

Stability and Deferred Acceptance: Strategic Behavior in Two-Sided Matching

By CLAYTON R. FEATHERSTONE AND ERIC MAYEFKY*

Draft: February 28, 2010

When matching mechanisms yield assignments that are not stable, offers unravel, and, eventually, participants abandon the match. This logic is thought to provide a reasonable explanation of why, empirically, the stable Deferred Acceptance (DA) mechanism persists where non-stable alternatives, such as Priority mechanisms, are abandoned. Theory, however, tells us that both Deferred Acceptance (DA) and Priority mechanisms can yield unstable matches in incomplete information equilibrium. If, however, under DA, match participants deviate from equilibrium by truthfully revealing, then the empirical difference between DA and Priority can be explained. In an experiment, we find evidence that, while participants are able to learn to deviate from truth-telling under a Priority mechanism, they fail to do so under DA. This implies that out-of-equilibrium truth-telling might help to explain why DA persists in spite of being unstable in equilibrium.

JEL: C78, C92, D02

Keywords: Matching, deferred acceptance, stability, experiments

The fundamental question this paper examines is why some two-sided matching mechanisms are used persistently from year to year while others are abandoned. Although the usual distinction concerns whether a mechanism is stable with respect to the reported preferences¹, such an explanation is incomplete without also considering whether preferences are truthfully revealed. Previous theoretical literature has looked at large markets to do this; however, we take a different tack by observing strategic preference revelation

* Featherstone: Department of Economics, Stanford University. Mayefsky: Department of Economics, Stanford University.

¹One might also bypass truthful preference revelation entirely and simply look at whether a mechanism yields a stable allocation in equilibrium. See, for instance, Ergin and Sönmez (2006).

in the lab. Our evidence suggests that an important part of the difference between mechanisms that persist and mechanisms that don't is how well match participants are able to learn that deviation from truthful preference revelation can be profitable.

Two-sided matching mechanisms are widely used in the field. The most well-known example is the National Residency Matching Program (NRMP) which every year makes about 25,000 matches between newly-minted doctors and residency programs in the United States (NRMP 2009). Once participants have formed their preferences, they submit rank-order lists of acceptable match partners to the NRMP clearinghouse, which then runs those lists through an algorithm, outputting a match. The true preferences of match participants remain private information. Other examples of two-sided matching include the Association of Psychology and Post-doctoral Internship Centers (APPIC) match (about 2,800 clinical psychologists matched to internship programs per year (APPIC 2009)), and the New York City Department of Education public high school match (about 90,000 high school students per year (NYC-DOE 2009)).

When deciding which mechanism to use in a matching market, the literature has consistently come back to the idea of stability. A **blocking pair** consists of two agents who prefer each other to their assigned matches. A **stable** match has no blocking pairs (pair-wise stability) and no agents who would prefer to remain unmatched. If either of these conditions fails to hold, there is an intuitive worry that participants will refuse to abide by the match or will sidestep it altogether. If a match participant knows that there is a sufficiently high probability that his assigned partner will renege on the clearinghouse assignment, then he might also endeavor to find a partner before the centralized match assigns him one. Such actions reinforce each other, leading to the well-known phenomenon of unraveling. The link between instability and unraveling has been demonstrated in the lab (Kagel and Roth 2000). We say that a matching mechanism **persists** if it continues to be used without being undermined by unraveling. Stability and persistence are intuitively linked. Even when participants can be nominally forced to go through the clearinghouse, matches can still be pre-arranged outside of the mechanism almost all “reasonable” matching mechanisms will, with high probability, match two participants if they both rank each other first.

Most matching schemes we see in the field can be classified as either **priority** mechanisms or **deferred acceptance (DA)** mechanisms². DA mechanisms are based on the Gale-Shapley algorithm. One such mechanism, ***M*-Proposing DA**, is implemented in

²Another important class of mechanisms, based on linear programming optimization, is not considered here.

the following way, denoting the members of the two sides of the market M s and W s (Gale and Shapley 1962):

Step 1

- All M s make an offer to their first-choice W .
- W s hold their favorite acceptable offer, rejecting all others.

Step t

- Rejected M s make an offer to their favorite acceptable W that has not yet rejected them.
- W s hold their favorite acceptable offer from this round and previous rounds, rejecting all others.

STOP in the first round where no new offers are made. All held offers become finalized matches.

Priority mechanisms instead use the preferences submitted by the participants to order the set of all possible match pairs. They then try to implement those match pairs in that order, skipping those that are not feasible due to previous implemented matches (Roth and Sotomayor 1990). For concreteness, consider the **M -Proposing Priority** mechanism implemented by the following algorithm³:

Step 1

- All M s make an offer to their first-choice W .
- W s are permanently matched to their favorite acceptable M who made an offer, rejecting all other offers.

Step t

- Rejected M s make an offer to their favorite acceptable W that has not yet rejected them.
- Matched W s reject all offers.

³This mechanism ranks potential match pairs in the order of M s preferences, with ties broken by W s preferences.

- Unmatched W s are permanently matched to their favorite acceptable M who made an offer.

STOP in the first round where no new offers are made.

The immediate difference between the M -Proposing DA and M -Proposing Priority algorithms is that under the priority mechanism, matches can be finalized in any step of the implementing algorithm, while under the DA mechanism, matches are only finalized in the last step of the algorithm. A less obvious difference is that DA mechanisms yield matches that are stable with respect to the reported preferences, while priority mechanisms generally do not. Since the literature looks for stable mechanisms, it has tended to look to DA.

The theoretical preference for DA also seems to be empirically justified. Unlike in the U.S., residency matches in the United Kingdom are organized at the regional level. Policy variation across regions then provides a nice natural experiment concerning which matching mechanisms persist. Roth (1991) looks at this natural experiment and finds that regions that adopt DA mechanisms tend to keep using them, while regions that adopt priority mechanisms tend to abandon them after a few years. This result seems to line up well with the intuition that participants will sidestep a matching procedure that does not yield a stable outcome. An interesting nuance of the U.K. study is that, due to the nationalization of healthcare in that country, doctors and hospitals had no choice but to go through the regional match clearinghouses. Instead of explicitly refusing to use the mechanisms, participants who wanted to match outside of them simply made pre-arrangements to rank each other first.

Our story thus far seems to nicely follow the arc from theory to the field. Lack of stability intuitively undermines persistence; in fact, we see evidence for this empirically. There is, however, a problem with this logic. While DA yields a stable allocation relative to the submitted preferences, generically, some participants can do better by deviating from truthful reporting. More specifically, although the proposing side finds truthful preference revelation to be a dominant strategy, the receiving side can fail to truthfully reveal in Bayes-Nash equilibrium (Roth and Rothblum 1999, Coles 2009). Under a priority mechanism it is also generally true that the equilibrium strategy for the receiving side involves deviation from truth-telling (Ehlers 2008). Furthermore, equilibrium predicts that neither DA nor priority mechanisms should yield matches that are stable relative to true preferences. Why then do we see an empirical divergence between the persistence of

priority and DA mechanisms?

Other papers provide possible answers to this question and are discussed in the next section. The answer explored in this paper is best prefaced by a look at the intuitive reasons that participants can profit by deviating from truth-telling. It can be shown that *M*-Proposing DA chooses the stable (relative to the submitted preferences) match that is most preferred by the *M*s and least preferred by the *W*s (Roth and Sotomayor 1990). We then might ask what happens when a *W* declares its least preferred stable match partner to be unacceptable. If the *W* has at least one other stable match partner available, the *W* will be matched to the least preferred of these partners and be better off than before. If the *M* declared unacceptable was the only stable match partner, however, the *W* will now be unmatched. In a complete information environment, which of these two things will happen is known, while in an incomplete information environment, a decision concerning whether to declare a potential match partner unacceptable is essentially a decision concerning whether to match to a low-ranked partner with some probability or to increase the likelihood of both obtaining a better match partner and remaining unmatched (Coles 2009). Another way to think about what happens when an agent declares a stable match partner unacceptable is that this action might launch a successful rejection chain a series of rejections and subsequent new proposals that eventually brings a better offer back to the original agent. At heart then, the decision to deviate from truth-telling under DA concerns a probabilistic evaluation of the set of stable matches, i.e. the likelihood of initiating a successful rejection chain.

