Truncation strategies in matching markets--In search of advice for participants
Alvin E Roth; Uriel G Rothblum
Econometrica; Jan 1999; 67, 1; ABI/INFORM Global
pg. 21

Econometrica, Vol. 67, No. 1 (January, 1999), 21–43

TRUNCATION STRATEGIES IN MATCHING MARKETS—IN SEARCH OF ADVICE FOR PARTICIPANTS[1]

BY ALVIN E. ROTH AND URIEL G. ROTHBLUM

We consider the strategic options facing workers in labor markets with centralized market clearing mechanisms such as those in the entry level labor markets of a number of professions. If workers do not have detailed information about the preferences of other workers and firms, the scope of potentially profitable strategic behavior is considerably reduced, although not entirely eliminated. Specifically, we demonstrate that stating preferences that reverse the true preference order of two acceptable firms is not beneficial in a low information environment, but submitting a truncation of the true preferences may be. This gives some insight into the successful operation of these market mechanisms.

KEYWORDS: Stable matching, labor markets.

1. INTRODUCTION

THE STUDY OF TWO-SIDED MATCHING MARKETS is an area in which the theoretical and empirical literature make close contact (see Roth and Sotomayor (1990)). Many such markets—particularly entry level professional labor markets—have developed centralized market clearing mechanisms in response to a variety of market failures (see Roth and Xing (1994)). It has proved possible to analyze these mechanisms, and explain to a large degree why some mechanisms have succeeded and others have failed. Mechanisms that implement stable matchings succeed very much more often than those that do not. (Stable matchings will be defined shortly.) The theory and evidence thus allow us to offer practical advice to market organizers who have reason to contemplate adopting such market clearing mechanisms.[2]

However the existing theoretical results do not generally allow us to address the considerable demand for practical advice about how to *participate* in such markets, once they are established. It is difficult to advise participants in markets that use stable matching mechanisms when to behave straightforwardly (i.e. in a way that reveals their true preferences) and when there might be opportunities to behave strategically, and if so, how. This also suggests that there are some gaps in our understanding of why stable matching mechanisms work so well in practice.

The lack of advice to individual participants is all the more disturbing (to us as game theorists) because the form of some of the existing theoretical results

[1]This work was partially supported by NSF Grant No. SES-9121968 and ONR Grant N00014-92-J1142. We are grateful to Professor Ulrich Kamecke for a helpful early conversation, and to some anonymous referees for their suggestions. This work was conducted while Roth was at the Dept. of Economics of the University of Pittsburgh.

[2]See, e.g., the section "The growing (consulting) business of economic design" at http://www.economics.harvard.edu/faculty/roth/roth.html.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

would appear to suggest that they *do* offer advice to participants. For markets organized to produce stable outcomes in the way most commonly observed, theoretical results demonstrate that, except when there is a unique stable outcome, there are always players who could do better than to straightforwardly reveal their true preferences. Furthermore, at equilibrium (at least in the simplest models), when agents behave strategically (so as to misrepresent their true preferences), the mechanisms continue to produce outcomes that are stable with respect to the true preferences, just as they do when all agents behave straightforwardly. Finally, it is possible to determine precise equilibrium strategies. The problem is that these equilibrium strategies require more information than participants typically have.[3] Furthermore, the multiplicity of equilibria means that the existence results give no clue to the *form* that sensible strategies might take.

The gap between these kinds of (conventional game theoretic) observations and practical advice is large. For example, one of the markets in question is the entry level market for American physicians, the National Resident Matching Program. When some of the results referred to above were brought to the attention of participants in that market, a number of questions were raised. Representative of these was a letter from a dean at a prestigious New England medical school, who wrote (7 January, 1991, letter to Roth):

> "My purpose in writing is to explore your statement that the student's self interest may be served by not ranking hospitals in his true preference order. As one who has advised students in this function for over 15 years, I have regularly told them that that is not the case... Could I ask you for any material you have that would illuminate this point? I don't like to bother you, but in fact a great deal is riding on this...."[4]

This paper takes a step towards providing a formal analysis of this question, beginning with the simple but robust model of one-to-one matching called the "marriage model." We will observe that, in order for participants to identify some kinds of strategies that perform better than straightforward behavior, they require a lot of information about other participants' preferences. It might be conjectured, therefore, that when participants have very little information about others' preferences, they will never be able to identify profitable strategic manipulations. This turns out to be false—we will show that even with very little information about others' preferences, just by consulting his own utility function a participant may sometimes be able to identify a better strategy than to state his true preferences. However we will show that when participants have very

[3] For formal statements of these results, see Roth and Sotomayor (1990, Theorems 4.6, 4.16 (Roth (1984b)), and 4.15 (Gale and Sotomayor (1985))).

[4] This question continues to play a substantial role in contemporary debate within the medical community concerning the organization of the NRMP; see, e.g., the articles and replies by Williams (1995a, 1995b) and Peranson and Randlett (1995a, 1995b), and the document put out by the American Medical Students Association in conjunction with Ralph Nader's organization, Public Citizen, AMSA/Public Citizen (1995). In response to these calls for change, the Board of Directors of the NRMP commissioned the design of a new algorithm and a study of possible changes, outlined in Roth (1995, 1996b). In May of 1997 the NRMP decided to adopt the newly designed algorithm for all matches starting in 1998 (Roth and Peranson (1997, 1998)).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

little information about the preferences of other participants, the class of strategies that it may be profitable to employ is a very simple one. It consists of what we will call *truncation strategies* that, loosely speaking, are strategies in which applicants restrict the number of positions for which they apply, but faithfully transmit their true preferences about those positions for which they do apply. Furthermore, when truncation strategies are used judiciously, the outcome of the match will in fact be stable, and the instabilities which may result from excessive truncation will be of a kind that are difficult to detect.

2. TWO-SIDED MATCHING AND STABILITY: A SIMPLE MODEL

Formally, a *marriage market* is a triplet (F, W, P) where F and W are disjoint finite sets of *firms* and *workers*, respectively, and P is a preference profile, i.e., a collection of preferences,[5] such that for each $f \in F$ P_f is a preference relation over the set $\{f\} \cup W$ and for each $w \in W$ P_w is a preference relation over the set $\{w\} \cup F$. (We will sometimes assume preferences are strict, i.e. that agents are not indifferent between distinct alternatives.) We refer to $V \equiv F \cup W$ as the set of *agents*. We write $v' >_v v''$ when v' is preferred to v'' under the preference relation P_v and in this case we say that v *prefers* v' to v''. The relations $<_v, \geq_v, \leq_v$ are derived in the standard way. We sometimes add superscripts and write, for example, $v' >_v^P v''$ rather than $v' >_v v''$ to indicate particular preference profiles. A preference profile P will sometimes be written $P = (P_{-v}, P_v)$, where P_{-v} denotes the preferences of all agents other than some agent v whose preferences P_v have been singled out.

For $v \in V$ we define the *acceptable set of* v *under* P to be the set of agents with which v would rather be matched than remain unmatched, i.e., $A_v(P) \equiv \{v' \in V: v' >_v v\}$. Of course, if $v \in F$ then $A_v(P) \subseteq W$ and if $v \in W$ then $A_v(P) \subseteq F$. We say that v is *acceptable to* v' if $v \in A_{v'}(P)$. A pair $(f, w) \in F \times W$ is called *acceptable* if f and w are acceptable to each other. The set of all acceptable pairs under P is denoted by $A(P)$. Generally we represent the preferences of the agents by lists for each agent v, with the members of $A_v(P)$ listed in decreasing order of P_v.

A *matching* can be most easily thought of as a subset of $F \times W$ (i.e. a subset of pairs) such that any agent v appears in at most one of the pairs. To formally represent both the matched agents and those who may be unmatched, we represent a matching by a one-to-one correspondence $\mu: V \to V$ where $\mu(f) = w$ and $\mu(w) = f$ if (f, w) is a matched pair, and $\mu(v) = v$ if v is unmatched. Given a matching μ, we call $\mu(v)$ the *outcome for* v *under* μ. For example, if $F = \{f_1, f_2, f_3\}$ and $W = \{w_1, w_2, w_3\}$, then a matching might be $\mu = [(w_1, f_2), (w_2, f_1), (w_3), (f_3)]$, indicating that w_1 and f_2 are matched to one another, w_3 is unmatched, etc. A matching is called *acceptable* if its matched pairs are a subset of $A(P)$.

