.41	152.14	681.79	386.80
6	198.02	158.48	707.38	395.01
7	198.37	163.37	727.89	401.10
8	198.54	167.66	745.23	406.29
9	198.68	171.46	761.06	410.70
10	198.84	174.77	775.07	414.65
11	198.97	177.59	787.29	417.85
12	199.05	180.32	798.49	421.03
13	199.18	182.78	808.49	423.75
14	199.29	184.76	817.30	425.99
15	199.41	186.72	824.77	428.12
16	199.44	188.26	831.41	429.84
17	199.51	189.75	837.30	431.51
18	199.57	191.04	842.61	432.89
19	199.62	192.19	847.21	434.17
20	199.67	193.11	851.38	435.20
21	199.69	193.91	854.99	436.09
22	199.71	194.70	858.47	436.96
23	199.76	195.47	861.63	437.78
24	199.77	195.98	864.35	438.32
25	199.80	196.56	866.80	438.92
26	199.82	197.07	869.00	439.45
27	199.85	197.52	870.94	440.24
28	199.87	197.81	872.50	440.63
29	199.87	198.18	873.98	440.93
30	199.90	198.47	875.13	441.23
31	199.90	198.77	876.16	441.49
32	199.90	199.01	877.22	441.73
33	199.91	199.23	878.12	441.87
34	199.92	199.37	879.03	442.03
35	199.93	199.53	879.80	442.18
36	199.94	199.67	880.40	442.28
37	199.94	199.77	880.99	442.40
38	199.97	199.89	881.39	442.40
39	199.97	199.95	881.62	442.46
40	199.99	199.99	881.71	442.50

of the 100 simulations terminated by the fortieth hour, at which point the average number of students who had received at least one offer was 199.99. This reflects that in one simulation only 199 students were matched at the employer-optimal stable matching; in all the other simulations all 200 students were matched. Column 2 shows that, of course, by the time the markets had terminated, every student had received an offer from the employer with whom he was ultimately matched. Column 3 shows that, on average, 882 offers were made to reach the stable outcome (i.e., 4.4 offers per position, or almost 18 offers per firm), and column 4 shows that almost exactly half of these offers were rejected immediately whereas half were held for at least some time.

Now looking at the top of table 2, we see that 45 percent (400/882) of the average number of offers eventually made in 40 hours were in fact made in the *first* hour. Since each offer takes 5 minutes, this means that many offers were made in parallel. On average, just over 178 distinct students received offers in the first hour, that is, almost 90 percent (recall from panel A of table 1 that the 90 percent mark is actually reached at 1:02). But fewer than half of the students who received offers in the first hour (and at 86/200 only 43 percent of all students) had yet heard from the employer to whom they would ultimately be matched. So the market still has considerable work to do after the first hour.

However, the pace at which it accomplishes this work slows down dramatically. The reason is that, after the first hour, most firms have already offered all four of their positions to someone and must wait for a rejection before they can make any new offers. So on average only 132 new offers (532 − 400) are made from hour 1 to hour 2, and they reach only 13 of the students who had not yet received any offer (191 − 178). And only 30 of these new offers reach students who will ultimately accept them (i.e., at the end of hour 2, only 116 students have received an offer from the employer to whom they would ultimately be matched). So there is still much further to go before the market clears.

The rate at which offers are made slows still further as the market progresses: on average, only 70 offers (602 − 532) are made from hour 2 to hour 3, only 47 (649 − 602) from hour 3 to hour 4, and only 33 (682 − 649) from hour 4 to hour 5. By hour 5 almost 99 percent of the students have received at least one offer. But while only 1 percent of the students have yet to receive an offer at hour 5, 24 percent of the students ([200 − 152]/200) have yet to receive an offer from the employer with whom they will be matched when the market clears and a stable matching has been reached.

Recall that when 99 percent of the students in these particular

markets have received an offer, at most two students can be holding more than one offer. Thus (at a moment when this is the case) there will be only two calls being placed, after which at most two firms will have vacancies; so only two more calls will be placed and so forth. And when 99.5 percent of the students have received offers (which happens, on average, after hour 8), there will be only one phone call going at a time. Only between 10 and 12 new offers can be made per hour at this point (since offers that are rejected immediately have a turnaround time of 5 minutes, whereas an offer that causes another offer to be rejected causes a new offer to be completed after only 6 minutes). But after hour 8 there remain, on average, 137 offers (882 − 745) to be made before the market clears.[17] Thus the bulk of the time before the market ends is spent when offers must be made serially.

By the end of hour 7, the time at which the psychology market closes, on average fewer than 2 percent of the students in the medical model simulations do not have offers. But 16 percent of the students have yet to receive the offers they would finally accept. We turn next to consider the consequence of terminating the market at this point. We begin with a simulation in which all agents continue to consult their preferences straightforwardly and without delay.

2. The Simulated Psychology Market with Straightforward Behavior

These simulations of the psychology market follow exactly the rules of the previous simulations, up until the end of hour 7. At that point, all offers that are still being held are accepted (by default, which takes no time), and any firm that has not filled one of its positions continues to call the candidates remaining on its list of preferences until an unmatched candidate accepts the offer or until it runs out of candidates to call. Any unmatched candidate accepts the first offer he receives.[18]

Column 1 of table 3 gives the results of these simulations. Naturally, the times at which 90 percent and 99 percent of the students have received at least one offer are like those in the previous simulations, and well before the hour 7 deadline. But since all outstanding offers are accepted after the deadline is reached, the resulting

[17] Note that the average number of offers made in col. 3 stops being so informative at later times since many simulations terminate before the final hours; e.g., 15 of these 100 simulations terminated in under 10 hours.

[18] After the end of hour 7 no more information calls are made in the simulation, so calls are made only by firms with vacant positions.