The intuition for deviating from truth-telling under *M*-proposing Priority seems simpler. Since matches can be finalized in any step of the implementing algorithm, a *W* might worry about matching early on and missing out on a better offer that comes too late. By declaring a low-ranked match partner unacceptable, a *W* leaves itself open for better offers that come later. The downside to such a strategy is that the better offer may never come. The tradeoff that a *W* must navigate in deciding whether to deviate from truth-telling under *M*-Proposing Priority is between leaving oneself open to better offers that come later and potentially ending up unmatched.

Although it is far from clear, we feel that the rationale for deviating from truthful preference revelation is intuitively more transparent under priority mechanisms. If, in fact, match participants fail to deviate from truth-telling under DA mechanisms, even though they could gain by doing so, then we would expect an out-of-equilibrium stable matching. The difficulty of gaming DA could help to explain why it persists.

To confirm this intuition, we will look at strategies used by experimental participants under M -Proposing Priority and M -Proposing DA in a stylized, incomplete information environment. Specifically, we look at receiving side deviations from truth-telling and show that while participants learn to deviate from truth-telling under the priority mechanism, they fail to do so (or learn to do so much more slowly) under the DA mechanism. We interpret this to mean that out-of-equilibrium truth-telling is a possible contributor to the success of DA mechanisms in the field.

We continue below by outlining related previous literature and then describing the formal environment being studied. Next, we describe the details and design of the experiment itself and present our results. Finally, we conclude with a discussion of the data and practical application of the experiment

I. Related literature

The first two-sided matching experiments date to the early nineties (Sondak and Bazerman 1991, Harrison and McCabe 1996). An experiment that explicitly compares priority and DA mechanisms is described in Kagel and Roth (2000), although their paper focuses more on unraveling behavior than on strategic preference revelation. They do, however, provide a nice demonstration of the intuitive link between stability and persistence. Our experiment focuses on the receiving side of the market; however, several other experiments focus on strategies used by the proposing side (in the context of school choice) (Chen and Sönmez 2006, Pais and Pintér 2008, Featherstone and Niederle 2009).

Roth and Sotomayor (1990) provides much of the theory undergirding two-sided matching. Concerning incomplete information settings, Roth and Rothblum (1999) and Ehlers (2008) look at conditions on the reported preferences of other agents that make truncation strategies optimal. Results in Coles (2009) provide intuition on the mechanics of what happens when acceptable match partners are truncated.

There are also a few theory papers that provide other explanations for the persistence of DA mechanisms. Roth and Peranson (1999) takes preferences submitted to the NRMP and calculates the set of stable allocations. It finds that most participants in the match have the same match partner in all stable matches, which implies that either the set of stable matches is small or that match participants have gamed the system almost to the point of complete information equilibrium. Kojima and Pathak (2009) takes a more theoretical cut at the problem by looking at incentives to deviate from truth-telling under DA as the number of match participants goes to infinity. It finds that, for large enough

markets, truthful preference revelation becomes an e-equilibrium. Both of these papers closely relate to the idea of core convergence, which states that in large markets, the set of stable matches is small, leaving little leeway for participants to profitably deviate from truthful preference revelation under a stable mechanism like DA.

Core convergence as expressed in these papers relies on very large markets; however, we also see DA mechanisms being persistently used in smaller markets. For instance, in many of the markets examined in Roth (1991), the number of participants is less than 100. We think of the out-of-equilibrium truth-telling explanation proffered by this paper as a complement, rather than a replacement for, the previous explanations in the literature. The persistence of DA even in small markets implies that there is something else going on besides the core convergence explanations which have previously been put forward, and we seek to address this gap in understanding.

II. One-to-one, two-sided matching under incomplete information

We will present the relevant theory in terms of the two one-to-one matching mechanisms mentioned in the introduction; however, the distinctions we make between our two mechanisms hold in spirit whenever comparing any DA mechanism to any priority mechanism. We will use these two mechanisms, M -Proposing DA and M -Proposing Priority, in the lab, in conjunction with two different market structures. Under one structure, theory predicts that the receiving side will truncate its preferences under both mechanisms, while under the other structure, theory predicts truth-telling. Also, since our experimental design uses truth-telling robots for the proposing side, we demonstrate that this indeed the behavior we expect in equilibrium.

A. The formal setup

Formally, a **one-to-one, two-sided matching market under incomplete information** is a quadruple $(M, W, \mathcal{P}, \lambda)$, where M and W are sets of agents on the two sides of the market, \mathcal{P} is the set of all possible preference profiles for the agents, and λ is a measure over \mathcal{P} . An element of \mathcal{P} is a vector $(P_i)_{i \in M \cup W}$ of individual **preference profiles**. P_m for some $m \in M$ is an ordering over $W \cup \{\emptyset\}$, where \emptyset represents the outcome of being unmatched; P_w for some $w \in W$ is defined similarly. Hence, we can think of some W 's preference ordering as an $(|M| + 1)$ -vector whose elements are \emptyset and the members of M . A **matching** is a function $\mu : M \cup W \mapsto M \cup W \cup \{\emptyset\}$ such that for any $m \in M$ and $w \in W$, we have $\mu(m) \in W \cup \{\emptyset\}$, $\mu(w) \in M \cup \{\emptyset\}$, and $\mu(m) = w \Leftrightarrow \mu(w) = m$. A

revelation strategy for an agent i is a function $\sigma_i : \mathcal{P}_i \mapsto \mathcal{P}_i$ where \mathcal{P}_i denotes the projection of \mathcal{P} onto only agent i 's preference profile. Next, we define an important concept which we use to analyze the information structure of the matching market. Consider the following alteration of a preference profile $P \in \mathcal{P}$:

- For all $w' \in W \setminus \{w\}$, switch the positions of m and m' in $P_{w'}$.
- Let the preference of m change to the preference of m' and vice-versa.

Denote this new altered profile by $P_{-w}^{m \leftrightarrow m'}$. If for any $m, m' \in M$ and $w \in W$, $\lambda(P) = \lambda(P_w, P_{-w}^{m \leftrightarrow m'})$, then we say that the measure λ is **M -symmetric**. **W -symmetric** is defined similarly. We think of M -symmetry as a way of mathematically codifying the idea that W s have little information about preferences in the market. Essentially M -symmetry says that one M is just as likely as another to prefer any given W .

Briefly, let us introduce a few important classes of revelation strategies. We call a revelation strategy **truthful** if $\sigma_i(P_i) = P_i$. A revelation strategy is called a **truncation** if, for some k , $\sigma_i(P_i)(j) = P_i(j)$ for all $j \leq k$, and $\sigma_i(P_i)(k+1) = \emptyset$. Note that under this definition, a truthful strategy is an extreme point of the set of truncations.

Finally, since the environments we consider in this experiment are very symmetric, we want to be able to rule out equilibria we think of as artificial⁴. To do this, we call a strategy **label-independent** if its ranking of a given agent is solely a function of that agent's true ordinal ranking. Mathematically, this means that for some permutation π_i , $\sigma_i(P_i) = \pi_i(P_i)$. Note that truth-telling is the label-independent strategy associated with the identity permutation, and a truncation after the i^{th} ranked agent of a preference that declares exactly j ($> i+1$) agents acceptable is the label-independent strategy associated with a permutation that switches the $(i+1)^{\text{th}}$ entry and the $(j+1)^{\text{th}}$ entry, while mapping the k^{th} entry to the k^{th} entry for $k = i$.⁵

B. A concrete example: the uncorrelated market

Consider a small matching market that with $|M| = |W| = 5$. The true ordinal preferences of each participant are drawn independently from the uniform distribution over rank-order lists that find all members of the other side of the market acceptable. Cardinal

⁴For example, consider a market with $|M| = |W| = 5$. If all members of W find all members of M acceptable, but only rank m_1 acceptable, then we would have an equilibrium where m_1 ranks all members of W acceptable, and all members of $M \setminus \{m_1\}$ rank all members of W unacceptable.

⁵Note that there is a bit of redundancy here for any individually rational matching mechanism; for example, two different permutations produce $(m_1, m_2, m_3, \emptyset, m_4, m_5)$ and $(m_1, m_2, m_3, \emptyset, m_5, m_4)$.

payoffs are a decreasing function of ordinal rank only. We call this the **uncorrelated market**.