[5] It will sometimes be useful to think of a preference profile P as a function which maps each player $v \in V$ into a preference ranking. This will be helpful later, when we generalize profiles into random profiles, which can then be viewed as random variables.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Gale and Shapley (1962) originally studied a marriage problem with equal numbers of agents on both sides, and in which no one was ever unmatched. This can be viewed as arising from a restriction on the allowable preferences, so that being unmatched is always less preferred than being matched, i.e. a restriction so that if $v \in F$ then $A_v(P) = W$ and if $v \in W$ then $A_v(P) = F$. (Such a preference restriction would seem to be justified in certain special matching markets, such as might occur within a firm, in which all agents are committed to remaining with the firm. For example the matching process that assigns new graduates of the Naval Academy to their first positions (see Roth and Sotomayor (1990, p. 86)) has the property that all graduates are assigned positions.) Although we will consider unrestricted preferences, we will note that this "all-acceptable" restriction of preferences will permit even stronger results.

We are going to consider matching games whose basic rules are that any firm and worker may be matched together if and only if they both agree. With this in mind, a *blocking pair* for a matching μ is a pair $(f,w) \in F \times W$ such that $w >_f \mu(f)$ and $f >_w \mu(w)$. Of course, if μ is acceptable then each blocking pair for μ is in $A(P)$. A matching is called *stable* if it is acceptable and has no blocking pairs. The idea is that if μ is a matching that admits a blocking pair (f,w), then it is unstable, since f and w would prefer to be matched to each other, and the rules allow them to arrange this. In the marriage model the set of stable matchings equals the core of the game. (In models of many to one matching, the stable matchings are a subset of the core.) We will in fact be considering markets, like the market for new physicians, whose detailed rules involve submitting preference rankings to a centralized matching mechanism. But because firms and workers can recontract, matching mechanisms that produce unstable matchings have typically failed (see Roth (1991)).

We denote the set of all stable matchings by $S(P)$, and say that v is *achievable for* v' if $\mu(v') = v$ for some $\mu \in S(P)$, i.e. if v and v' can be matched at a stable matching. Gale and Shapley (1962) proved that a stable matching exists for any preference profile. They further observed that a deferred acceptance process, in which firms make offers to workers, who can wait and take the best offer they receive, will produce a stable matching. Roth (1984a) observed that an algorithm equivalent to this deferred acceptance algorithm had in fact been incorporated into the centralized matching mechanism developed in the market for new physicians in 1952.[6] Numerous other markets have since adopted equivalent mechanisms (see Roth (1990, 1991) and Roth and Xing (1994)).

The centralized markets that motivate the present paper typically work by having each employer and each worker submit a rank ordered preference list of acceptable matches to a centralized clearinghouse, which then produces a match by processing all the preference lists according to some algorithm such as the

[6] In subsequent years, as medical residencies have become differently structured, modifications in the algorithm to accommodate the market have moved it away from a purely deferred acceptance algorithm (Roth (1996a)).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

following (which is the version studied by Gale and Shapley (1962)). We state the algorithm as if it were decentralized (i.e. as if firms made offers rather than submitted preference lists) in order to emphasize how the algorithm resembles a decentralized procedure for producing matchings. (It is this resemblence that accounts for the independent development of this algorithm in a number of centralized markets, and of related decentralized procedures in other markets (Roth and Xing (1997).)

Firm-Proposing Deferred Acceptance (DA) Algorithm:

Step 1: Each firm makes an offer to the first worker on its preference list of acceptable workers. Each worker rejects the offer of any firm that is unacceptable to her, and each worker who receives more than one acceptable offer rejects all but her most preferred of these, which she "holds."

Step k: Any firm whose offer was rejected at the previous step makes an offer to its next choice (i.e. to its most preferred worker among those who have not yet rejected it), so long as there remains an acceptable worker to whom it has not yet made an offer. If a firm has already offered a position to, and been rejected by, all of the workers it finds acceptable, then it makes no further offers. Each worker receiving offers rejects any from unacceptable firms, and also rejects all but her most preferred among the set consisting of the new offers together with any offer she may have held from the previous step.

Stop: The algorithm stops after any step in which no firm's offer is rejected. At this point, every firm's position is either being held by some worker or has been rejected by every worker on the firm's list of acceptable workers. The output of the algorithm is the matching at which each worker is matched to the firm she is holding when the algorithm stops. Workers who did not receive any acceptable offer, and firms who were rejected by all workers acceptable to them, remain unmatched.

This completes the description of the algorithm, except that we have described it as if all agents have strict preferences. The modification required in case some firm or worker is indifferent between two or more possible matches is simple. At any step of the algorithm at which some agent must indicate a choice between two mates who are equally well liked, introduce some fixed "tie breaking" rule (such as, when a firm is indifferent between making its next offer to either of two workers, break the tie in favor of the worker who has experienced less unemployment, or who is younger, etc.). Such a tie breaking rule therefore specifies to which worker a firm will propose when it is indifferent to whom to make its next offer, and which offer a worker will hold when she is indifferent among more than one most-preferred offer.

The algorithm must eventually stop because there are only a finite number of firms and workers, and no firm offers its position more than once to any worker.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The outcome it produces is a matching, since each firm has an offer out at any step to at most one worker, and each worker is holding at most one offer. This matching is individually rational, since no firm ever makes an offer to an unacceptable worker, nor does any worker ever hold an offer from an unacceptable firm. To see that the matching μ produced by the algorithm is stable, suppose some firm f and worker w are not matched to each other at μ, but f prefers w to its outcome at μ. Then f must have made an offer to w at some step of the algorithm, and eventually been rejected. Therefore w is matched at μ to a firm she likes at least as well as f (because preferences are transitive), and so f and w do not block μ.

Gale and Shapley (1962) further showed that, when all preferences are strict, the matching μ is the firm-optimal stable matching, in the sense that each firm is matched to its most preferred achievable worker. That is, when preferences are strict, the output matching gives each firm the most preferred worker with which it can be matched at any stable matching; this matching is denoted μ_F. There is of course also a worker proposing version of the algorithm, which produces the worker-optimal stable matching.

Among the reasons to study the Firm-Proposing Deferred Acceptance Algorithm is that a number of markets have adopted versions of it as centralized market clearing mechanisms, as an antidote to various kinds of market failure; see Table I.[7]

There are many equivalent versions of the deferred acceptance algorithm (i.e. variations that produce the same matching from the same stated preferences). In some of the proofs that follow it will be convenient to consider a variation in which, at any step of the algorithm, only one firm at a time makes a proposal (see, e.g., McVitie and Wilson (1970)).

As in the centralized procedures we observe in some labor markets, we will assume that while the true preferences of the participants may have some indifferences, their stated preferences must be strict rank orderings. When the stated (strict) preferences are P we denote by $DA(P)$ the outcome of the firm-proposing deferred acceptance procedure, namely the firm-optimal stable matching with respect to the stated preferences.

3. THE INFORMATION REQUIRED FOR STRATEGIC MISREPRESENTATION OF PREFERENCES

In the examples that follow, an agent's preference ordering will be represented by the ordered list of acceptable matches. Thus $P_v = v', v''$ denotes that according to the preference P_v agent v's first choice is to be matched to v', his

[7]The markets mentioned are drawn from Roth (1984, 1990, 1991) and Roth and Xing (1994), and from unpublished notes from those and subsequent investigations. A number of markets have recently been involved in discussions of whether to follow the NRMP and switch to versions of the Worker-Proposing DA Algorithm (see Roth (1996b), Roth and Peranson (1997, 1998)). The results for the marriage model apply equally well to that algorithm, but in the actual markets being studied the fact that firms employ many workers would require a somewhat different treatment.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE I

STABLE AND UNSTABLE CENTRALIZED MECHANISMS

Market	Stable/Based on DA Algorithm	Still in Use
Entry level medical markets:		
U.S. (NRMP)	yes/yes	yes
Edinburgh ('69)	yes/yes	yes
Cardiff	yes/yes	yes
Canada	yes/yes	yes
Cambridge	no	yes
London Hospital	no	yes
Birmingham	no	no
Edinburgh ('67)	no	no
Newcastle	no	no
Sheffeld	no	no
Other markets:		
Medical Specialties		
(approximately 30 markets)	yes/yes	yes
Canadian Lawyers (articling positions)		
Toronto	yes/yes	yes
Vancouver	yes/yes	no (abandoned in '96)
Calgary & Edmonton	yes/yes	yes
Dental Residencies (5 specialties,		
2 general programs)	yes/yes	yes
Osteopaths (< '94)	no	no
Osteopaths (> '94)	yes/yes	yes
Pharmacists	yes/yes	yes
Sororities	yes (at equilibrium)/no	yes

second choice is v'', and his third most preferred alternative is to be unmatched. (Agent v's preferences among unacceptable matches is suppressed in this notation.) Indifference will be denoted by brackets with v, if needed, representing the outcome of being unmatched; e.g., $P_v' = v', [v'', v]$ denotes preferences that differ from P_v in that now agent v is indifferent between being matched to v'' or being unmatched (i.e. to having the outcome $\mu(v) = v''$ or $\mu(v) = v$).