TABLE 3

TELEPHONE MARKET WITH 7 HOURS ENFORCED TERMINATION TIME (100 Simulations)

	PSYCHOLOGY MODEL	20 STUDENTS MAY HOLD TWO OFFERS FOR 2 HOURS	EVERY STUDENT MAY HOLD TWO ADJACENT OFFERS		EVERY FIRM FIRST ISSUES OFFERS TO STUDENTS WHO LIKE IT BEST
			Until 1 Hour before the Deadline	Until the Deadline	
Mean time to termination	7:43	7:53	8:01	8:08	7:36
	(:22)	(:10)	(:10)	(:07)	(:37)
Mean time by which 90% of students have received at least one offer	1:02	2:11	2:22	2:33	:57
Mean time by which 99% of students have received at least one offer	5:07	7:06	7:37	7:51	5:23
Mean number of blocking firms	1.58	3.25	6.32	12.77	2.34
	(.74)	(1.26)	(1.61)	(2.27)	(1.02)
Mean number of blocking students	16.67	29.88	48.74	77.76	15.74
	(7.73)	(9.80)	(11.26)	(11.57)	(8.06)
Mean number of unmatched students	.88	1.09	1.52	1.69	.78
Mean number of unmatched firms	.87	1.07	1.41	1.52	.78

NOTE.—Standard deviations are in parentheses.

matching of students to programs need not be stable. Typically there will be blocking pairs, that is, a student and an internship program that are not matched to one another but each ranked one another higher than the outcome they received at the end of the market.

Column 1 of table 3 shows that at the end of these markets (i.e., after all calls following the expiration of the deadline have been completed), on average, almost two firms and 17 workers can take part in such blocking pairs, that is, about 4 percent of the firms and 8 percent of the workers. Now, at the moment that the deadline expires, any firm that can be part of a blocking pair must have at least one position unfilled.[19] So these firms immediately start calling the workers who have not yet rejected them. Since 8 percent of the workers can take part in blocking pairs, it is not at all a rare event that a firm finds itself talking to a worker with whom it could form a blocking pair. The vast majority of such workers have already accepted an offer from another firm (since on average fewer than 1 percent of the workers received no offer by the deadline, so more than 99 percent have accepted offers). This is the origin of the incentives to break the rules, in each of the ways so carefully enumerated in parts *a*–*e* of rule 7.

The incentives to break the rules live on beyond the end of the telephone calls that follow the expiration of the deadline. Firms and workers that were unmatched at the deadline, and especially those that remain unmatched when all remaining transactions have been exhausted, have been badly hurt by the fact that there is a deadline (i.e., in comparison to the outcome they could have expected if the market were conducted as in the medical market). Even if these dissatisfied individuals and firms are fully bound by their verbal commitments, firms have an incentive (the following year) to try to avoid the risk of being caught short at the deadline by breaking the rules against early offers or against pressuring students into revealing their preferences so that offers can be concentrated on students who will accept promptly.

Students, who also face the risk of being caught short at the deadline, may be willing to go along with the attempt to arrange early matches or to signal their preferences in order to attract prompt offers. Indeed, column 1 of table 4, which shows the hourly progress of offers in this market, shows that any student who does not get an

[19] Because any firm *F* that has no positions vacant when the deadline expires has already been rejected by each worker it prefers to its current assignments, and each worker who made such a rejection must be holding (or have already accepted) an offer it prefers to firm *F*.

TABLE 4

HOURLY PROGRESS OF THE PSYCHOLOGY MODEL TELEPHONE MARKET (Means of 100 Simulations)

Hour	Number of Students with at Least One Offer	Number of Students with an Offer from the Firm They Will Ultimately Match With	Number of Offers Made	Number of Offers Not Rejected Immediately
0	.00	.00	.00	.00
1	178.47	104.13	400.08	278.06
2	191.24	140.52	531.96	333.90
3	194.83	161.12	602.36	360.04
4	196.50	174.59	648.58	375.70
5	197.41	184.64	681.79	386.80
6	198.02	192.46	707.38	395.01
7	198.37	198.37	727.89	400.99
8	199.11	199.11	786.35	401.73
9	199.12	199.12	786.79	401.74

offer in the first hour has a substantial (9 percent) risk of getting no offer by the deadline. So there is reason for students to worry about what will happen to them if they do not get an offer in the early, parallel processing phase of the market.

Note also, by comparing the hourly transactions in this market with those in the market with no deadline (tables 2 and 4), that there are, on average, almost 100 fewer offers made when the market has a 7-hour deadline. That is, although there is only a small difference between the two markets in terms of how many positions are filled, there is a substantial difference in terms of how many potential transactions are evaluated to determine which individuals will fill which positions. Thus the instability of the final outcome and the incentives for firms to identify workers who will accept immediately arise from the fact that this market closes before it clears.

We next briefly consider the effects when some firms and workers act on these incentives.

3. The Simulated Psychology Market in Which Employers Seek Out Those Who Will Accept Their Offers Immediately

These simulations follow exactly the rules of the previous psychology market simulations, except that instead of making offers strictly in

order of preference, each firm makes its first offers to any students among its 10 most preferred students for whom that firm is the first choice. (This can be interpreted as the case in which *every* student identifies himself to his first-choice firm and each firm acts on this information only if the students who identify themselves are among its top choices.)[20]

The aggregate results of these simulations are presented in column 5 of table 3. The differences between these simulations and those of the psychology market with straightforward behavior are quite modest, both because only a minority of firms change their behavior (and even then only in regard to their initial offers) and because this change of behavior slows down some transactions at the same time that it speeds others up. That is, even though certain students who have indicated that they would immediately accept an offer receive one sooner than if firms acted straightforwardly on their preferences, other students receive offers later than they would have, and not merely because the firms have changed the order of their offers. The parallel processing phase of the market more quickly gives way to the serial phase, since some firms have one of their positions accepted faster and thus cannot do as much parallel processing of offers. Thus in these simulations the mean time at which 90 percent of students have received offers comes 5 minutes sooner than when behavior is straightforward, but it takes 16 minutes longer before 99 percent of the students have received offers.