Without defining cardinal payoffs, we can already start to say things about the structure of equilibrium. To start, we look at behavior under M-Proposing DA (defined in the introduction) in markets where the true preferences are M -symmetric.

Theorem 1. *In any two-sided, one-to-one market where the true preferences are M -symmetric, under M -Proposing DA, any Bayes-Nash equilibrium in label-independent, weakly undominated strategies has any $m \in M$ truthfully revealing and any $w \in W$ playing a truncation.*

Intuitively, we know that members of M must truthfully reveal if they are constrained to weakly undominated strategies (Dubins and Freedman 1981). If this is true, then we know that the reported preferences will inherit M -symmetry from the true preferences, so long as the W s play label-independent strategies, since the $m \leftrightarrow m'$ operator and any permutation commute. Previous work has assumed M -symmetry about the reported preferences; we are able to give a sufficient condition on strategies that makes M -symmetric true preferences yield M -symmetric reported preferences. We then know that non-truncations are weakly dominated (Roth and Rothblum 1999). Intuitively, the members of W are truncating for the reasons discussed in the introduction and in Coles (2009). Once preferences are realized, M -Proposing DA matches a member of W to its least-preferred stable match partner (relative to the submitted preferences). By declaring an acceptable stable match partner unacceptable, a member of W either goes unmatched (if the stable partner in question was the only stable match partner of the member of W in question) or is matched to a more preferred stable match partner. Under M -symmetric incomplete information, all members of M are potential stable match partners with equal probability, so there is no reason to truncate a high-ranked member of M instead of a lower ranking one. Switching the reported ranks of two members of M can only hurt a member of W under any preference environment, which leaves us with truncation. Without understanding that M-Proposing DA picks the W -pessimal match relative to the submitted preferences, it is hard to understand how truncation could be profitable. Clearly, since the true preferences in the uncorrelated market are M -symmetric, the previous theorem has an immediate corollary.

Corollary (Theorem 1). *In the uncorrelated market, under M -Proposing DA, any Bayes-Nash equilibrium in label-independent, weakly undominated strategies has any $m \in M$ truthfully revealing and any $w \in W$ playing a truncation.*

Now, let's compare this to what we expect to see under M -Proposing Priority (again, defined in the introduction).

Theorem 2. *In the uncorrelated market, under M -Proposing Priority, any Bayes-Nash equilibrium in label-independent, weakly undominated strategies has any $m \in M$ reporting all members of W as acceptable and any $w \in W$ playing a truncation.*

So, we still have truncation by all members of W , although the intuition is quite different. Since M -Proposing Priority potentially finalizes matches after every step of its implementing algorithm, it is possible to match to a low ranked match partner in an early step and be forced to refuse an offer from a better ranked match partner in a later step. By dropping the low ranked partner from its list, a member of W avoids this problem but also opens itself to the possibility of never getting a better offer. At its heart, truncation under M -Proposing Priority entails understanding that there is an opportunity cost to being locked into a match too early in the implementing algorithm. Finally, note that the difference between Theorem 2 and the Corollary to Theorem 1 is that under M -Proposing Priority, the members of M can be playing any label-independent strategy that declares all members of W acceptable, while under M -Proposing DA, the members of M are truth-telling. Two theorems and a corollary can be stated that bring Theorem 2 and Corollary 1 into closer accord.

Theorem 3. *In a one-to-one, two-sided matching market with W -symmetric true preferences, if all members of W report the same number of M s acceptable, then under any Bayes-Nash equilibrium in label-independent, weakly undominated strategies, the members of M must be truth-telling.*

Corollary (Theorem 3). *In the uncorrelated market, if all members of W report the same number of M s acceptable, then under any Bayes-Nash equilibrium in label-independent, weakly undominated strategies, the members of M must be truth-telling.*

The corollary follows directly from the fact that true preferences are W -symmetric in the uncorrelated market. The big implication here is that if an M believes that the equilibrium played will be a symmetric truncation equilibrium, then truth-telling is the best response. This theorem extends work done in Roth and Rothblum (1999) and Ehlers (2008) to conditions that lead to truth-telling for the proposing side under a priority

mechanism⁶. Its proof method is essentially identical to that used in Roth and Rothblum (1999). In a broader sense, though, it turns out not to matter whether the M s truthfully reveal, as is expressed in the next theorem.

Theorem 4. *In the uncorrelated market, a member of M who plays a label-independent, weakly undominated strategy reports all complete preference lists that declare every member of W acceptable with equal probability. Thus, the interim payoff probability distribution for any strategy played by any W is independent of the strategies being played by the M s, so long as those M s are playing label-independent, weakly undominated strategies.*

The intuition here is that not ranking all W s is weakly dominated for the M s, and that running the true, uniform preference of each member of M through some permutation will simply yield the uniform distribution back again.

In the uncorrelated market, then, the unifying principle is that, under both mechanisms, we expect to see the members of W playing truncation strategies and that truth-telling can be rationalized for the M s. The intuitive reasons for this behavior, however, are quite different across the two mechanisms.

C. A contrasting example: the correlated environment

A natural question is then whether we always expect the members of W to truncate beyond truth-telling, or whether there are plausible environments where they might truthfully reveal. To address the question, we introduce a new environment by slightly altering the distribution of preferences in the uncorrelated market. Instead of drawing preferences independently for the members of M , draw only one preference and give it to all members of M . Continue to draw a new preference for each member of W . We call this the **correlated market**. Since it is still both M -symmetric and W -symmetric, Theorems 1 and 3 apply to the correlated market as well. A few more theorems demonstrate that we expect truth-telling for the members of W under both mechanisms.

Theorem 5. *In the correlated market, under M -Proposing DA, the unique Bayes-Nash equilibrium in label-independent, weakly undominated strategies entails truth-telling by all agents.*

⁶Roth and Rothblum (1999) and Ehlers (2008) focus on incentives for the receiving side. These papers also assume that reported preferences are M -symmetric instead of assuming that the true preferences are M -symmetric and backing out sufficient conditions to ensure that the reported preferences inherit M -symmetry as well.

Theorem 6. *In the correlated market, if all members of M have the same weakly undominated, label-independent revelation strategy, then all members of W best respond by truthfully revealing under M -Proposing Priority.*

Theorem 5 follows from realizing that if the members of M must truth-tell, then there is a unique stable match relative to the reported preferences. With a unique stable match, there is no reason to deviate from truth-telling. Theorem 6 follows from realizing that if all members of M play the same revelation strategy, then they will all submit the same reported preferences, which means that a member of W receives all offers in the same round of the M -Proposing Priority algorithm.

To conclude, we might worry that it is unrealistic that all members of M should use the same revelation strategy. The next theorem addresses this concern.

Theorem 7. *In the correlated environment, there exist cardinal payoffs such that there exists an equilibrium where all M s and W s truthfully reveal their preferences.*

Intuitively, we know this is so by thinking of a case where the payoff for getting a first-ranked W is more than 5 times the payment for getting a second-ranked W , which in turn is more than 4 times the payment for getting a third-ranked W , etc.

In the present papers experiment, which we will outline in the next section, the contrast between these two environments will give us the treatment effect we are looking for. The theorems also serve to justify the fact that our experiment fills the role of M s with truth-telling robots. Though theory predicts that W s should exhibit different strategic behavior across markets and matching mechanisms, we will find that, experimentally, this is not the case.

III. Experimental Setup

Table 1 shows the four treatments which comprise the experiments 2×2 design. As described in the previous section, the perfectly correlated M priorities in the truth telling treatments serve as a control for the truncation treatments, allowing us to measure truth telling rates against a benchmark. Each session consists of 40 rounds. In a given round, each participant plays the role of one of five W s potentially being matched to one of five M s. In all treatments, for each W , the computer independently draws a preference ordering over the five M s from the set of all possible orderings that list each M as acceptable. Each W then privately learns their preferences and submits a ranking of some, all or none of the M s.

TABLE 1—EXPERIMENTAL TREATMENTS.