We will be considering the strategic options facing players in a revelation game that employs the firm-proposing deferred acceptance algorithm on the stated preferences. Example 1 shows that some potentially profitable strategic opportunities require detailed knowledge of others' preferences.

EXAMPLE 1: Detailed information needed to manipulate by changing the order of preferences.

Consider a market with three firms and workers, whose true preferences P are

$$P_{f_1} = w_2, w_1, w_3, \qquad P_{w_1} = f_1, f_2, f_3,$$
$$P_{f_2} = w_1, w_2, w_3, \qquad P_{w_2} = f_2, f_1, f_3,$$
$$P_{f_3} = w_1, w_3, w_2, \qquad P_{w_3} = f_1, f_2, f_3.$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

If all agents state their true preferences, the outcome is $DA(P) = [(w_1, f_2), (w_2, f_1), (w_3, f_3)]$, at which w_1 is matched to her second choice. If w_1 instead stated $P'_{w_1} = f_1, f_3, f_2$, then the stated preferences would be $P' = (P_{-w_1}, P'_{w_1})$ with outcome $DA(P') = [(w_1, f_1), (w_2, f_2), (w_3, f_3)]$, at which w_1 is matched to her first choice. Furthermore, P' is a Nash equilibrium of stated preferences—no agent has any incentive (in terms of the true preferences P) to deviate from the stated preferences P'.

To see what is going on, look at the deferred acceptance algorithm operating on P or on P'. Under either set of preferences, f_1 makes its initial offer to w_2, and both f_2 and f_3 make their initial offers to w_1. Under the (true) preferences P, w_1 rejects the offer from f_3, which then makes an offer to w_3, at which point the algorithm stops, at $DA(P)$. But under the preferences P', w_1 instead rejects the offer from f_2, which (unlike f_3) prefers to make its next offer to w_2, the worker who is holding the offer from f_1 that w_1 would really like to get herself. Since (unlike the other workers) w_2 prefers f_2 to f_1, she now rejects f_1, who makes an offer to w_1, who only now rejects f_3. Thus, to detect the profitable opportunity to mis-state her preferences in a way that could otherwise leave her unmatched, w_1 needs detailed information both about firms' preferences and about other workers' preferences.

In view of Example 1, we might hope to be able to show that when workers' information about other workers' and firms' preferences is sufficiently limited, we could always advise them to straightforwardly reveal their true preferences. However Example 2 below shows that certain kinds of profitable strategic manipulations can sometimes be identified even when workers have essentially no information that allows them to distinguish among others' preferences.

EXAMPLE 2: Manipulating by submitting a truncated preference may possibly be profitable even with little information about others' preferences.

Consider a market with two firms and two workers, in which the Firm-Proposing DA Algorithm is employed on the stated (strict) preferences. Suppose that the true preference of w_1 is $P_{w_1} = f_1, f_2$ and that w_1's beliefs about the preferences that might be stated by the other agents w_2, f_1 and f_2 are that they will be independently chosen, each agent will state two acceptable matches, and that each of the two possible orderings will be equally likely.[8]

Suppose (that w_1 believes that) there are eight equally likely possible preference profiles for w_2, f_1 and f_2. The outcomes of w_1 under these preference profiles when she, respectively, submits her true preference ordering $P_{w_1} = f_1, f_2$ and the preference ordering $P'_{w_1} = f_1$ are given by Table II.

[8]It would not change the example in any essential way if we looked at beliefs with positive support on the larger (and perhaps more natural) set of preferences that include those in which some of the other agents may state only one acceptable match. We keep the set small here just so that the set of all possible combinations is simple to enumerate.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE II

$P(w_2), P(f_1), P(f_2)$	Outcome of $w_1(DA(P)(w_1))$ with $P_{w_1} = (f_1, f_2)$	Outcome of $w_1(DA(P')(w_1))$ with $P'_{w_1} = (f_1)$
$(f_1, f_1), (w_1, w_2), (w_1, w_2)$	f_1	f_1
$(f_1, f_2), (w_1, w_2), (w_2, w_1)$	f_1	f_1
$(f_1, f_2), (w_2, w_1), (w_1, w_2)$	f_2	w_1
$(f_1, f_2), (w_2, w_1), (w_2, w_1)$	f_2	w_1
$(f_2, f_1), (w_1, w_2), (w_1, w_2)$	f_1	f_1
$(f_2, f_1), (w_1, w_2), (w_2, w_1)$	f_1	f_1
$(f_2, f_1), (w_2, w_1), (w_1, w_2)$	f_2	f_1
$(f_2, f_1), (w_2, w_1), (w_2, w_1)$	f_1	f_1

So, if the expected utility function of w_1 is $u_{w_1}(.)$, we have that her expected utility when submitting the preferences P_{w_1} and P'_{w_1} is, respectively, $(\frac{5}{8})u_{w_1}(f_1) + (\frac{3}{8})u_{w_1}(f_2)$ and $(\frac{6}{8})u_{w_1}(f_1) + (\frac{2}{8})u_{w_1}(w_1)$; so, w_1 would benefit from submitting $P'(w_1)$ whenever $u_{w_1}(f_1) + 2u_{w_1}(w_1) > 3u_{w_1}(f_2)$. In particular, whenever $u_{w_1}(f_1) - u_{w_1}(f_2)$ is very much larger than $u_{w_1}(f_2) - u_{w_1}(w_1)$ then the incentive for w_1 to list only her first choice will persist for a very wide range of beliefs over the preferences stated by the other market participants.

Finally suppose that instead of having strict preferences, w_1 is indifferent between being matched to f_2 or being unmatched; i.e. suppose her true preferences are given by $P_{w_1} = f_1, [f_2, w_1]$. Then there are *no* preferences P_{-w_1} that the other players might state such that w_1 prefers to reveal her true preferences rather than state P'_{w_1}, but of course the table continues to show that there are preferences P_{-w_1} such that w_1 strictly prefers to state P'_{w_1}. That is, in this case w_1 can identify a strategy that dominates truthful revelation just by examining her own preferences, i.e. regardless of the other agents' preferences.[9]

In Example 2 the profitable manipulation P'_{w_1} is a truncation of the true preferences P_{w_1}. Formally, a *truncation* of a preference list P_w containing k acceptable firms is a list P'_w containing $k' \leq k$ acceptable firms such that the k' elements of P'_w are the first k' elements of P_w, in the same order.

Examples 1 and 2 show that it requires more information to identify profitable opportunities to misstate your preference by changing the order of preference than by truncating your preference, i.e. by shortening the list of acceptable matches without changing their order. In the context of the deferred acceptance procedure, it is easy to get some intuition about why this is the case.

When you change the order of your stated preferences, you are choosing which of two firms f or f' to reject when you have offers from both. Whichever

[9] It might even be argued that in low information environments indifference is not as "knife edge" as in complete information environments, since indifference between two alternatives might be a natural response to lack of information about them. However we will not pursue this argument here, since the kind of low information with which we will be concerned is about other agents' announced preferences, not about their desirability as possible matches.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

firm you reject will cause a chain of further rejections, which may lead to new offers for you. To see if a given misstatement will be profitable, you need to be able to evaluate both chains of rejections; the chain that arises when you reject your more preferred firm, and the chain that arises when you reject your less preferred firm. If you cannot distinguish between the string of new offers that might come to you from the two possible rejection chains that you could set in motion, then the decisive factor will be those cases in which you will get no new offers you prefer, in which case you will be matched to the firm you did not reject, and so should have kept the one you prefer and rejected the less preferred offer.

But misstating your preferences by truncating them is different, because if you hold an offer from f instead of rejecting it (when it is your only offer), you do not generate any chain of further rejections (with the potential to bring new offers your way), as you do when you reject it. This is why it might be profitable to reject an offer from f, even when it is acceptable, and this is what is done in a centralized deferred acceptance algorithm when you submit a preference order that does not include f.