This kind of behavior yields more firms that can take part in blocking pairs with respect to the final outcome. This reflects that a firm that makes an offer to, say, its ninth-ranked candidate because that candidate happens to rank it first may be missing a chance to hire its fourth- or fifth-ranked candidate. But as we have seen, firms may be prepared to pay this cost to avoid the risk of being caught with a vacant position when the market closes.

We shall return to this when we consider the differential welfare effects of these different market rules and behaviors. But first we consider the effect on the market when some students may delay before rejecting some offers, that is, when they may sometimes hold more than one offer.

[20] In these simulations each student who applies to a firm has a .05 probability of ranking that firm first, so the probability that none of a firm's top 10 candidates will rank it first is .6. In the (unlikely) event that more than four of a firm's top 10 candidates rank it first, the simulation has the firm make offers to the four most preferred of these candidates. Since firms have four positions, each firm would ordinarily make its initial offers to its first four choices, so the only change of behavior will occur when a firm makes one of its initial offers to the candidates it ranks 5–10.

4. Simulated Psychology Markets in Which Some Students May Sometimes Hold Two Offers

There have been persistent complaints that students may sometimes hold more than one offer, that is, that they may delay in rejecting offers (see, e.g., Belar and Orgel 1980). Rule 6 is now specifically meant to prevent this. Here we investigate the consequences of such delays.

Whereas the previous subsection investigated behavior that agents might exhibit in response to the incentives they face in the market, the behavior we study here is maladaptive. In the deferred acceptance procedure, students can only do better as they receive more offers. So a student who slows the market by holding multiple offers can only reduce the number of offers he will get by the time the market closes. This does not mean, of course, that there are not reasons that students might hold multiple offers, having to do either with mistakes or with (hard to model but real) costs of decision making.[21] Delayed decision making might take different forms, depending on its causes. We therefore model delays in two ways.

In the first set of simulations, 10 percent of the students are randomly selected as being potential delayers, and the first time any one of these students receives an offer when she is already holding one (i.e., when she receives her second offer), she delays responding for 2 hours (or until the deadline, whichever comes first) and thereafter reverts to straightforward behavior without any further delays. The delays in these simulations can be thought of as arising from simple mistakes or carelessness.[22]

In the second and third sets of simulations, every student has the potential to have a delay, which occurs the first time she finds herself holding two *adjacent* offers neither of which is her first choice. The idea is that if she is holding, for example, her third- and fourth-choice offers, then a delay offers the possibility that she will get an offer from her (clear) first choice and not have to decide.[23] In the

[21] Consider a married student whose first choice is clear, but for whom the differences between the second and third choices are less clear, with the third choice being near the spouse's family. To make a decision between the second and third choices might involve a family fight, which can be avoided by delaying in the hope that an offer from the first choice will arrive. Note that there would be additional reasons for delay in a market in which the terms of employment can be negotiated. Then multiple offers could be held to improve the negotiating position of a worker in relation to each of the firms making offers.

[22] Although 20 students are potential delayers, the mean number of actual delayers in the simulations reported below is 16.88, since some potential delayers never receive a second offer.

[23] If a student holding, e.g., her third and fourth choices receives an offer from her second choice, the simulation has her continuing to hold two offers, but now her second and third. This happens, on average, only 1.12 times per simulation. A

second set of simulations we suppose that (perhaps out of fear that rule 6*b* may be activated) students end their delay an hour before the deadline, after which they reject the less preferred of the offers and continue in a straightforward manner with no further delays. In the third set of simulations we suppose that the delay continues all the way to the deadline at the end of hour 7, at which point the student accepts the more preferred offer and rejects the other, and the simulation of the aftermarket proceeds straightforwardly.[24]

Table 3 shows the results of these simulations and allows comparison with the psychology market simulations with no delays. In all these markets, the 7-hour deadline means that all transactions remaining after the deadline are quickly concluded, so there is not too much difference in mean times to termination. Similarly, the fact that the initial part of the market can process offers in parallel means that the time by which 90 percent of the students have received offers is delayed by much less than the length of each individual's delay (e.g., in the market with 2-hour delays, the 90 percent mark comes only 69 minutes after the market with no delays). The story is different with the 99 percent mark, which in each of the markets with delays now comes after the deadline for the deferred acceptance part of the market, so that many more students are faced with exploding offers in the aftermarket than in the market with no delays. (In 100 simulations of the market in which students may hold adjacent offers until the deadline, not a single offer was made between hours 4 and 7, and very few offers were made after hour 2.)[25] And there is an enormous difference in the stability of the final outcome, compared to the market with no delays. In the market with the longest delays, a quarter of the firms and 40 percent of the workers can be involved in blocking pairs with respect to the final outcome. So in the markets with delays, the incentives for breaking the rules have risen enormously.

---

student holding two offers immediately releases them both if she gets a nonadjacent preferred offer or if she gets an offer from her first choice.

[24] These simulations are different from the first set of simulations with delays in two ways. First, more students may potentially be involved in delays. Second, the delays may be longer. The second factor has a greater effect than the first because although all students may potentially be involved in the second kind of delay, they never in fact participate in a delay unless they get an offer adjacent in their preferences to the one they are holding. The mean number of students to whom this happened in the simulations was 23.51.

[25] Recall that, on average, only 24 students were involved in such delays. But there is sufficiently little parallel processing going on in the later hours of the market that this completely shuts it down. To take the phase change metaphor further, these markets start out liquid (in the parallel processing phase) and then ''freeze'' as they become serial. When workers hold multiple offers, the market may freeze solid.

5. Welfare Comparisons among Different Rules and Behaviors

Table 5 allows us to compare the welfare effects of the different regimes we have considered. For the simulations of each model (of the rules of the market and the behavior of the participants) it shows how many students are matched with their first-choice firm, their second choice, and so forth. Dividing these numbers by 200 yields the probability that a random student would have ended up in each position on his preference list, including the possibility of being unmatched.