	Truncation (uncorr. M and W prefs.)	Truth-telling (perfectly corr. M prefs., uncorr. W prefs.)
Priority	9 groups	8 groups
DA	9 groups	8 groups

When all 5 W s submit their preferences, the computer generates preferences for each of the M s in one of two ways. In the truncation treatments, M priorities are drawn independently and uniformly from the set of all possible orderings that list all W s as acceptable. In the truth telling treatments, a single set of preferences is drawn from the same uniform distribution, and this single ordering is applied to all five M s. The computer then generates a match outcome according to the rules of the appropriate mechanism Priority or DA using the generated M priorities and submitted W preferences as inputs, with M s acting as the proposing side in all cases. W s then learn their match outcome, as well as the outcomes of all other W s. W s gain points based on where their match partner appeared in their true preference list for that round, according to payoffs given in Table 2.

TABLE 2—PAYOFF TABLE.

Match	1 st choice	2 nd	3 rd	4 th	5 th	No match
Payoff	32 points	16	8	4	2	0

All treatments were run at Stanford University during the Spring of 2009. Participants were recruited for the experiment using existing dorm and course email lists and advertisements on facebook.com. Each session consisted of one or two groups of 5 participants. In sessions with two groups, groups were not mixed during the session, and participants were not informed which other participants were in their group.

At the start of each session, participants were read detailed instructions which included the overall experiment structure in terms of number of rounds and payment, the way in which W preferences and M preferences were to be generated in their treatment, and the way the mechanism used in their treatment generated a match from the preferences

reported by the M s and the W s.⁷ Participants worked through an exercise consisting of manually working through the steps of the appropriate mechanism for an example set of reported preferences. Actual play commenced only after all participants completed the exercise and indicated they understood the mechanism rules. Nothing was done to overtly suggest what the treatment variables were, i.e., there was no mention of matching mechanisms or preference distributions other than the ones in use in that particular treatment.

During the experimental session, participants could see their preferences for a given round on their computer screen and were reminded of payments for all possible match outcomes. They were then directed to click on radio buttons to rank each of the M s⁸. After all participants submitted rankings, a results screen showing the participants match for that round, their point accrual for that round and their total cumulative points would be displayed. At all times, a participant had the ability to see, for all prior rounds, the match outcomes for all participants, her own true preferences, and the rank list she submitted in that round.

At the conclusion of the 40th round, participants saw a closing screen which indicated their final payment amount. Participants were individually paid, asked to fill out a short optional exit survey, and dismissed.

We discuss the motivation behind some of our design decisions below.

A. *Payoff Choice*

One of the largest difficulties in simulating real world markets in a lab setting is the fact that cardinal preferences are not usually observable in the field. Thus, the magnitude of differences in payoffs assigned to different ordinal outcomes in the lab can be difficult to calibrate effectively. When designing payoffs for our treatments, our goal was to find a payoff scheme which provided behavioral incentives that were as comparable as possible between treatments, thus preventing a difference in the magnitude of incentives from driving our observed behavior. Of course, the precise incentive scheme each individual participant truly faces depends on the behavior of her peers, rendering precise comparability impossible *ex ante*. However, as shown in Figures 1 and 2, based on the actual

⁷In the lab, we provide a specific context in the hopes of making understanding easier for participants. Proposing side agents (referred to here as M s) are referred to as “Schools” and the agents receiving offers (here, W s) are referred to as “Students.”

⁸We did this so that participants would have to click the same number of times regardless of what preference they wished to report. If declaring all M s unacceptable were too easy, some participants might choose to do this in order to save time and effort.

behavior which occurred, individual participants did in fact face incentive magnitudes which were roughly comparable across treatments. Note that a simple reinforcement learning model would predict that the slopes of the curves are much more important than the levels.

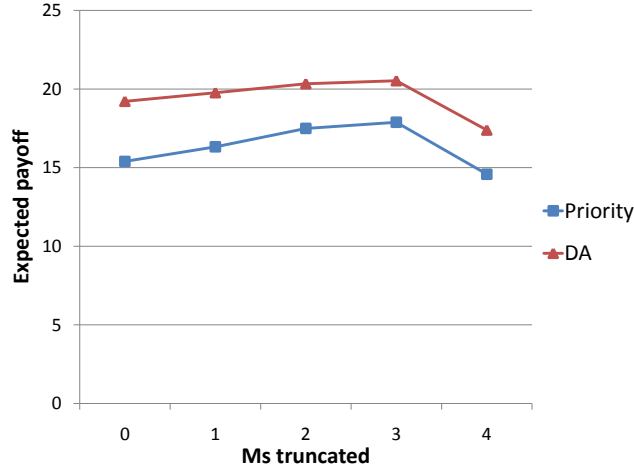


FIGURE 1. EXPECTED PAYOFF FOR TRUNCATION STRATEGIES IN THE TRUNCATION TREATMENTS.

B. Use of Repetition

In reality, most individuals participate in a matching process such as the one we hope to emulate only once (or perhaps a handful of times in some applications). However, we nevertheless argue that repeating the match a large number of times with the same group of participants improves the applicability of the result to real markets. One primary reason for this common experimental practice is that we can adequately mimic neither the stakes participants in real matching markets face nor the efforts those participants can undertake to improve their outcome (such as observing prior match outcomes or talking to prior participants). Furthermore, we cannot realistically allow experimental participants as much time to consider their prospects as they have in real markets. Instead, by having them participate in repeated trials, we allow for participants to learn about the environment and possibly alter their strategy as they progress. One could argue that this

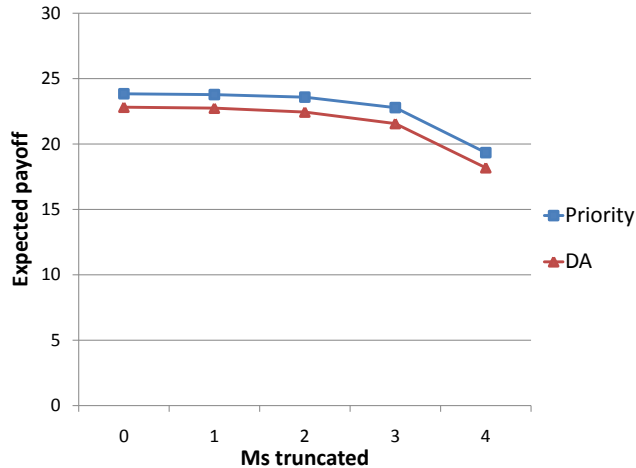


FIGURE 2. EXPECTED PAYOFF FOR TRUNCATION STRATEGIES IN THE TRUTH-TELLING TREATMENTS.

makes participants better able to understand the mechanism and behave strategically than in real world markets; however, if this is the case, and, as we anticipate, subjects nonetheless have difficulty successfully manipulating effectively, we can be confident that manipulation is even more difficult in the field.

C. Use of Automated Ms

In real life two sided matching markets, the proposing side’s report to the matching mechanism is not automatic as it is in our experiment. In priority mechanisms, proposers do not necessarily have dominant strategy incentives to report their preferences truthfully (although as discussed in the theory section, this behavior can occur in equilibrium). In DA mechanisms, truthful reporting is a dominant strategy, but there is some evidence that proposing side agents may not propose to all agents in order in a sequential matching market, even when they have dominant strategy incentives to do so (Echenique, Wilson and Yariv 2009).

We nevertheless use automated proposers playing fixed strategies for two primary reasons. First, as discussed previously, in both the Priority and DA mechanisms, proposers do have dominant strategy incentives to list all acceptable match partners. Thus, even if proposers do not list partners in truthful order in real markets, we can reasonably expect

them to list all agents. In symmetric environments such as the ones we use, responders’ strategies would not be affected by any reordering of true preferences a proposer might choose to report (so long as proposers’ strategies are label independent). The second reason to use automated Ms is that doing so reduces the complexity of the decision the subjects face. If, as we anticipate, subjects have difficulty learning to successfully manipulate the mechanism in this simplified environment, we are confident they will also have trouble in the more complicated real world markets of interest.

IV. Results

A. Overall Truthtelling Rates

We are most interested in the rate of truth telling across the four primary treatments. This value is significantly higher in the DA truncation treatment than in the Priority truncation treatment; however, for the two truth telling treatments, the differences between the DA and Priority treatments are not statistically significant. Furthermore, the rate difference between the two DA treatments is not statistically significant, while the difference between the two Priority treatments is highly significant.