Thus, to truncate or not is a choice between holding some offer or rejecting it (and therefore a choice between holding an offer or initiating a chain of rejections), while to change the order of your stated preferences is a choice between which of two offers to reject (and therefore which of two rejection chains to initiate). To make this intuition precise, we formalize in the next section the idea that matchings are random variables determined by (subjectively) random preference profiles.

4. A MODEL OF SYMMETRIC INFORMATION

In this section we outline a model in which it will be possible to make statements about what a given player knows, or, more precisely, does not know, about other players' announcements. We want to be able to consider cases in which players may not have the information necessary to distinguish among the preferences and strategies of other players in the manner required, e.g., to identify the profitable strategic manipulations of the kind discussed in Example 1. The game we have in mind will continue to be a centralized matching game that produces a stable matching by applying the firm-proposing DA algorithm to participants' stated preferences.

Some notation will be useful, to allow us to speak both about a worker w's decision whether to reverse her stated preference for a given pair of firms, and to model her uncertainty about other workers' preferences for these firms, and about these firms' preferences for workers.

For a given preference profile, denote by P_S the preference orders of the players in the subset $S \subseteq V \equiv F \cup W$. Denote by $P_S^{f \leftrightarrow f'}$ the preference order of the players in S obtained from P by switching f and f', i.e., each worker in S

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

exchanges the places of f and f' in her preference list and if f is in S its preference is $P_{f'}$ and if f' is in S its preference is P_f. For a given worker w, we will continue to use the subscript w for $\{w\}$ and $-w$ for $V \setminus \{w\}$, and so we write $P_w, P_{-w}, P_w^{f \leftrightarrow f'}$, and $P_{-w}^{f \leftrightarrow f'}$. Also, if $S = V$, we omit the subscripts S, e.g., we write P and $P^{f \leftrightarrow f'}$ for P_V and $P_V^{f \leftrightarrow f'}$, respectively.

Note that if worker w's true preferences are given by P_w, then $P_w^{f \leftrightarrow f'}$ is the preference in which she reverses the order of f and f' (but otherwise states her true preferences). Similarly, P_{-w} and $P_{-w}^{f \leftrightarrow f'}$ are assessments by player w of the preferences of all other agents that are identical except that the roles of f and f' are everywhere exchanged.

We will examine random variables $\tilde{P}_{-w}$ that have preference-profiles of all the players in V other than w as their range, and we refer to such a $\tilde{P}_{-w}$ as a *random preference profile* for the players in $V \setminus \{w\}$. We will interpret the random variable $\tilde{P}_{-w}$ as representing w's beliefs about the *stated* preference of the other players, i.e., about their announcements to the centralized mechanism.[10] We will typically consider a fixed worker w with a given true preference ranking P_w. We do not assume that w treats other players as independent; so, the distributions of the component preferences $\tilde{P}_v$ and $\tilde{P}_{v'}$ may be correlated.

We model player w's uncertainty about differences in the preferences of firms f and f', and about other workers' preferences for those firms, as follows. For distinct firms f and f', we say that the random variable $\tilde{P}_{-w}$ is $\{f, f'\}$-*symmetric* if the distributions of $\tilde{P}_{-w}$ and $(\tilde{P}_{-w})^{f \leftrightarrow f'}$ coincide, i.e., if for every specific realization of P_{-w}, $\Pr\{\tilde{P}_{-w} = P_{-w}\} = \Pr\{\tilde{P}_{-w}^{f \leftrightarrow f'} = P_{-w}\}$, or equivalently, $\Pr\{\tilde{P}_{-w} = P_{-w}\} = \Pr\{\tilde{P}_{-w} = P_{-w}^{f \leftrightarrow f'}\}$.

For example, one way in which this situation could arise is if the distributions of the random preferences $\tilde{P}_v$ are independent for $v \in W \cup \{f\} \cup \{f'\}$, the distributions of $\tilde{P}_f$ and $\tilde{P}_{f'}$ coincide, and for each worker $w' \in W \setminus \{w\}$, the distribution of $\tilde{P}_{w'}$ is $\{f, f'\}$-*symmetric*, that is, the distribution of $\tilde{P}_{w'}$ and $(\tilde{P}_{w'})^{f \leftrightarrow f'}$ coincide. Notice that w may know a good deal about f and f' in such a case, e.g. w may know that both firms prefer some w' to w'' with certainty, and that they both have a .8 probabililty of preferring w'' to w herself. What w doesn't know about f and f', if her beliefs are $\{f, f'\}$-symmetric, are any *differences* in their preferences, or in other workers' preferences between them. And of course w knows P_w, so w knows enough about f and f' to form her own preferences between them.[11]

[10] The simplest interpretation is that we are modeling a game in which players have complete information about one anothers' preferences, but nevertheless entertain strategic uncertainty about what others will announce.

[11] For example, a new assistant professor candidate in economics might have {Harvard, MIT}-symmetric beliefs, if, despite knowing which of the two she preferred, she couldn't say which of the two was more likely to rank her highly compared to other top candidates, or which of the two would likely be preferred by other candidates.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Given a worker $w \in W$ and a set of firms $U \subseteq F$, we say that a random preference profile $\tilde{P}_{-w}$ for $V \setminus \{w\}$ is *U-symmetric* if it is (f, f')-symmetric for each pair (f, f') of distinct members of U.

We are now in a position to model complex information structures. For example, let $\{F_1, \ldots, F_p\}$ be a partition of F, i.e. a collection of disjoint sets whose union is F. Then we can consider a worker w whose beliefs $\tilde{P}_{-w}$ about others' preferences are F_k-symmetric for each $k = 1, \ldots, p$. That is, w may have lots of information with which to differentiate firms in different sets F_i and F_j, but insufficient information to differentiate between the preferences to be expected by and about firms within the same set. This information structure is potentially quite general, since the partition in which each set is a singleton firm allows the worker to have highly differentiated information about each other firm and worker, while the partition $\{F\}$ allows us to consider the case of a worker (with $\{F\}$-symmetric beliefs) who cannot distinguish among any of the firms.

A *random matching* is a random variable whose range is the set of all matchings; for example, one gets random matchings as the random outcome of the deferred acceptance algorithm under random preferences of the players. For each random matching $\tilde{\rho}$ we obtain random variables $\tilde{\rho}(v)$ for each player $v \in V$, where each $\tilde{\rho}(v)$ is the (random) assignment of v under $\tilde{\rho}$. The range of $\tilde{\rho}(v)$ is the union of v and the set of members of V from the set of agents opposite to that of v. Given a random matching $\tilde{\rho}$ and $v \in V$, we denote the expectation with respect to $\tilde{\rho}(v)$ of a real-valued function u_v on the range of $\tilde{\rho}(v)$ by $E_{\tilde{\rho}(v)}(u_v)$.

Given two random matchings $\tilde{\rho}^1$ and $\tilde{\rho}^2$, a member w of W and a preference ranking P_w over $F \cup \{w\}$, we say that $\tilde{\rho}^2(w)$ *stochastically P_w-dominates* $\tilde{\rho}^1(w)$, written $\tilde{\rho}^2(w) \succcurlyeq_{P_w} \tilde{\rho}^1(w)$, if for every $v \in F \cup \{w\}$, $\Pr\{\tilde{\rho}^2(w) \geq_{P_w} v\} \geq \Pr\{\tilde{\rho}^1(w) \geq_{P_w} v\}$. Also, a function $u_w\colon F \cup \{w\} \to R$ is called *P_w-monotone* if it is monotone with respect to P_w. Any P_w monotone function can be the expected utility function of some worker whose ordinal preferences are P_w. It is well known that if $\tilde{\rho}^1$ and $\tilde{\rho}^2$ are two random matchings, then $\tilde{\rho}^2(w) \succcurlyeq_{P_w} \tilde{\rho}^1(w)$ if and only if for every P_w-monotone function $u_w\colon F \cup \{w\} \to R$, $E_{\tilde{\rho}^2(w)}(u_w) \geq E_{\tilde{\rho}^1(w)}(u_w)$. That is, $\tilde{\rho}^2(w)$ stochastically P_w-dominates $\tilde{\rho}^1(w)$ if and only if worker w prefers $\tilde{\rho}^2(w)$ to $\tilde{\rho}^1(w)$ regardless of her expected utility function u_w (corresponding to the preferences P_w).