We shall defer for a moment the case in which firms do not behave straightforwardly, and concentrate first on the comparisons among the other five models. Here the comparisons are striking and unambiguous, since we can order these five models (as in the table) so that the distribution of outcomes for each one stochastically dominates the distribution for the next. That is, the students do best under the medical model (with no termination time), next best under the psychology model with straightforward behavior, and increasingly worse as the length of the delay during which two offers may be held increases. For example, a student has a higher probability of getting each of his first four choices when there is no termination time (the medical model) than when there is (the psychology model) and a lower probability of getting each of his choices 5–20 or of being unmatched.

The reason is that, in the deferred acceptance procedure, a worker can only be helped by getting an additional offer, so workers' success in the market is monotonic in the number of offers they receive. Introducing a deadline or increasing the length of delays acts to reduce the expected number of offers a worker will receive. That is, these simulations show that the conclusions of theorem 3 apply. As we have already noted in table 3, the events on which the conclusions of theorem 3 are conditioned (that any particular student has received at least one offer by the deadline) are highly probable, and the inequalities in the conclusion of theorem 3 are likely to be strict since the effective lengths of these markets, as measured by the number of offers for which there is time, are strictly different.

As in theorem 4, it is more complicated to evaluate the welfare of the firms, primarily because, in contrast to the workers, the decreasing number of offers made and the increasing chance of having a position unfilled work in *opposite* directions for the firms. We can separate out the two effects by considering the expected success of a firm in each of the five models, conditional on the firm's filling all its positions. In this case, the welfare of the firms is ordered in

TABLE 5

NUMBERS OF STUDENTS AND THE CHOICES THEY MATCH WITH (Means of 100 Simulations): NUMBER OF STUDENTS WHO MATCH WITH CHOICE i

CHOICE i THAT STUDENTS MATCH WITH	MEDICAL MODEL	PSYCHOLOGY MODEL	MODIFIED PSYCHOLOGY MODEL: 20 STUDENTS MAY HOLD TWO OFFERS FOR 2 HOURS	PSYCHOLOGY MODEL		
				Students May Hold Two Adjacent Offers until 1 Hour before the Deadline	Every Student May Hold Two Adjacent Offers until the Deadline	Firms First Issue Offers to Students Who Like Them Best
1	50.50	41.02	36.24	30.61	23.98	42.57
2	36.78	31.57	28.74	24.84	20.01	32.05
3	27.62	25.54	24.12	22.06	19.07	25.33
4	20.97	20.74	20.07	18.49	16.55	20.84
5	16.17	16.88	16.50	16.24	15.10	16.38
6	11.75	13.21	13.58	13.59	13.19	12.67
7	9.03	10.57	11.14	11.78	11.43	10.17
8	6.53	8.22	8.88	9.60	10.50	8.20
9	5.20	6.62	7.33	8.39	9.22	6.79
10	3.93	5.40	6.29	7.14	8.42	4.85
11	2.91	4.30	5.20	6.29	7.69	4.60
12	2.30	3.63	4.50	5.95	7.31	3.55
13	1.71	2.70	3.63	4.58	5.92	2.70
14	1.27	2.30	2.94	3.95	5.34	2.11
15	.86	1.53	2.34	3.26	5.41	1.72
16	.80	1.44	2.13	3.12	4.63	1.42
17	.50	1.04	1.53	2.69	4.23	.95
18	.44	.85	1.31	2.14	3.63	.98
19	.37	.76	1.23	2.07	3.93	.83
20	.35	.80	1.21	1.69	2.75	.51
Unmatched	.01	.88	1.09	1.52	1.69	.78

exactly the opposite order of the welfare of the workers. (For example, to choose a simple measure, the probability that a firm will fill its four positions with its first four choices is lowest in the medical model, higher in the psychology model with straightforward behavior, and higher still in the models with increasing delays.)[26] This is so even though the medical model produces the firm-optimal stable outcome; the other markets produce *un*stable outcomes that are better for those firms that have all their positions filled.

Of course, the very worst thing that can happen to a firm is that one or more of its offers should be rejected just before the deadline, so it does not have time to get new offers out before all workers accept the offers they are holding. When the market rules are followed, this means that such a firm must either remain unmatched or find a match with a worker who did not receive any offers before the deadline. If this is a sufficiently undesirable outcome (as it seems to be to many market participants we have spoken to), the increasing risk of remaining unmatched may even cause firms' preferences to coincide with workers' preferences over these five market regimes.[27] Another factor working in this direction is that increased instability of the final market outcome presumably causes more violations of the rules, and insofar as the firms have a long-term interest in the orderly operation of the market, they may prefer those regimes that offer rule-breaking incentives to the fewest potential blocking pairs.

As already noted, comparing these five models to the remaining model, in which the firms do not act straightforwardly on their preferences, is more nuanced, because the firms' actions cut different ways in the parallel and serial phases of the market. Table 5 shows that the distribution of students over the choices they match to in this model neither stochastically dominates nor is dominated by the distribution from the psychology model with straightforward behavior. The distribution is stochastically dominated by the distribution of the medical model (without a termination time), and it dominates the distributions in the three models with delays. This simply confirms the large effects that introducing a deadline and experiencing delays have on reducing the number of offers in the market (and therefore reducing the welfare of the workers).

[26] More comprehensive measures are necessarily a little complex since we have made no assumptions that permit us to compare the welfare of a firm when it fills its positions with, say, choices 2, 3, 4, and 5 and when it fills them with choices 1, 2, 3, and 7.

[27] Exactly which programs are unmatched is random, and even highly regarded programs are not immune (e.g., a program that is the second choice of every student can still be rejected right at the deadline by a student who has just gotten an offer from his first choice).