TABLE 3—TRUTH-TELLING RATES.

	DA		Priority
Truth-telling	66.0%	\leftrightarrow (0.382)	58.4%
	\downarrow (0.200)		\downarrow (0.001)**
Truncation	56.7%	\leftrightarrow (0.000)**	25.3%

Numbers in parentheses are p-values from two-tailed Mann-Whitney tests with session-level averages as the units of observation.

When we restrict attention to the last ten periods, focusing on the behavior of subjects when they are more experienced, we find qualitatively similar effects. Statistically, there is a mildly significant difference between the two DA treatments, as well as the high significance between the Priority treatments and the truncation treatments seen in the data for all 40 periods.

Note that for DA, truth telling rates are slightly lower in the last 10 periods (2% lower) in the truncation treatment, but also 4% higher in the truth telling treatment. Thus, the significance of the difference in truth telling rates between the two groups is in some sense

TABLE 4—TRUTH-TELLING RATES.

	DA		Priority
Truth-telling	70.3%	\leftrightarrow (0.328)	60.8%
	\downarrow (0.046)*		\downarrow (0.001)**
Truncation	54.7%	\leftrightarrow (0.000)**	19.3%

Numbers in parentheses are p-values from two-tailed Mann-Whitney tests with session-level averages as the units of observation.

as much due to participants in the truth telling treatment learning to tell the truth as it is those in the truncation treatment learning to truncate. In sum, we only see a significant deviation from the benchmark truth-telling rate under the Priority truncation treatment. Under DA, participants do not respond to the truncation treatment by deviating from truth-telling.

Of course, failure to tell the truth is not synonymous with truncation, and although truncation weakly dominates other non-truth telling strategies, we do observe some portion of suspects employing “switching” or “dropping” strategies in some rounds.⁹ Frequency of this behavior, however, is not significantly different between any of the treatments.

TABLE 5—TRUNCATION (INCLUDING TRUTH-TELLING) RATES.

	DA		Priority
Truth-telling	16.3%	\leftrightarrow (0.234)	11.1%
	\downarrow (0.673)		\downarrow (0.200)
Truncation	14.3%	\leftrightarrow (0.489)	17.9%

Numbers in parentheses are p-values from two-tailed Mann-Whitney tests with session-level averages as the units of observation.

B. Blocking Pairs and Overall Match Stability

For practical market design, we may be primarily concerned not with the rate at which participants tell the truth, but rather with how successfully a mechanism generates

⁹The characterization of this other behavior as “non-truthful, non-truncation” is redundant, as truth-telling is one extreme of the set of truncation strategies for participants. We nevertheless use the terminology to ensure clarity.

desirable (i.e., stable) match outcomes. One measure of this is the number of blocking pairs present in any given assignment. Since the outcome is never 100% stable in any treatment at any time, the number of blocking pairs is one measure of the degree of stability of a match outcome: a mechanism which generates an outcome that is stable for most participants may still work well enough to be persistent.

TABLE 6—NUMBER OF BLOCKING PAIRS PER PERIOD.

	DA		Priority
Truth-telling	0.47	\leftrightarrow (0.574)	0.59
	\downarrow (0.815)		\downarrow (0.000)**
Truncation	0.49	\leftrightarrow (0.000)**	1.87

Numbers in parentheses are p-values from two-tailed Mann-Whitney tests with session-level averages as the units of observation.

Blocking pairs were found to occur significantly more often in the Priority truncation treatment than in the DA truncation treatment or the Priority truth telling treatment. The two DA treatments were not significantly different in blocking pair frequency; nor were the two truth telling treatments.

Note that the same M or W can be involved in multiple blocking pairs if there is more than one attainable match partner that they prefer to their actual match partner. However, we do not observe any interesting asymmetries in terms of which unique agents are involved in multiple blocking pairs: the number of unique Ms involved in blocking pairs is not significantly different than the number of unique Ws for any treatment, and the between-treatment differences are similar qualitatively and in terms of statistical significance when the number of unique Ms and Ws in blocking pairs are considered separately.

C. Best Response Frequencies

Truth telling rates establish how apt participants are to manipulate, and low non-truth, non-truncation rates establish that these manipulations are, for the most part, some sort of truncation. However, participants who truncate are not automatically maximizing their expected payoff: they may be truncating too much or too little. For the set of payoffs used in the experiment, we can find an equilibrium where all agents truncate symmetrically; however, as out-of-equilibrium strategies may be a best response to other

out of equilibrium strategies, we would not necessarily expect even the most sophisticated participants to truncate as if in equilibrium. We instead look at the ability of participants to find the strategy which is a best response to the environment in which they find themselves. If a significant proportion of subjects are able to achieve this in a significant portion of sessions for a certain mechanism, we might reach different conclusions as to the sophistication of subjects than we would looking strictly at truth telling rates (or looking at the frequency of play consistent with theoretical equilibrium, for that matter). Also, we might wonder if there is a great deal of heterogeneity in participant sophistication, or if all participants reported optimal truncations about the same fraction of the time.

However, simply comparing subjects' behavior in an individual round to the optimal behavior possible in that period ex post fails to capture the uncertainty which is inherent in truncation strategies in the truncation treatments, for example, it is optimal ex ante to truncate in each period, even though it may be suboptimal ex post. Thus, we consider the participant to be playing optimally in their "environment" if they play the truncation strategy which generates the highest expected utility across all the rounds they played, given the actual behavior of other participants and generated proposer preferences. We then look at the frequency with which each individual subject behaved optimally; the results are summarized in Figure 3.

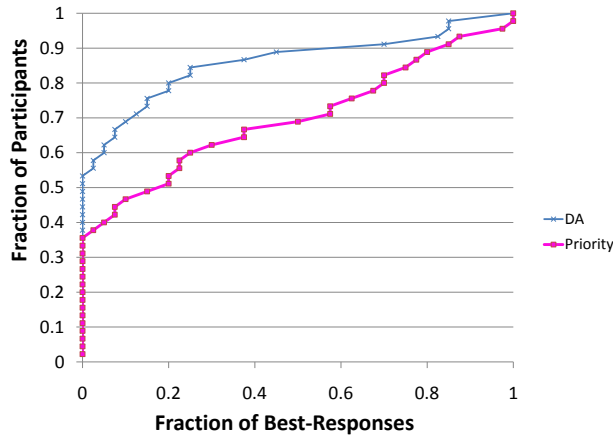


FIGURE 3. BEST RESPONSE FREQUENCY CDF FOR TRUNCATION TREATMENTS, ALL PERIODS.

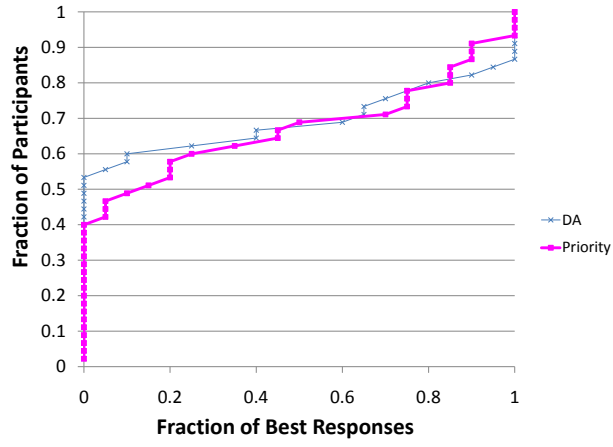


FIGURE 4. BEST RESPONSE FREQUENCY CDF FOR TRUNCATION TREATMENTS, LAST 20 PERIODS.