In what follows, since the players will compare risky outcomes, they can be assumed to have expected utility functions. But our principal results will be that certain strategies will stochastically dominate others. Consequently these results will hold for arbitrary utility functions, and so we will be able to avoid assumptions beyond the ordinal preferences of the players.

5. STRATEGIC BEHAVIOR WITH SYMMETRIC INFORMATION

The labor market mechanisms discussed earlier all require participants to submit rank orderings (i.e. strict preferences), and it will be convenient to

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

continue to assume in what follows that all preferences are strict.[12] Some of our results depend on lemmas about the mechanics of matching, independent of agents' information, and these are presented in the Appendix.

Our first result says that a worker whose information about two firms is symmetric can never, regardless of her attitude towards risk, improve her outcome in the match by simply misstating her preferences between them. Its proof is immediate from Lemmas A3 and A4 in the Appendix.

THEOREM 1: *Let $w \in W$, P_w be a preference ranking over $F \cup \{w\}$, and let f and f' be two firms such that $f' <_{P_w} f$. Then for every $\{f, f'\}$-symmetric random preference profile $\tilde{P}_{-w}$ for the players in $V \setminus \{w\}$, $DA[P_w, \tilde{P}_{-w}](w) \succcurlyeq_{P_w} DA[P_w^{f \leftrightarrow f'}, \tilde{P}_{-w}](w)$.*

Theorem 1 provides sufficient conditions for a worker not to have an incentive to misrepresent her preferences by a simple switch of two positions. It demonstrates that it is never profitable for a worker whose information is symmetric with respect to two firms f and f' to simply misstate her stated preferences between them: the (random) outcome of such a strategy is stochastically dominated by the truthful revelation of preferences. Much of the intuition behind this result can be gotten by examining the proof of Lemma A3 in the Appendix. Basically what is going on is that whichever of the firms $\{f, f'\}$ the worker places before the other on her preference list, there are some states of the world (i.e. preferences of the other agents) in which she will do better than if she had announced the opposite ordering of the two firms. But because her information about f and f' is symmetric, the more favorable outcomes are as likely to come from putting the firms in the order of her true preferences as in the reverse order. And the tie-breaker comes from the fact that there are states of the world in which she will be matched with whichever of the two firms she has announced is her more preferred, in which case she does better with her true preferences.

We turn next to consider the simple special case in which a worker's beliefs are F-symmetric, i.e. in which they are $\{f, f'\}$-symmetric for all firms $f, f' \in F$.

5.1. *Strategic Behavior with Completely Symmetric Information*

A corollary of the proof of Theorem 1 is that when worker w's information is $\{F\}$-symmetric, she will never have an incentive to change her preference ordering of firms.

COROLLARY 1: *For a worker with $\{F\}$-symmetric information, any strategy that changes her true preference ordering of firms is stochastically dominated by a strategy*

[12] It appears that this assumption can be interpreted in what follows as if individual agents may in fact be indifferent between some alternatives, but break ties at random when called upon to submit a strict preference.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

that states the same number of acceptable firms in their correct order. That is, let $w \in W$ *and let* P_w, P'_w, *and* P''_w *be preference rankings for w such that:*

(i) *the restriction of* P_w *and* P''_w *to F coincide, and*

(ii) $|\{f \in F: f >_{P'_w} w\}| = |\{f \in F: f >_{P''_w} w\}|$.

Then for every random preference profile $\tilde{P}_{-w}$ *for the players in* $V \setminus \{w\}$ *that is* $\{F\}$*-symmetric,* $DA[P''_w, \tilde{P}_{-w}](w) \succcurlyeq_{P_w} DA[P'_w, \tilde{P}_{-w}](w)$.

In addition, a corollary of a known result (Theorem 2.24 in Roth and Sotomayor (1990)), allows us to state the following lemma, saying that if a worker is contemplating stating a preference that doesn't change the relative position of any firms, then she can never profit from ranking as acceptable any firm that is in fact unacceptable. (This result does not depend in any way on the worker's information or beliefs.)

LEMMA 1: *Suppose* $w \in W$, $P = (P_w, P_{-w})$ *is a preference-profile and* P''_w *is a preference ranking over* $F \cup \{w\}$ *such that:*

(i) *the restrictions of* P_w *and* P''_w *to F coincide, and*

(ii) $|\{f \in F: f >_{P''_w} w\}| \geq |\{f \in F: f >_{P_w} w\}|$.

Then $DA[P](w) \geq_{P_w} DA[P''_w, P_{-w}](w)$.

Taken together, Corollary 1 and Lemma 1 say that, if a worker with $\{F\}$-symmetric information isn't going to truncate her preferences, she had better state her true preferences.

COROLLARY 2: *Let* $w \in W$ *and let* P_w *and* P'_w *be preference rankings for w where* $|\{f \in F: f >_{P'_w} w\}| = |\{f \in F: f >_{P_w} w\}|$. *Then for every random preference* $\tilde{P}_{-w}$ *for the players in* $V \setminus \{w\}$ *that is* $\{F\}$*-symmetric,* $DA[P_w, \tilde{P}_{-w}](w) \succcurlyeq_{P_w} DA[P'_w, \tilde{P}_{-w}](w)$.

PROOF: The conclusion of the corollary is immediate from Corollary 1 with $P''_w = P_w$. *Q.E.D.*

We can now state our chief result for workers with completely symmetric information.

THEOREM 2: *For a worker with* $\{F\}$*-symmetric information, any non-truncation strategy is stochastically dominated by a truncation of the true preferences. That is, let* $w \in W$ *and let* P_w *and* P'_w *be preference rankings for w. Let* P^*_w *be the truncation of* P_w *with*

$$|\{f \in F: f >_{P^*_w} w\}| = \min\left\{|\{f \in F: f >_{P_w} w\}|, |\{f \in F: f >_{P'_w} w\}|\right\}. \tag{5.1}$$

Then for every random preference $\tilde{P}_{-w}$ *for the players in* $V \setminus \{w\}$ *that is* $\{F\}$*-symmetric*

$$DA\left[P^*_w, \tilde{P}_{-w}\right](w) \succcurlyeq_{P_w} DA\left[P'_w, \tilde{P}_{-w}\right](w). \tag{5.2}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

PROOF: Let $k \equiv |\{f \in F: f >_{P'_w} w\}|$ and consider the unique P''_w such that the restrictions of P''_w and P_w to F coincide and such that $|\{f \in F: f >_{P''_w} w\}| = k$. It then follows by Corollary 1 that

$$DA\left[P''_w, \tilde{P}_{-w}\right](w) \succcurlyeq_{P_w} DA\left[P'_w, \tilde{P}_{-w}\right](w).$$

Now, if $k \leq |\{f \in F: f >_{P_w} w\}|$, then P''_w is a truncation of P_w, and $P^*_w \equiv P''_w$ satisfies (5.1) and (5.2). Alternatively, if $k \geq |\{f \in F: f >_{P_w} w\}|$, then Lemma 1 implies that $DA[P_w, P_{-w}](w) \geq_{P_w} DA[P''_w, P_{-w}](w)$ for every preference profile P_{-w} of the players in $V \setminus \{w\}$, immediately implying that for every random preference $\tilde{P}_{-w}$ for the players in $V \setminus \{w\}$

$$DA\left[P_w, \tilde{P}_{-w}\right](w) \succcurlyeq_{P_w} DA\left[P''_w, \tilde{P}_{-w}\right](w).$$

It now follows, respectively, from the transitivity of $\succcurlyeq_{P_w}$ and the assumption $k > |\{f \in F: f >_{P_w} w\}|$ that $P^*_w \equiv P_w$ satisfies (5.2) and (5.1). *Q.E.D.*

While Theorem 2 is stated only for the simple special case of a worker with very limited information, it is a result that has at least the right *form* to allow us to begin to answer the question of the medical school dean quoted in our Introduction. We can advise such students that to rank hospitals in other than the order of their true preferences is a stochastically dominated action. That is, we can advise them that they should rank hospitals in the order of their preferences. This is true regardless of their attitudes towards risk.