C. *Sensitivity to the Joint Distribution of Preferences and to the Concentration of Positions*

1. Correlated Preferences

So far we have reported simulations in which all preferences are uncorrelated. Now we consider the robustness of the results when preferences are correlated, reflecting some agreement on which are the most desirable firms and workers. The chief result is that increasing the correlation among firms' preferences makes it even more difficult for the market to clear because the initial phase of the market becomes more congested, with many firms lining up to make offers to the most preferred workers. The simplest way to see this is to begin with the extreme cases in which preferences are either identical or completely independent. Varying this separately for the workers (students) and for the firms yields four cases, shown in table 6. (In order to allow preferences on each side of the market to be identical, each student has preferences over all firms, and vice versa, so table 6 is comparable to panel B of table 1.)

Case 1 in table 6 shows the results for both the medical and psychology markets in the case we have already considered, in which all preferences are uncorrelated. (The medical market numbers in case 1 reproduce those in col. 1 of table 1 [panel B].) Comparing the medical markets in case 1 and case 2 of table 6 shows that the change from uncorrelated to common preferences among the firms dramatically slows the critical early hours of the market (although the mean time to termination is only modestly longer in case 2 than in case 1). The mean time by which 90 percent of the students have received at least one offer goes from just over 1 hour in the medical market when firms have uncorrelated preferences to just over 22 hours when they have identical preferences. Because firms' preferences are perfectly correlated in case 2, they all attempt to make their first offers to the same students, and in the resulting congestion an average of only 11 students (5.5 percent) receive at least one offer in the first hour.

To put it another way, by the time the 7 hours available in the psychology market have expired, 99 percent of the students have received at least one offer when the firms have uncorrelated preferences (in case 1), but only 31 percent have when the firms' preferences are perfectly correlated (case 2). So in the psychology market, the exploding offer aftermarket makes most of the matches in case 2, in contrast to case 1, as can be seen by comparing the mean numbers of firms and students that can participate in blocking pairs to destabilize the final outcome in the two cases. When the firms' preferences are uncorrelated, on average, only two firms and 31 students

TABLE 6

MEDICAL AND PSYCHOLOGY MARKET SIMULATIONS: VARYING THE CORRELATION OF PREFERENCES
(Students Have Preferences over All 50 Firms; 100 Simulations)

	PREFERENCES							
	Case 1: Students Have Uncorrelated Random Preferences; Firms Have Uncorrelated Random Preferences		Case 2: Students Have Uncorrelated Random Preferences; Firms Have Identical Preferences		Case 3: Students Have Identical Preferences; Firms Have Identical Preferences		Case 4: Students Have Identical Preferences; Firms Have Uncorrelated Random Preferences	
	Medical Market	Psychology Market	Medical Market	Psychology Market	Medical Market	Psychology Market	Medical Market	Psychology Market
Mean* time to termination	22:53 (12:03)	8:39 (:43)	25:09 (:45)	18:10 (:14)	20:46 (:18)	17:12 (:17)	13:16 (2:18)	8:29 (:32)
Mean time by which 90% of students have received an offer	1:09	1:09	22:06	16:10	18:51	15:12	1:18	1:18
Mean time by which 99% of students have received an offer	7:02	6:21	24:50	17:57	20:36	16:58	7:53	6:52
Mean number of blocking firms	0	2.23 (.85)	0	47.75	0	37.1 (2.05)	0	.68 (.68)
Mean number of blocking students	0	31 (12.83)	0	151.31 (3.71)	0	156.13 (7.48)	0	1.72 (2.23)

NOTE.—Standard deviations are in parentheses.
* The corresponding medians are very close to the means.

can participate in blocking pairs. But when firms' preferences are perfectly correlated (and most matching is done by the aftermarket), an average of 48 firms and 151 students can participate in destabilizing blocking pairs.

Comparing cases 4 and 3 shows that going from uncorrelated to perfectly correlated firms' preferences has the same effect whether students' preferences are correlated or uncorrelated.

Correlation of students' preferences, in contrast, speeds up the serial part of the market. This is clearest comparing cases 1 and 4 in table 6, that is, comparing the effect of correlation of students' preferences when firms' preferences are uncorrelated. In both cases 1 and 4, the first hour of the market has a big parallel processing component. Both when students' preferences are uncorrelated and when they are perfectly correlated, it takes a little over an hour for 90 percent of the students to receive their first offer, and in both cases the mean time by which 99 percent of the students have received at least one offer is under 8 hours. But when students' preferences are perfectly correlated, much more sorting of firms takes place in the early hours of the market. (For example, in case 1 at the end of hour 7 the mean number of students who have received at least one offer is 197.59, and in case 4 the number is 196.87, virtually identical. But in case 1, in which students' preferences are uncorrelated, the mean number of students who by this time have received an offer from the firm to which they will ultimately be matched in the medical market is only 155.4, whereas in case 4 this number is 196.15.) So the final, serial part of the market moves much faster when students' preferences are correlated because fewer offers need to be made before a stable matching is achieved.[28] This is reflected in the much longer mean (and median) times to termination in the case 1 than case 4 medical markets and in the larger number of potential blocking pairs in the case 1 than case 4 psychology markets.[29]

[28] The mean number of offers required to reach a stable outcome when all preferences are uncorrelated (case 1) is 1,015, but when students' preferences are perfectly correlated (case 4), it is 800.

[29] We also studied partially correlated preferences, as follows. For a firm's preferences, assign to each student j the number $j + R_j$, where R_j is a random number drawn from a normal distribution with mean zero and variance M. The firm prefers a student with a lower number $j + R_j$ to one with a higher number $k + R_k$. When the variance $M = 0$, this gives us the identical, perfectly correlated preferences considered in table 6. As M goes to infinity, the original numbers j diminish in importance, and in the limit we get the case of uncorrelated random preferences. The results of these simulations show, as we might expect, that as the correlation of the preferences increases from zero to one, the behavior of the markets moves continuously from one extreme to the other in table 6. The most marked effects occur with the highest correlations of firms' preferences.