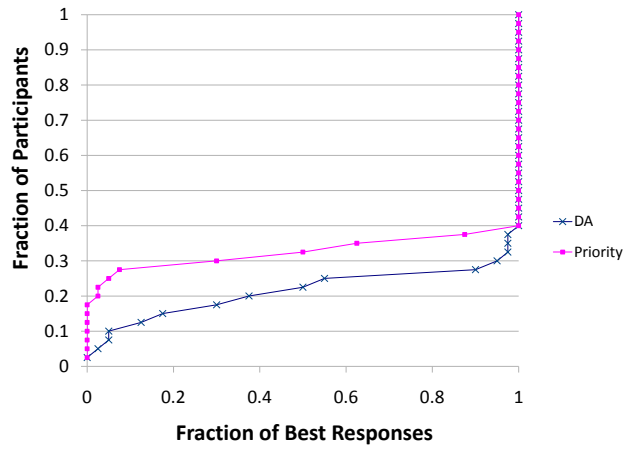


FIGURE 5. BEST RESPONSE FREQUENCY CDF FOR TRUTH-TELLING TREATMENTS, ALL PERIODS.

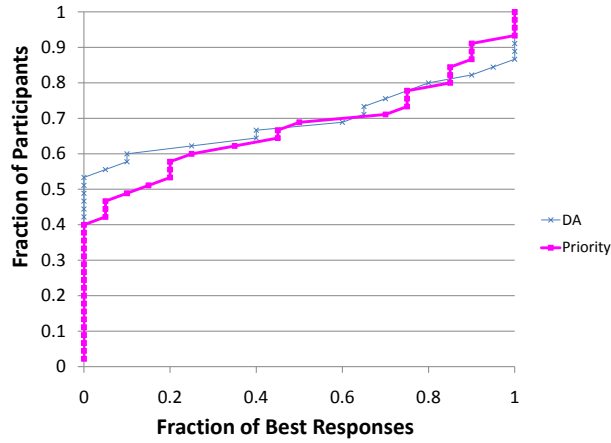


FIGURE 6. BEST RESPONSE FREQUENCY CDF FOR TRUTH-TELLING TREATMENTS, LAST 20 PERIODS.

Figure 3 indicates the proportion of participants playing an overall best response at most the indicated proportion of the time for the truncation treatments. For example, approximately 36% of Priority participants never played a best response (compared with about 52% for DA), and 50% of participants played a best response no more than 20% of the time (compared with around 75% for DA). Note that the Priority treatment first order stochastically dominates the DA treatment: for any level of frequency of best response play we consider, more participants best respond at least that frequently in the Priority treatment than in the DA treatment. However, this gap closes when only the last 20 periods are considered, as seen in Figure 4. Note that this closing of the gap simply implies that under both mechanisms, participants have converged to similarly bad distributions of sub-optimal play.

In the Truth Telling treatments, truthful reporting is always the unique best response, and much as there was no significant difference in the overall truth telling rates between DA and Priority in these treatments, there is no noticeable difference in the frequency with which individual subjects play this best response, either in the whole sample or restricting attention to the last 20 periods.

V. Results Discussion (working title)

The major conclusion our data suggests is that a substantially greater number of participants reap the benefits of truncation in the Priority mechanism compared with the DA mechanism. Rates of straightforward play are substantially lower under Priority, but only for the treatments where truncation was a rational response. This suggests that participants’ ability to comprehend the Priority mechanism exceeds their ability to comprehend the DA mechanism, rather than there being something else inherent in the description or operation of the mechanism which leads to truncation for some a-rational reason. This notion is further supported by the lack of a significant difference in the level of “Truth Mod Trunc” behavior: neither mechanism drives participants to behave in “crazy” fashion significantly more than the other.

In the last ten periods, we do see a significant difference between truth telling rates in the DA treatments; however, this is due more to increased truth telling in the truth telling treatment than to an increase in truncation in the truncation treatment. Furthermore, the drop in truth telling is substantially smaller in the DA truncation treatment (2%) than in the Priority treatment (6%) — despite the fact that a much larger portion of the observations in Priority were already truncations. This suggests that while some subjects may be learning to truncate with experience in DA, any learning effect seems to be as strong or stronger under the Priority mechanism.

VI. Discussion

We begin by briefly discussing the external validity of the experiment. Our general idea was to make it relatively easy for experimental participants to discover profitable deviations from truth-telling (but not so easy as to make the experiment completely removed from reality). Since the real world is more complicated than our lab setup, we think it quite plausible that if participants fail to discover profitable deviation from truth-telling in the lab, then they would also fail to discover it in the field. Several design features of the experiment make it simple relative to the real world. First, at the beginning of every session, we gave participants twenty minutes of training on how the relevant mechanism and environment work. Second, by having proposer roles be played by truth-telling robots, we decreased the outcome noise that might otherwise hamper learning for the receiving side. Third, we allowed participants to repeat the game many times with feedback. In the field, at best, hospitals get once-a-year, partial feedback about the match, while doctors mainly get feedback through their interactions

with other doctors who have previously been through the match.

We also chose a small environment for our experiment. Intuitively, a very large market seems like it should be more difficult to understand; however, under some assumptions, asymptotic results tell us that truth-telling in large markets can be very close to equilibrium play. Our experiment works with either side of this coin. Even in environments where we might expect asymptotic truth-telling, we are stacking the cards against our result by allowing deviation from truth-telling to yield significant gains. If we believe that smaller markets should be intuitively simpler, then we can continue to hold with the “from minor to major” argument of the previous paragraph. Although we have already stated this, we feel it bears repeating that our design does nothing to refute the importance of asymptotic results. We simply show that, in our chosen environment, where deviation from truth-telling is still profitable, participants fail to realize potential gains.

Along the same lines, we should also mention that our experiment only looks at certain types of deviations from truth-telling in very low information environments. We believe it plausible that these low information truncations are present in some sector of almost any real-world matching market, although there are certainly other environments we could have looked at. Non-truncation deviations from truth-telling can be found in the literature (e.g. Ex. 3 in Roth and Rothblum (1999)), although it seems difficult to come up with examples of this sort that do not require a high level of information to be known by match participants. Also, we could have changed the structure of our experimental market to make truncation even easier. One obvious way to do this would have been to put a marked discontinuity in our payoff structure, i.e. have payoff as a function of rank go something like 100, 99, 98, 21, 20, 0. Another, perhaps more realistic, way that could perhaps make truncation more easily realized would have been to use a tiered uncorrelated environment, where, for instance, all W s agree that m_1 , m_2 , and m_3 are preferred to m_4 and m_5 , but that within these two tiers, preferences are still uncorrelated. What sorts of manipulations we see in this sort of environment is an open question; regardless, we can view our experiment as an examination of within-tier truncations. Moreover, when we look at successful real-world instances of DA, there is often some sort of pre-match interview process which could serve to create within tier matching before the centralized procedure even begins. In this sort of world, again, we feel that the low-information environments that lead to truncation deviations from truth-telling are particularly pertinent.

We feel we have chosen a simple experimental setup which bears marked similarities

to real world markets. Since DA is known to persist, even in some small markets, we think of our out-of-equilibrium truth-telling finding as an additional part of story of DAs persistence. We make no claim that it is the entire story, nor can we claim that our experiment does much to explain behavior in market sectors about which agents hold a high level of information concerning each others idiosyncratic preferences.

In addition to the external validity of our experiment, we also think an experiment such as ours feeds into the broader concerns of market design. Whenever a matching mechanism is strategy-proof, it is straightforward for designers to predict agent behavior in the field, since both focality and optimality push towards truth-telling. Sometimes though, strategy-proofness is either not desired or cannot be achieved due to other design goals. Consider the job of a market designer who has been tasked with creating a two-sided matching mechanism that persists. We can view the current paper as an experiment that would help inform our theoretical designer. Persistence can be intuitively linked to ex post stability, so DA is a natural candidate. Unfortunately, under DA, truth-telling is generically not an equilibrium. Theory provides a set of strategies which could outperform truthful preference revelation: the question is then whether our designer should expect market participants to use these deviations from truth-telling, which is the clear focal strategy. If agents use these profitable deviations from truth-telling, then DA will not yield an ex post stable outcome, but if they dont, then it will. To determine which is the more likely outcome, the present lab experiment becomes very informative.