What Theorem 2 does not allow us to assert, and what Example 2 shows that we cannot assert, is that it is a stochastically dominant strategy for (even) such a student to straightforwardly reveal her full true preferences. Rather, such a student is left with a balance of risks, the resolution of which does depend on her risk posture. To submit a shorter preference list increases the risk of being unmatched. But to submit a longer preference list also has risks, as it may lower the probability of being matched to a more favored outcome.[13]

This balance of risks is of course absent in any market in which workers are committed to taking one of the jobs on offer, and must therefore rank all jobs as acceptable. (Recall the earlier discussion of the situation facing new graduates of the Naval Academy.) If such a market is organized by a firm-optimal stable matching mechanism, then truthful revelation of preferences becomes the unique stochastically dominant strategy for workers with $\{F\}$-symmetric information. That is, we have the following corollary of Theorem 2, which is similar in form to restricted preference results in the mechanism design and social choice literature, except that it also includes restrictions on players' information.

[13] The computational experiments discussed in Roth (1996b) were designed to allow the magnitudes of these different risks to be assessed in the market for new physicians. It was found that the size and transaction costs (of interviewing) in that market combine to make the latter risk negligible (Roth and Peranson (1997, 1998).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

COROLLARY 3: *In games that use the F-optimal stable mechanism with the restriction that all positions must be ranked (i.e. all firms are acceptable), truthful revelation of preferences P_w is a stochastically dominant strategy for a worker with $\{F\}$-symmetric information. That is, for any strategy P'_w,*

$$DA\left[P_w, \tilde{P}_{-w}\right](w) \succcurlyeq_{P_w} DA\left[P'_w, \tilde{P}_{-w}\right](w).$$

We next informally consider more general information structures and misrepresentations.

6. STRATEGIC BEHAVIOR IN GENERAL INFORMATION STRUCTURES

It is possible to show that a player with $\{F_k\}$-symmetric beliefs cannot do better than to submit a preference whose restriction to F_k is a truncation of the true preferences restricted to F_k, but this says nothing about the ordering of firms in different parts of the partition $\{F_1, \ldots, F_p\}$. We might hope to show that a worker should confine her attention to some kind of truncation strategy that preserved the relative order of firms about which she has different information. But the following example shows that we cannot get such a result, because this kind of information structure is so general that it allows a great deal of information to be conveyed.

EXAMPLE 3: A game in which listing an unacceptable candidate (a nontruncation strategy) is optimal.

Consider a game with three firms and two workers, in which worker w_1's preferences are $P(w_1) = f_1, f_2$, and she knows that the others' (stated) preferences are given by $P(f_2) = w_1, w_2$, and $P(w_2) = f_2, f_1, f_3$, and that the preferences of firms f_1 and f_3 are always identical and are given by

$$P(f_1) = P(f_3) = w_2, w_1 \text{ with probability } 1/2$$

("the good state of the world for w_1"); and

$$P(f_1) = P(f_3) = w_2 \text{ with probability } 1/2$$

("the bad state of the world for w_1").

When the outcome is determined by the deferred acceptance procedure with firms proposing, the optimal misrepresentation for w_1 is to state $P'(w_1) = f_1, f_3, f_2$ which involves listing the unacceptable f_3 as acceptable. $P'(w_1)$ is optimal because w_1's outcome, $DA(P_{-w_1}, P'(w_1))(w_1)$ is f_1 in the good state of the world and f_2 in the bad state. (If f_3 ever proposes to w_1, then so will f_1 if rejected by w_2, since f_1 and f_3 have perfectly correlated preferences. And if w_1 rejects f_2 when f_3 proposes, then w_2 will reject f_1. So listing f_3 is better for w_1 than truncating at f_1 or stating her true preferences, since she is matched to f_1 whenever it is willing to employ her, without foregoing f_2 otherwise.)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Note that if we modified the example slightly so that worker w_2 has the preferences $P(w_2) = f_2, f_1, f_3$, with probability $1/2$ and f_2, f_3, f_1 with probability $1/2$, then w_1's optimal misrepresentation would remain the same, but now worker w_1 would have $\{f_1, f_3\}$-symmetric information, and indeed the optimal strategy, while not a truncation, preserves w_1's true preference order on the set $\{f_1, f_3\}$.

While Example 3 shows that truncation strategies may not always be optimal for workers with more complex information and beliefs, the proofs and results about match mechanics in the Appendix can begin to give us some informal idea of what kinds of advice might be appropriate when information is more structured.

For example, during the 1997 NRMP match one of us was contacted by a student who was applying for residencies in a highly competitive (i.e. low match rate) speciality with very few accredited positions (Dermatology[14]), who had also interviewed at some positions in a much less competitive speciality (Internal Medicine). He wanted advice on whether he should truncate his preference list so as to include only Dermatology choices, or whether he should also include some Internal Medicine positions on the end of his list.

It might be a reasonable approximation of this student's information to say that he had symmetric beliefs within the set of Dermatology positions to which he had applied, and within the set of Internal Medicine positions. Given the considerable over demand for Dermatology positions, he might reasonably approximate the preferences of other students with the assumption that no student preferred any Internal Medicine position to any Dermatology position (however this kind of information would not be publicly available). This assumption is almost sufficient to rule out the possibility that, by deleting an Internal Medicine position from the end of his list, he could match to a Dermatology position to which he would not otherwise match. (Formally, this assumption rules out certain two-player cycles in the preferences, and a sufficient condition would require ruling out similar cycles among any number of players.) This would mean that deleting an Internal Medicine position could not cause the student to match to a Determatology position to which he would not have matched without the deletion, but might cause him to be unmatched instead of matched to the Internal Medicine position.

Now, there is a very small number of Dermatology positions that are still vacant after the match, and an unmatched student could hope to obtain one of these while a matched student cannot. So there might be a very small probability of a benefit from deleting an Internal Medicine position from the end of the submitted preference list, and this can be weighed against the much larger probability that this deletion would leave the student unmatched. The balance

[14] Dermatology is regarded as a very desirable specialty for "lifestyle" reasons, in that. unlike other specialties with tight limits on the number of new entrants, the nature of the diseases treated mean that a dermatologist can hope to work regular hours, with few midnight emergencies.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

of these risks cannot be judged without knowing the student's expected utility function: a student whose preferences for Dermatology are lexicographic should truncate his list after the last Dermatology position, while one who finds an Internal Medicine position much more desirable than scrambling for any position in the secondary market following the match should not. But such a student can be advised that this is the risk he must weigh, and can concentrate on the alternatives of a truncation that includes only Dermatology positions and one that includes Internal Medicine positions at the end as well.

7. DISCUSSION

Adaptations of the firm-proposing deferred acceptance algorithm have been successfully implemented in several dozen professional labor markets and submarkets. Why does this kind of market mechanism work so well? The theory suggests that the key to success is that it produces stable matchings, and the empirical evidence supports this, since unstable mechanisms typically fail.

But how should workers behave in such markets, and why do such markets perform as they do? Theorem 2 suggests an approach to these related questions. On the matter of advice to individual workers, it says that, when they have little information about differences among other players, they can do no better than to reveal a truncation of their true preferences.[15] And this offers a suggestion about what might account for the success of the markets organized via variants of the deferred acceptance procedure.

If every worker submits a truncation of her true preferences, then there will not be any instabilities involving blocking pairs of matched firms and workers.[16] There may however be blocking pairs involving a firm and an unmatched worker (see Blum, Roth, and Rothblum (1997)). But these may be difficult to detect, particularly if the way workers' truncate their preferences involves not interviewing with less preferred firms. In such a case, unmatched workers would have

[15]A complementary approach is explored in Roth and Peranson (1997b), where it is shown that, in large markets with high interviewing costs, the probability is small that even a completely informed agent will have a profitable strategic opportunity to misstate his preferences.

[16]Of course Theorem 2 hardly allows us to *advise* every worker (or even most workers) to submit a truncation strategy, since taken literally its assumption about a worker's information about other agent's preferences will not apply to most workers. But we use the assumption of symmetric information to model the case in which workers cannot calculate that one chain of offers is more likely than another to result from one rejection than from another. This inability could arise from other causes than symmetric information about preferences; e.g., it could arise from lack of information about the details of the algorithm. This however is difficult to model, and it is customary in the game theoretic literature to assume that players know both the rules and all logical inferences that can be derived from these rules. This assumption is far from compelling in describing observed behavior in labor markets. To put it another way, when we are not advising workers, but are simply trying to explain their behavior, we may want to consider the assumption of symmetric information not merely literally, but also as a metaphor for a wider class of uncertainty. It is this more inclusive interpretation that may apply to a large set of workers. (See Rubinstein (1991) for some more extended reflections on modeling, which we read in this spirit, and see Barbera and Dutta (1995) for a model of truth telling as very risk averse behavior.)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

exhausted all of their immediate opportunities to match, and would have to enter the secondary market which follows the centralized match.[17] Thus one of the properties of the deferred acceptance procedure as a market mechanism is that it has good performance properties even when participants have little information. This is likely one of the key ingredients of its success.