Note that the case in which both students and firms have perfectly correlated preferences is a good example in which the assumptions and conclusions of theorem 3 do not hold. Because the firms all have common preferences, we can identify, for example, the least desirable student, who is matched at the unique stable matching (and hence at the outcome of the medical market) to the least desirable firm (which is also well defined since students have identical preferences). Because of the congestion caused by the correlation of the firms' preferences in this market, this least desirable student is virtually certain to be unmatched at the end of hour 7, which means that in the psychology market he is matched during the exploding offer aftermarket. This introduces a large random component, so this student's distribution of possible matches is better in the short (psychology) market than in the longer medical market.

In summary, the more highly correlated with one another the firms' preferences are, the more congestion there is at the beginning of the market, so the longer it takes for workers to receive offers. When the deferred acceptance part of the market has a 7-hour termination time, this means that the number of instabilities increases as firms' preferences become more highly correlated. In contrast, more highly correlated workers' preferences speed the sorting of firms, shorten the length of the market, and reduce the number of instabilities. But there are substantial bottlenecks even in the best case (when firms' preferences are uncorrelated and workers' preferences are identical). And when both firms and workers have highly correlated preferences, the congestion due to the firms' preferences predominates because it slows down the rate at which workers receive offers (and because a worker's preferences do not begin to have any effect until she receives at least two offers).

2. Concentration of Positions

So far we have reported simulations in which the 200 positions are always offered by 50 firms. To see the effect of concentration, we consider a range of markets, starting with much more concentrated markets (two firms each with 100 positions) and ending with less concentrated markets (200 firms each with one position). Of course, if we continued to model each firm as having only one telephone for outgoing calls, we would introduce an arbitrary kind of serial processing in the most concentrated markets. To make these comparisons informative about the transition between parallel and serial processing, we therefore relax any constraint on the number of simultaneous offers a given firm can make. (This would be natural, for example, in a market in which offers were made by mail.) This

is equivalent to modeling each firm as having ''infinitely many'' phones, that is, a large enough staff of telephonists to offer all its positions simultaneously. Having no constraint on the number of simultaneous offers a firm can make speeds up the parallel processing part of the market (so the times for the next simulations are not comparable to those reported above). But having many phones does not speed the market at all when most firms must wait for an offer to be rejected before they can issue a new one.

Table 7 reports the results as we vary the number n of firms: $n = 2, 5, 10, 20, 50, 100$, and 200. In each market, all firms and students have uncorrelated random preferences. Each firm has infinitely many phones, whereas each student has only one phone. The number of students and the number of total positions are constant at 200; each firm in the market has $200/n$ positions to fill.

The simulations show that the mean time to termination increases with the number of firms, as do the times by which 90 percent and 99 percent of the students are matched. For each concentration of positions, the substantial differences between each of these three times show that there is a clear transition between the early, parallel processing part of the market in which the bulk of the students receive at least one offer and the later, much more extended serial processing part of the market.

III. Concluding Remarks

Even the extremely short turnaround times characteristic of the entry-level market for clinical psychologists can cause bottlenecks, which impede market clearing and promote strategic behavior. This is robust to changes in the correlation of preferences and in the concentration of positions, and seems likely to have implications not only for other labor markets but for markets generally.

The turnaround time makes itself felt when the market enters a phase in which most potential transactions must be processed serially. It is useful to distinguish between congestion-based serial processing and serial processing that arises because of the nature of the transactions.

It appears that any market can be forced by congestion into serial processing. Even in stock exchanges, which have perhaps the highest degree of parallel processing of proposed transactions (in the sense that every bid or asked price can be offered to the whole market), great congestion may cause serial processing (e.g., in the specialist's book), which may contribute to market ''meltdowns.'' But if the market and specialist firms hired more staff, more transactions could be processed simultaneously.

TABLE 7

MEDICAL MODEL TELEPHONE MARKET WITH VARYING NUMBER OF FIRMS AND INFINITELY MANY PHONES: 200 STUDENTS AND 200 POSITIONS TO BE FILLED (100 Simulations)

	NUMBER OF FIRMS						
	2	5	10	20	50	100	200
Mean time to termination at a stable outcome	1:20	3:40	5:27	7:17	8:47	10:31	11:13
	(:24)	(1:22)	(2:34)	(3:27)	(3:19)	(5:41)	(4:49)
Median time to termination	1:20	3:33	4:46	6:23	8:18	9:14	10:07
Mean time by which 90% of students have received an offer	:10	:21	:27	:31	:34	:35	:35
Mean time by which 99% of students have received an offer	:40	1:32	1:59	2:32	3:10	3:22	3:58
Longest time to termination	3:05	7:45	15:08	24:14	18:39	30:51	26:52
Shortest time to termination	:40	1:26	2:32	2:04	3:52	2:53	3:29

NOTE.—Standard deviations are in parentheses.

The serial processing we have been considering arises in a more fundamental way. It arises even in markets in which firms have no communications constraints and can offer all their available positions simultaneously. The bottlenecks arise not from a lack of processing capacity, but from the fact that, after most workers have received at least one offer, most firms must wait for an offer to be rejected before they can issue a new offer, and this takes time. It seems likely that any market in which agents propose transactions by making offers that must be left open for at least a short specified period has the potential to experience this phase transition.[30]

The turnaround time itself does not cause markets to end prematurely, but must be considered along with the duration of the market to determine the market's *effective length,* which (for a labor market) we can roughly define as the average number of sequential offers a firm can expect to have time to make for a given position.[31] In markets in which salaries (and other dimensions of the job) are negotiated, there is more reason for workers to hold multiple offers (which may be used for leverage in negotiation), and the resulting delays in rejecting offers will decrease the effective length of the market.

It is not simple to extend the effective length of a market. For example, in previous years the psychology market had a longer duration. But the short turnaround time in the current market is related to its 7-hour duration, which allows everyone to plan to spend the day next to the telephone. In a market conducted over 5 days, for example, people could not wait by the phone all the time, so turnaround time would likely go up by more than a factor of five (i.e., to more than ½ hour from the time someone decides to reject an offer until another candidate has received it). Thus this is a market in which extending the duration of the market might *shorten* its effective length.