In demonstrating that agents learn to play some deviations from truth-telling, but not others, we bring up the idea that not all equilibria are equal in their predictive power. Depending on the mechanism and environment, agents are sometimes very close to equilibrium play and sometimes not. Some intuitive factors that seem like they should be important for whether a theoretical equilibrium will be realized in the field are focality of truth-telling, obviousness that deviation from truth-telling will be profitable, difficulty of finding the optimal deviation, and profitability of deviation. Unfortunately, although these factors may guide us intuitively, there is no formal theory for how they might trade off in determining the accuracy of an equilibrium prediction; in fact, most of them are difficult even to define. This is where lab experiments can prove most useful for design. In short, although the main contribution of this experiment is to show how out-of-equilibrium truth-telling could contribute to the ex post stability of DA in the field, we also feel that the experiment is the sort of inquiry that should be used in practical market design.

VII. Conclusion

While DA is stable with respect to reported preferences, equilibrium does not generically predict a stable ex post allocation under incomplete information. In spite of this, DA persists in the field — a fact that is intuitively linked to stability. In the lab, we show that in an incomplete information environment, participants fail to significantly deviate from truth-telling, even though they could gain by doing so. To demonstrate that this is not just a lab effect, we look at behavior in the same preference environments under a priority mechanism and find that we do see significant deviation from truth-telling. These findings suggest that DA could be close to ex post stable in the field, in spite of equilibrium predictions to the contrary. The findings also help to explain why DA succeeds even in smaller markets, where asymptotic explanations have less bite, or in markets where the assumptions needed to attain truth-telling as an ε -equilibrium do not hold. Even in larger markets where these assumptions do hold, the tendency of the receiving side towards out-of-equilibrium truth-telling might help to neutralize what small gains to truth-telling remain. Finally, we feel that the present experiment provides an example for how market designers could use experiments to understand when they are bound by equilibrium predictions, and when they can expect focal truth-telling to win out over optimality.

REFERENCES

- APPIC. “2009 APPIC Match Statistics.” appic.org. 2009. http://www.appic.org/match/5.2.2.1.11_match_about_statistics_general.2009.html (accessed February 18, 2010).
- Chen, Yan, and Tayfun Sönmez. “School choice: an experimental study.” *Journal of Economic Theory*, 2006: 202-231.
- Coles, Peter. “Optimal truncation in matching markets.” mimeo, 2009.
- Dubins, L.E., and A. Freedman. “Machiavelli and the Gale-Shapley algorithm.” *The American Mathematical Monthly*, 1981: 485-494.
- Echenique, Federico, Alistair Wilson, and Leeat Yariv. “Clearinghouses for two-sided matching: an experimental study.” mimeo. 2009.
- Ehlers, Lars. “Truncation strategies in matching markets.” *Mathematics of Operations Research*, 2008: 327-335.
- Ergin, Haluk, and Tayfun Sönmez. “Games of school choice under the Boston mechanism.” *Journal of Public Economics*, 2006: 215-237.

- Featherstone, Clayton, and Muriel Niederle. "School choice mechanisms under incomplete information: an experimental study." mimeo, 2009.
- Gale, D., and L.S. Shapley. "College admissions and the stability of marriage." *The American Mathematical Monthly*, 1962: 9-15.
- Harrison, Glenn, and Kevin McCabe. *Stability and Preference Distortion in Resource Matching: An Experimental Study of the Marriage Market*. Vol. 8, in *Research in Experimental Economics*, edited by R.M. Isaac. Greenwich, CT: JAI Press, 1996.
- Kagel, John H., and Alvin E. Roth. "The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment." *The Quarterly Journal of Economics*, 2000.
- Kojima, Fuhito, and Parag Pathak. "Incentives and Stability in Large Two-Sided Matching Markets." *American Economic Review*, 2009: 608-627.
- NRMP. "Results and Data: 2009 Main Residency Match." [nrmp.org](http://www.nrmp.org). 2009. <http://www.nrmp.org/data/resultsanddata2009.pdf> (accessed February 18, 2010).
- NYC-DOE. "Statistical summaries: register by grade." schools.nyc.gov. 2009. <http://schools.nyc.gov/AboutUs/data/stats/Register/CurrentRegisterbyGrade/default.htm> (accessed February 18, 2010).
- Pais, Joana, and Ágnes Pintér. "School choice and information: an experimental study on matching mechanisms." *Games and Economic Behavior* 64, no. 1 (2008): 303-328.
- Roth, Alvin E. "A Natural Experiment in the Organization of Entry-Level Labor Markets: Regional Markets." *American Economic Review*, 1991: 415-440.
- Roth, Alvin E., and Elliott Peranson. "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design." *American Economic Review*, 1999: 748-780.
- Roth, Alvin E., and Marilda Sotomayor. "Two-Sided Matching: A Study in Game Theoretic Modeling and Analysis." Cambridge University Press, 1990.
- Roth, Alvin E., and Uriel G. Rothblum. "Truncation strategies in matching markets — in search of advice for participants." *Econometrica*, 1999: 21-43.
- Sondak, Harris, and Max H. Bazerman. "Power Balance and the Rationality of Outcomes in Matching Markets." *Organizational Behavior and Human Decision Processes*, 1991: 1-23.

Proofs to Theorems

Lemma 1. *Any strategy except for truth-telling is weakly dominated by truth-telling for any $m \in M$ under M -Proposing DA.*

PROOF:

Dubins and Freedman (1981).

Lemma 2. *In a one-to-one, two-sided matching market $(M, W, \mathcal{P}, \lambda)$, for any $w \in W$, if λ is M -symmetric, all agents in M truthfully reveal, and all agents in $W \setminus \{w\}$ play label-independent strategies, then under M -Proposing DA, any strategy where w does not submit a truncation of the true preference is weakly dominated.*

PROOF:

Fix $m, m' \in M$ and $w' \in W$. Let the label-independent strategy of w' be given by $\pi_{w'}$. Given the assumed strategies, we can map from reported preferences to true preferences in order to determine the probability of a certain profile of reported preferences. Note that we make this mapping one-to-one by considering reports like $(m_1, m_2, m_3, \emptyset, m_4, m_5)$ and $(m_1, m_2, m_3, \emptyset, m_5, m_4)$ to be distinct. Let the mapping from reported preferences to true preferences be given by $\text{True}(\cdot)$. Note that since $\pi_{w'}$ was defined as the label-independent strategy used by w' , $\pi_{w'}^{-1}(\widetilde{P}_{w'})$ must be in the support of the true preference distribution of w' . Hence the probability that the reported preference is \widetilde{P}_{-w} must be equal to the probability that the true preference is $\text{True}(\widetilde{P}_{-w})$.

$$\begin{aligned}
 \widetilde{P}_{-w} &= \left(\widetilde{P}_{m_1}, \widetilde{P}_{m_2}, \left(\widetilde{P}_m \right)_{m \in M \setminus \{m_1, m_2\}}, \left(\widetilde{P}_{w'} \right)_{w' \in W \setminus \{w\}} \right) \\
 \text{True}(\widetilde{P}_{-w}) &= \left(\widetilde{P}_{m_1}, \widetilde{P}_{m_2}, \left(\widetilde{P}_m \right)_{m \in M \setminus \{m_1, m_2\}}, \left(\pi_{w'}^{-1}(\widetilde{P}_{w'}) \right)_{w' \in W \setminus \{w\}} \right)
 \end{aligned}
 \tag{1}$$

Now, we do the same for \widetilde{P}_{-w} with the appropriate symmetry operator applied:

$$\begin{aligned}
(2) \quad \widetilde{P}_{-w}^{m_1 \leftrightarrow m_2} &= \left(\widetilde{P}_{m_2}, \widetilde{P}_{m_1}, \left(\widetilde{P}_m \right)_{m \in M \setminus \{m_1, m_2\}}, \left(\widetilde{P}_{w'}^{m_1 \leftrightarrow m_2} \right)_{w' \in W \setminus \{w\}} \right) \\
\text{True} \left(\widetilde{P}_{-w}^{m_1 \leftrightarrow m_2} \right) &= \left(\widetilde{P}_{m_2}, \widetilde{P}_{m_1}, \left(\widetilde{P}_m \right)_{m \in M \setminus \{m_1, m_2\}}, \right. \\
&\quad \left. \left(\pi_{w'}^{-1} \left(\widetilde{P}_{w'}^{m_1 \leftrightarrow m_2} \right) \right)_{w' \in W \setminus \{w\}} \right) \\
&= \left(\widetilde{P}_{m_2}, \widetilde{P}_{m_1}, \left(\widetilde{P}_m \right)_{m \in M \setminus \{m_1, m_2\}}, \right. \\
&\quad \left. \left(\pi_{w'}^{-1} \left(\widetilde{P}_{w'}^{m_1 \leftrightarrow m_2} \right) \right)_{w' \in W \setminus \{w\}} \right)
\end{aligned}$$