In general, to understand what kind of mechanisms function well in the field, we may need to learn more about their *robustness* to the assumptions we make about what participants know, and how they behave. As mechanism design moves from the realm of pure theory into the realm of a practical design technology, there will be much to learn.

Dept. of Economics, Harvard University, Cambridge, MA 02138, U.S.A., and Harvard Business School, Boston, MA 02163, U.S.A.

and

Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 32000, Israel

Manuscript received September, 1996; final revision received January, 1998.

APPENDIX: THE MECHANICS OF SIMPLE SUBSTITUTIONS

Throughout we assume that the McVite-Wilson version of the deferred acceptance algorithm is applied; that is, at each stage an arbitrarily selected firm whose offer in not being held makes an offer to its most preferred acceptable worker to whom it has not yet made an offer. The algorithm terminates at a stage at which all firms either have their offer held by some worker, or else have proposed to all their acceptable workers. The outcome is then independent of the particular sequence of proposals, i.e. the outcome is the same for all executions of the algorithm.[18]

In the following two lemmas and their corollaries we examine how the outcome matching is influenced by switching two firms in some worker's preferences. (The results we obtain do not depend on participants' information or beliefs.)

Lemma A1 first considers what happens if two firms f and f' are interchanged at every place in which they play a role, i.e. if the preference profile P is replaced by $P^{f \leftrightarrow f'}$. This affects only those workers who are matched to one of f or f'.

LEMMA A1: *Let P be a profile, let w be a worker, let f and f' be distinct firms, and let $v \in F(\cup\{w\}) \setminus \{f, f'\}$. Then:*

(a) *$DA[P^{f \leftrightarrow f'}](w) = v$ if and only if $DA[P](w) = v$, and*

(b) *$DA[P^{f \leftrightarrow f'}](w) = f$ if and only if $DA[P](w) = f'$.*

PROOF: The Lemma follows immediately from the fact that any execution of the deferred acceptance algorithm with the profile P is also feasible for the profile $P^{f \leftrightarrow f'}$ except that players f and f' will exchange roles. *Q.E.D.*

[17] See Theorem 3 in Roth and Vande Vate (1991), which concerns the stability properties of truncation strategies in the context of decentralized stable matching of the sort proposed in Roth and Vande Vate (1990). The market forces that act on workers in labor markets do not reach out and correct workers who search too little.

[18] Because when preferences are strict there is a unique firm optimal stable matching, and this is selected regardless of the order in which proposals are made.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

As a corollary, we can now see what happens when a given worker w changes the position of two firms in her preference ordering, and compare this with the situation in which the two firms are switched in the preferences P_{-w}. (This will set the stage for considering the problem facing worker w when she doesn't have the information to distinguish between P_{-w} and $P_{-w}^{f \leftrightarrow f'}$.)

COROLLARY A1: *Let P be a profile, let w be a worker, let f and f' be distinct firms, and let $v \in (F \cup \{w\}) \setminus \{f, f'\}$. Then:*

(a) *$DA[P_w^{f \leftrightarrow f'}, P_{-w}](w) = v$ if and only if $DA[P_w, P_{-w}^{f \leftrightarrow f'}](w) = v$, and*

(b) *$DA[P_w^{f \leftrightarrow f'}, P_{-w}](w) = f$ if and only if $DA[P_w, P_{-w}^{f \leftrightarrow f'}](w) = f'$.*

PROOF: Observing that $[P_w^{f \leftrightarrow f'}, P_{-w}]^{f \leftrightarrow f'} = [P_w, P_{-w}^{f \leftrightarrow f'}]$, the conclusion of the corollary follows from the application of Lemma A1 to the profile $[P_w^{f \leftrightarrow f'}, P_{-w}]$. *Q.E.D.*

The next Lemma and its Corollary state that if a worker w states a preference which switches the position of two firms in her true preferences, this cannot cause her to be matched to the more preferred of those firms. (Recall that the profitable misstatement of the preferences in Example 1 involved switching the position of two firms in order to be matched to a third firm that was more preferred than either.)

LEMMA A2: *Let P be a profile, let w be a worker, and let f and f' be firms such that $f' <_{P_w} f$. If $DA[P](w) = f'$, then $DA[P_w^{f \leftrightarrow f'}, P_{-w}](w) = f'$.*

PROOF: Suppose $DA[P](w) = f'$ and consider a given execution of the deferred acceptance algorithm under profile P. It terminates with w being matched to f', hence, during the execution, w will not receive any offer from a firm which is ranked higher than f' according to P_w. It follows that the same sequence of offers, acceptances and rejections can be executed when w submits the preference $P_w^{f \leftrightarrow f'}$ rather than P_w. In particular, once w is matched with f', she will not get better offers with respect to $P_w^{f \leftrightarrow f'}$ as any firm which is preferred to f', according to $P_w^{f \leftrightarrow f'}$ is also preferred to f' according to P_w. So, we have an execution of the deferred acceptance algorithm which matches w with f', implying that $DA[P_w^{f \leftrightarrow f'}, P_{-w}](w) = f'$. *Q.E.D.*

Lemma A2 can be easily extended to show that if $v <_{P_w} f' <_{P_w} f$, then $DA[P_w^{f \leftrightarrow f'}, P_{-w}](w) = v$ if and only if $DA[P](w) = v$. But, this is not used in the forthcoming development.

COROLLARY A2: *Let P be a profile, let w be a worker, and let f and f' be firms such that $f' <_{P_w} f$. If $DA[P_w^{f \leftrightarrow f'}, P_{-w}](w) = f$, then $DA[P](w) = f$.*

PROOF: Consider the profile $P' \equiv [P_w^{f \leftrightarrow f'}, P_{-w}]$ and observe that the assumption $w <_{P_w} f' <_{P_w} f$ implies that $w <_{P'_w} f <_{P'_w} f'$. The conclusion of the corollary now follows from an application of Lemma A2 to P' with an exchange of the roles of f and f'. *Q.E.D.*

We next show that given a profile P, the above results allow one to determine the outcome of the deferred acceptance algorithm with profiles $(P_w, P_w^{f \leftrightarrow f'})$ and $(P_w^{f \leftrightarrow f'}, P_{-w}^{f \leftrightarrow f'})$ from the outcome of the algorithm on the profiles $P = (P_w, P_{-w})$ and $(P_w^{f \leftrightarrow f'}, P_{-w})$. Specifically, let w be a member of W, and let f and f' be two distinct members of F with $f' <_{P_w} f$. Suppose

$$DA[P_w, P_{-w}](w) = u \quad \text{and} \quad DA[P_w^{f \leftrightarrow f'}, P_{-w}](w) = v. \tag{7.1}$$

Lemma A2 and Corollary A2 imply that if $u = f'$ then necessarily $v = f'$, and if $v = f$ then necessarily $u = f$. With $*$ denoting excluded situations, the possibilities are summarized by the six cases listed in Table III.

Lemma A1 and Corollary A1 allow us to determine $DA[P_w, P_{-w}^{f \leftrightarrow f'}]$ and $DA[P_w^{f \leftrightarrow f'}, P_{-w}^{f \leftrightarrow f'}]$ for each of these six cases. The conclusions are summarized in Table IV.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE III

CASES FOR $u = DA[P_w, P_{-w}](w)$ AND $v = DA[P_w^{f \leftrightarrow f'}, P_{-w}](w)$

u \ v	$\in (F \cup \{w\}) \setminus \{f, f'\}$	$= f'$	$= f$
$\in (F \cup \{w\}) \setminus \{f, f'\}$	Case A	Case B	*
$= f$	Case C	Case D	Case E
$= f'$	*	Case F	*

To understand Table IV, first observe that in Case A, w experiences a loss of u and a gain of v as a result of declaring her preferences as $P_w^{f \leftrightarrow f'}$ rather than the true preferences P_w, when the preferences of the others are represented by P_{-w}. Table IV also shows that w experiences the reverse change as a result of that misrepresentation when the preferences of the others are given by $P_{-w}^{f \leftrightarrow f'}$, namely a loss of v and a gain of u. So, her net gain/loss from the misrepresentation when the preferences of the others are represented by P_{-w} is exactly offset by a net loss/gain when the others' preferences are given by $P_{-w}^{f \leftrightarrow f'}$. Similar opposite influences occur in Cases B, C, and D, except that a gain of f' is offset by a loss of f or a loss of f is offset by a gain of f' (and recall that $f' <_{P_w} f$). Misrepresentation causes no changes in cases E or F.