Finally, we have discussed how there are particular incentives for strategic behavior in markets whose effective length is insufficient to guarantee market clearing. In the clinical psychology market, these incentives lead to an emphasis on identifying candidates who will quickly accept an offer. In less centralized markets, firms with many different positions (e.g., universities that recruit in many departments) may make simultaneous multiple offers for each position. Similarly, firms have incentives to try to ''capture'' their top choices (and also to increase the effective length of their market) by making offers of very short duration, which shorten the effective length of

[30] Consider, e.g., U.S. markets for residential housing.

[31] Of course different agents in a market may face different effective lengths, depending on their own turnaround time.

the market faced by other firms by removing candidates from the market quickly. Thus a short effective length of a market can give agents incentives to behave in ways that further shorten the market's effective length.[32]

This in turn makes it increasingly difficult for the market to clear fully and arrive at a stable outcome. And as was observed in Roth (1984, 1990, 1991), Mongell and Roth (1991), and Roth and Xing (1994), unstable outcomes can give agents the incentive to ''jump the gun'' and make very early offers, or try to disrupt agreements with late offers.[33] Thus there is a relation between the turnaround times studied in this paper and the timing of early and late offers studied in those earlier papers. And even in markets that do not experience overt timing problems, firms have incentives to ''target'' the candidates to whom they will make their first offers, taking into account the probability that an offer will be accepted (i.e., not necessarily attempting to hire first the most preferred candidate). These strategic considerations make efficient outcomes more difficult to achieve.[34]

Methodologically, this paper is part of a body of work that seeks to understand markets on the basis of a detailed understanding of their rules and how they have evolved over time. (If we wish game theory to become as integral a part of applied economics as it is of economic theory, it cannot be said too often that rules are data and have to be collected and analyzed.) When we study naturally occurring games instead of simple stylized models, it is to be expected that the size of the strategy space may preclude analysis by currently available theoretical tools. It is for this reason that we have relied here on computation and have concentrated on straightforward behavior. The set of feasible strategies in the market studied here is too large to even sample sensibly without some further understanding of the strategy space being sampled. Some thoughts on sampling strategy spaces and exploring them computationally are found in Erev and Roth (1996).

Substantively, this paper is part of a body of work that seeks to

[32] For example, in recent years there has been a growth of ''early admission'' programs for college admission, in which colleges admit students in December (instead of in April) in return for a commitment to withdraw applications to other colleges and attend (see, e.g., Arenson 1996).

[33] Roth and Xing (1994) note that there are a number of other reasons for offers to be made earlier and earlier, and one of them has recently been explored at some length by Li and Rosen (1996).

[34] Thus, e.g., I am not sanguine about the prospects for the latest market reorganization proposed in the market for federal court clerks (see Becker, Breyer, and Calabresi 1994). The rules proposed there seem likely to promote behavior that will lead to a very short effective length of the market.

take seriously the role of time in markets. Considerations of timing play an important role in shaping the strategic environment facing market participants, and this in turn can have a profound influence on their behavior and on the performance of the market.

Appendix

APPIC Policy: Internship Offers and Acceptances (Revised 5/91)

Adherence to these policies is a condition of membership in APPIC. (Rules 1 and 11–14 are omitted.)

2. Internship program directors must inform applicants who are excluded from consideration as early as possible in the process, and no later than one week before selection day.
   a. Students who remain under consideration may be notified that they remain under consideration after others have been excluded.
   b. No other information (such as agency's ranking of the applicant; status as alternate/first choice, etc.) may be communicated to applicants prior to selection day.
3. No internship offers in any form may be extended by agencies before the beginning of selection day.
   a. The *only* information that agencies may communicate to applicants prior to this time is whether or not the applicant remains under consideration for admission (see item 2). The spirit of this item precludes any communication of an applicant's status prior to the time above, however ''veiled'' or indirect such communication might be.
   b. ''Alternates'' may be fully informed of their status any time after the start of selection day. Applicants may not be told whether they are considered alternates or first choices prior to that time.
   c. Internship programs may not solicit information regarding an applicant's ranking of programs or his/her intention to accept or decline an offer of admission until after that offer is officially tendered.
4. Applicants must reply to all offers no later than the closing time on selection day.
   a. This deadline applies to all offers including those to applicants who are initially considered ''alternates'' and are subsequently extended an offer any time prior to the end of selection day.
   b. Agencies may inquire as to the applicant's progress towards making a decision at any time after an offer is formally extended. Under no circumstances, however, may an agency implicitly or explicitly threaten to rescind an offer if a decision is not made prior to the end of selection day (except as noted in item 6).
   c. It is in everyone's best interest that applicants make and communicate decisions to accept or reject each offer as quickly as possible.
   d. Any offer that has not been accepted is void as of the ending hour of selection day.