We consider workers w who regard the preference profile $\tilde{P}_{-w}$ as a random variable whose distribution is symmetric between at least two firms.

The next lemma provides the first precise statement asserting that when a worker does not have enough information to distinguish between other workers' preferences concerning two firms f and f', and between the preferences of these firms, then it is never profitable for her to misstate her own preferences between them.

LEMMA A3: *For any worker w with preferences P_w over $F \cup \{w\}$ let f and f' be two firms such that $f' <_{P_w} f$. Suppose the profile of preferences $\tilde{P}_{-w}$ for all players other than w is an $\{f, f'\}$-symmetric random variable. Then for some $\gamma \geq 0$*

$$(7.2) \qquad \Pr\{DA[P_w^{f \leftrightarrow f'}, \tilde{P}_{-w}](w) = v\} - \Pr\{DA[P_w, \tilde{P}_{-w}](w) = v\}$$

$$= \begin{cases} 0 & \text{if } v \in (F \cup \{w\}) \setminus \{f, f'\}, \\ -\gamma & \text{if } v = f, \\ \gamma & \text{if } v = f'. \end{cases}$$

PROOF OF LEMMA A3: The conclusion follows from counting the relevant boxes of Table IV (and making the observation that, with P_{-w} and $P_{-w}^{f \leftrightarrow f'}(w)$ equally likely, these are balanced in the required way, as described in the paragraph that accompanies the Table). *Q.E.D.*

TABLE IV

$DA[P_w, P_{-w}^{f \leftrightarrow f'}](w)$ AND $DA[P_w, P_{-w}^{f \leftrightarrow f'}](w)$ AS DETERMINED BY $DA[P_w, P_{-w}](w)$ AND $DA[P_w^{f \leftrightarrow f'}, P_{-w}](w)$.

	$DA[P_w, P_{-w}](w)$	$DA[P_w^{f \leftrightarrow f'}, P_{-w}](w)$	$DA[P_w, P_{-w}^{f \leftrightarrow f'}](w)$	$DA[P_w^{f \leftrightarrow f'}, P_{-w}^{f \leftrightarrow f'}](w)$
Case A	$u \notin \{f, f'\}$	$v \notin \{f, f'\}$	v	u
Case B	$u \notin \{f, f'\}$	$v = f'$	f	u
Case C	$u = f$	$v \notin \{f, f'\}$	v	f'
Case D	$u = f$	$v = f'$	f	f'
Case E	$u = f$	$v = f$	f'	f'
Case F	$u = f'$	$v = f'$	f	f

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Lemma A4 will allow us to restate Lemma A3 in terms of stochastic dominance.

LEMMA A4: *Let $\tilde{\rho}^1$ and $\tilde{\rho}^2$ be two random matchings, let w be a member of W, let P_w be a preference ranking over $F \cup \{w\}$, and let f and f' be two members of F such that $f' <_{P_w} f$. If for some $\gamma \geq 0$,*

$$\Pr\{\tilde{\rho}^1(w) = v\} - \Pr\{\tilde{\rho}^2(w) = v\} = \begin{cases} 0 & \text{if } v \in (F \cup \{w\}) \setminus \{f, f'\}, \\ -\gamma & \text{if } v = f, \\ \gamma & \text{if } v = f', \end{cases} \tag{7.3}$$

then $\tilde{\rho}^2(w) \gg_{P_w} \tilde{\rho}^1(w)$.

PROOF OF LEMMA A4: Summations of (7.3) imply that for $v \in F \cup \{w\}$,

$$\Pr\{\tilde{\rho}^1(w) \geq_{P_w} v\} - \Pr\{\tilde{\rho}^2(w) \geq_{P_w} v\} \begin{cases} 0 & \text{if } f <_{P_w} v, \\ -\gamma & \text{if } f' <_{P_w} v \leq_{P_w} f, \\ 0 & \text{if } v \leq_{P_w} f'. \end{cases}$$

Hence the definition of stochastic dominance directly implies that $\tilde{\rho}^2(w) \gg_{P_w} \tilde{\rho}^1(w)$. *Q.E.D.*

REFERENCES

AMERICAN MEDICAL STUDENTS ASSOCIATION AND PUBLIC CITIZEN HEALTH RESEARCH GROUP (1995): "Report on Hospital Bias in the NRMP," http://pubweb.acns.nwu.edu/alan/nrmp2.html (accessed 1/6/96).

BLUM, YOSSI, ALVIN E. ROTH, AND URIEL G. ROTHBLUM (1997): "Vacancy Chains and Equilibration in Senior-Level Labor Markets," *Journal of Economic Theory*, 76, 362–411.

BARBERA, SALVADOR, AND B. DUTTA (1995): "Protective Behaviour in Matching Models," *Games and Economic Behavior*, 8, 281–296.

GALE, DAVID, AND LLOYD S. SHAPLEY (1962): "College Admissions and the Stability of Marriages," *American Mathematical Monthly*, 69, 9–15.

GALE, DAVID, AND MARILDA SOTOMAYOR (1985): "Ms Machiavelli and the Stable Matching Problem," *American Mathematical Monthly*, 92, 261–268.

MCVITIE, D., AND L. B. WILSON (1970): "Stable Marriage Assignment for Unequal Sets," *BIT*, 10, 295–309.

PERANSON, ELLIOTT, AND RICHARD R. RANDLETT (1995a): "The NRMP Matching Algorithm Revisited: Theory Versus Practice," *Academic Medicine*, 70, 477–484.

——— (1995b): "Comments on Williams' 'A Reexamination of the NRMP Matching Algorithm'," *Academic Medicine*, 70, 490–494.

ROTH, ALVIN E. (1984a): "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *Journal of Political Economy*, 92, 991–1016.

——— (1984b): "Misrepresentation and Stability in the Marriage Model," *Journal of Economic Theory*, 36, 383–387.

——— (1990): "New Physicians: A Natural Experiment in Market Organization," *Science*, 250, 1524–1528.

——— (1991): "A Natural Experiment in the Organization of Entry Level Labor Markets: Regional Markets for New Physicians and Surgeons in the U.K.," *American Economic Review*, 81, 415–440.

——— (1995): "Proposed Research Program: Evaluation of Changes to be Considered in the NRMP Algorithm," Consultant's Report, http://www.pitt.edu/alroth/nrmp.html.

——— (1996a): "Interim Report 1: Evaluation of the Current NRMP Algorithm, and Preliminary Design of an Applicant-Processing Algorithm," Consultant's Report, http://www.pitt.edu/alroth/nrmp.html.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

——— (1996b): "Report on the Design and Testing of an Applicant Proposing Matching Algorithm, and Comparison with the Existing NRMP Algorithm," Consultant's Report, http://www.pitt.edu/alroth/phase1.html.

ROTH, ALVIN E., AND ELLIOTT PERANSON (1997): "The Effects of the Change in the NRMP Matching Algorithm," *Journal of the American Medical Association*, 278, September 3, 729–732.

——— (1998): "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design," mimeo, Harvard University.

ROTH, ALVIN E., AND MARILDA O. A. SOTOMAYOR (1990): *Two Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Econometric Society Monograph Series. New York: Cambridge University Press.

ROTH, ALVIN E., AND JOHN H. VANDE VATE (1990): "Random Paths to Stability in Two Sided Matchings," *Econometrica*, 58, 1475–1480.

——— (1991): "Incentives in Two-Sided Matching with Random Stable Mechanisms," *Economic Theory*, 1, 31–44.

ROTH, ALVIN E., AND XIAOLIN XING (1994): "Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions," *American Economic Review*, 84, 992–1044.

——— (1997): "Turnaround Time and Bottlenecks in Market Clearing: Decentralized Matching in the Market for Clinical Psychologists," *Journal of Political Economy*, 105, 284–329.

RUBINSTEIN, ARIEL (1991): "Comments on the Interpretation of Game Theory," *Econometrica*, 59, 909–924.

WILLIAMS, KEVIN JON (1995a): "A Reexamination of the NRMP Matching Algorithm," *Academic Medicine*, 70, 470–476.

——— (1995b): "Comments on Peranson and Randlett's 'The NRMP Matching Algorithm Revisited: Theory versus Practice'," *Academic Medicine*, 70, 485–489.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.