5. An applicant must respond immediately to each offer tendered in one of three ways. The offer may be accepted, rejected or ''held.''
   a. *Accepting* the offer constitutes a binding agreement between applicant and internship program.
   b. *Refusing* the offer terminates all obligations on either side and frees the internship program to offer the position to another applicant.
   c. *Holding* the offer means that the offer remains valid until the applicant notifies the program of rejection or acceptance, or until the end of selection day.
6. Applicants may ''HOLD'' no more than one active offer at a time.
   a. If an applicant is holding an offer from one program and receives an offer from a more preferred program, s/he may accept or ''hold'' the second offer provided that the less preferred program is notified *immediately* that the applicant is rejecting the previously held offer.
   b. If a program confirms that an applicant is holding more than one offer, the program is free to withdraw their previously tendered offer of acceptance, and to offer that position to another applicant *after* the offending applicant is notified of that decision.
7. An offer of acceptance to an applicant is valid only if the applicant has not already accepted an offer of admission to another program.
   a. An applicant's verbal acceptance of an offer constitutes a binding agreement between the applicant and the program that may not be reversed unilaterally by either party.
   b. Before programs extend an offer, they must first explicitly inquire whether the applicant has already accepted an offer elsewhere. If so, no offer may be tendered.
   c. A program may in no way suggest that an applicant renege on previously accepted offers.
   d. If an applicant who has accepted an offer receives a second offer, s/he is obligated to refuse the second offer and inform the agency that s/he is already committed elsewhere.
   e. Any offer accepted subsequently to a prior commitment is automatically null and void, even if the offering agency is unaware of the prior acceptance and commitment.
8. When an applicant accepts an offer of admission, s/he is urged to immediately inform all other internship programs at which s/he is still under consideration that s/he is no longer available.
9. Applicants who have not accepted a position prior to the end of selection day may receive offers of admission after that deadline.
   a. Applicants should be prepared to accept or reject such late offers quickly, since most other deliberations should have already taken place.
   b. Programs may legitimately place short but reasonable deadlines for responses to such late offers.
10. Once a program has filled all available positions, all candidates remaining in their applicant pool must be notified that they are no longer under consideration.

a. Applicants who have not notified the agency that they have accepted a position elsewhere and who have not been selected by the agency should be notified by phone as soon as all positions are filled.

b. If an applicant cannot be reached by phone, s/he should be so notified by letter postmarked no later than 72 hours after the end of selection day.

References

Arenson, Karen W. ''Top Colleges Fill More Slots with Those Applying Early.'' *New York Times* (February 14, 1996), p. B8.

Becker, Edward R.; Breyer, Stephen G.; and Calabresi, Guido. ''The Federal Judicial Law Clerk Hiring Problem and the Modest March 1 Solution.'' *Yale Law J.* 104 (October 1994): 207–25.

Belar, Cynthia D., and Orgel, Sidney A. ''Survival Guide for Intern Applicants.'' *Professional Psychology* 11 (August 1980): 672–75.

Bergstrom, Theodore C., and Bagnoli, Mark. ''Courtship as a Waiting Game.'' *J.P.E.* 101 (February 1993): 185–202.

Blum, Yosef; Roth, Alvin E.; and Rothblum, Uriel G. ''Vacancy Chains and Equilibration in Senior-Level Labor Markets.'' *J. Econ. Theory* (in press).

Collins, Susan M., and Krishna, Kala. ''The Harvard Housing Lottery: Rationality and Reform.'' Working paper. Washington: Brookings Inst., 1993.

Crawford, Vincent P. ''Comparative Statics in Matching Markets.'' *J. Econ. Theory* 54 (August 1991): 389–400.

Erev, Ido, and Roth, Alvin E. ''Modeling How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria.'' Manuscript. Pittsburgh: Univ. Pittsburgh, 1996.

Gale, David, and Shapley, Lloyd. ''College Admissions and the Stability of Marriage.'' *American Math. Monthly* 68 (January 1962): 9–15.

Li, Hao, and Rosen, Sherwin. ''Unraveling in Assignment Markets.'' Manuscript. Chicago: Univ. Chicago, 1996.

Mongell, Susan, and Roth, Alvin E. ''Sorority Rush as a Two-Sided Matching Mechanism.'' *A.E.R.* 81 (June 1991): 441–64.

Osborne, Martin J., and Rubinstein, Ariel. *Bargaining and Markets*. San Diego: Academic Press, 1990.

Pollak, Robert A. ''For Better or Worse: The Roles of Power in Models of Distribution within Marriage.'' *A.E.R. Papers and Proc.* 84 (May 1994): 148–52.

Roth, Alvin E. ''The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory.'' *J.P.E.* 92 (December 1984): 991–1016.

———. ''The College Admissions Problem Is Not Equivalent to the Marriage Problem.'' *J. Econ. Theory* 36 (August 1985): 277–88.

———. ''On the Allocation of Residents to Rural Hospitals: A General Property of Two-Sided Matching Markets.'' *Econometrica* 54 (March 1986): 425–27.

———. ''New Physicians: A Natural Experiment in Market Organization.'' *Science* 250 (December 14, 1990): 1524–28.

———. ''A Natural Experiment in the Organization of Entry Level Labor Markets: Regional Markets for New Physicians and Surgeons in the United Kingdom.'' *A.E.R.* 81 (June 1991): 415–40.

———. ''Evaluation of Changes to Be Considered in the NRMP Algorithm.'' Consultant's report. Pittsburgh: Univ. Pittsburgh, 1995. http://www.pitt.edu/~alroth/nrmp.html.

———. ''The National Resident Matching Program as a Labor Market.'' *J. American Medical Assoc.* 275 (April 3, 1996): 1054–56. (*a*)

———. ''Report on the Design and Testing of an Applicant-Proposing Matching Algorithm, and Comparison with the Existing NRMP Algorithm.'' Manuscript. Pittsburgh: Univ. Pittsburgh, November 1996. (*b*) http://www.pitt.edu/~alroth/nrmp.html.

Roth, Alvin E., and Rothblum, Uriel G. ''The Information Requirements of Strategic Behavior in Labor Markets and Other Matching Processes.'' Working paper. Pittsburgh: Univ. Pittsburgh, 1996.

Roth, Alvin E., and Sotomayor, Marilda A. Oliveira. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge: Cambridge Univ. Press, 1990.

Roth, Alvin E., and Vande Vate, John H. ''Random Paths to Stability in Two-Sided Matching.'' *Econometrica* 58 (November 1990): 1475–80.

———. ''Incentives in Two-Sided Matching with Random Stable Mechanisms.'' *Econ. Theory* 1 (January 1991): 31–44.

Roth, Alvin E., and Xing, Xiaolin. ''Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions.'' *A.E.R.* 84 (September 1994): 992–1044.

Rubinstein, Ariel. ''Perfect Equilibrium in a Bargaining Model.'' *Econometrica* 50 (January 1982): 97–109.