The second expression for $\text{True}(\widetilde{P}_{-w}^{m_1 \leftrightarrow m_2})$ comes from the fact that the $m_1 \leftrightarrow m_2$ interchange operator commutes with permutations. We can then make the claim

$$(3) \quad \Pr \left\{ \widetilde{P}_{-w} \right\} = \Pr \left\{ \text{True} \left(\widetilde{P}_{-w} \right) \right\} = \Pr \left\{ \text{True} \left(\widetilde{P}_{-w}^{m_1 \leftrightarrow m_2} \right) \right\} = \Pr \left\{ \widetilde{P}_{-w}^{m_1 \leftrightarrow m_2} \right\}$$

where the inner equation comes from the M -symmetry of the true preferences, and the outer equations come from the fact that we have established a one-to-one mapping between reported and true preferences. So, we have shown that, under our assumptions about strategies, the reported preferences are $\{m_1, m_2\}$ -symmetric for w , which means that, since there was nothing special about m_1 , m_2 , or w , we have also showed that preferences are M -symmetric for all $w \in W$. By Theorem 2 of Roth and Rothblum (1999), truncations stochastically dominate non-truncations for any $w \in W$. Also note that the theorem handles the case of mixed strategies. Let a state of the world be defined by the pure strategies that are submitted by each agent. Conditional on a state of the world, the proof proceeds as above. Taking the expectation over states of the world establishes the theorem for mixed strategies.

Theorem 1. *In any two-sided, one-to-one market where the true preferences are M -symmetric, under M -Proposing DA, any Bayes-Nash equilibrium in label-independent, weakly undominated strategies has any $m \in M$ truthfully revealing and any $w \in W$ playing a truncation.*

PROOF:

The weakly undominated requirement means that each $m \in M$ must truthfully reveal,

by Lemma 1. Truthful preference revelation and each $w \in W$ playing a label-independent strategy combine through Lemma 2 to tell us that at equilibrium, each $w \in W$ must all be playing a truncation strategy.

Lemma 3. *Under M -Proposing Priority, any report for any $m \in M$ that does not list all truly acceptable $w \in W$ as acceptable is weakly dominated by one that does.*

PROOF:

Consider an arbitrary $m \in M$ submitting a list L of size N which excludes at least one acceptable $w' \in W$. Now consider L' , a list identical to L for the first N entries with w' listed in the $(N + 1)^{st}$ position and no entries thereafter.

Under both Priority and DA, repeat offers are disallowed for the M s and each M make offers to W s in the order given on their submitted preference list. Thus, any set of submissions for other agents resulting in m being matched to a given w when m^* submits L will also result in m being matched to that w when m submits L' , since neither algorithm conditions on M offers which have not yet been made and both would only generate an offer to w' when L' is submitted if all the offers in L have already been rejected. So L' never generates a worse outcome for m than L .

However, consider a set of submissions such that no member of W listed in L ranks m as acceptable, and the submitted preference list of w' lists only m as acceptable. In this case, M -Proposing Priority will match m and w' when L' is submitted and will match m to no one when L is submitted. Since w' is acceptable to m by construction, m achieves a better result in this case by submitting L' .

Lemma 4. *In the uncorrelated market, if all $m \in M$ play label-independent, weakly undominated strategies, and all agents in $W \setminus \{w\}$ play label-independent strategies, then under M -Proposing Priority, w not truncating is stochastically dominated by w truncating.*

PROOF:

By Lemma 3 combined with the requirement that weakly undominated strategies, all members of M must report all agents in W acceptable. Fix $m_1, m_2 \in M$ and $w' \in W$. Let the label-independent strategy of i be given by π_i . Given the assumed strategies, we can map from reported preferences to true preferences. Call this mapping $\text{True}(\cdot)$. Note that we make this mapping one-to-one by considering reports like $(m_1, m_2, m_3, \emptyset, m_4, m_5)$ and $(m_1, m_2, m_3, \emptyset, m_5, m_4)$ to be distinct. Also, note that since the p_i s were defined as the label-independent strategy used by agent i , $\pi_i^{-1}(\tilde{P}_i)$ must be in the support of the

true preference distribution of i . Hence the probability that the reported preference is \widetilde{P}_{-w} must be equal to the probability that the true preference is $\text{True}(\widetilde{P}_{-w})$. Then,

$$(4) \quad \begin{aligned} \widetilde{P}_{-w} &= \left(\widetilde{P}_{m_1}, \widetilde{P}_{m_2}, \left(\widetilde{P}_m \right)_{m \in M \setminus \{m_1, m_2\}}, \left(\widetilde{P}_{w'} \right)_{w' \in W \setminus \{w\}} \right) \\ \text{True}(\widetilde{P}_{-w}) &= \left(\pi_{m_1}^{-1}(\widetilde{P}_{m_1}), \pi_{m_2}^{-1}(\widetilde{P}_{m_2}), \left(\pi_m^{-1}(\widetilde{P}_m) \right)_{m \in M \setminus \{m_1, m_2\}}, \right. \\ &\quad \left. \left(\pi_{w'}^{-1}(\widetilde{P}_{w'}) \right)_{w' \in W \setminus \{w\}} \right) \end{aligned}$$

Now, we do the same for \widetilde{P}_{-w} with the appropriate symmetry operator applied:

$$(5) \quad \begin{aligned} \widetilde{P}_{-w}^{m_1 \leftrightarrow m_2} &= \left(\widetilde{P}_{m_2}, \widetilde{P}_{m_1}, \left(\widetilde{P}_m \right)_{m \in M \setminus \{m_1, m_2\}}, \left(\widetilde{P}_{w'}^{m_1 \leftrightarrow m_2} \right)_{w' \in W \setminus \{w\}} \right) \\ \text{True}(\widetilde{P}_{-w}^{m_1 \leftrightarrow m_2}) &= \left(\pi_{m_1}^{-1}(\widetilde{P}_{m_2}), \pi_{m_2}^{-1}(\widetilde{P}_{m_1}), \left(\pi_m^{-1}(\widetilde{P}_m) \right)_{m \in M \setminus \{m_1, m_2\}}, \right. \\ &\quad \left. \left(\pi_{w'}^{-1}(\widetilde{P}_{w'}^{m_1 \leftrightarrow m_2}) \right)_{w' \in W \setminus \{w\}} \right) \end{aligned}$$

The second equality can also be expressed in terms of the elements of $\text{True}(\widetilde{P}_{-w})$:

$$(6) \quad \begin{aligned} \text{True}(\widetilde{P}_{-w}^{m_1 \leftrightarrow m_2}) &= \left(\pi_{m_1}^{-1} \circ \pi_{m_2} \left(\pi_{m_1}^{-1}(\widetilde{P}_{m_2}) \right), \pi_{m_2}^{-1} \circ \pi_{m_1} \left(\pi_{m_2}^{-1}(\widetilde{P}_{m_1}) \right), \right. \\ &\quad \left. \left(\pi_m^{-1}(\widetilde{P}_m) \right)_{m \in M \setminus \{m_1, m_2\}}, \left(\pi_{w'}^{-1}(\widetilde{P}_{w'}^{m_1 \leftrightarrow m_2}) \right)_{w' \in W \setminus \{w\}} \right) \end{aligned}$$

Note that any $m \in M$ must rank any acceptable $w \in W$ as acceptable since we are not allowing agents to play weakly dominated strategies and that the true preferences in the uncorrelated market are such that any agent finds all members of the other side of the market acceptable. This means that $\pi_m(6) = 6$ for any $m \in M$. Additionally, this tells us that $\pi_m^{-1}(6) = 6$ for any $m \in M$. Compositions of permutations with this property are also permutations with this property. Finally, note that interchanging two elements of a preference that are ranked above \emptyset does not change the number of elements ranked above \emptyset . Hence

$$(7) \quad \Pr \left\{ \widetilde{P}_{-w}^{m_1 \leftrightarrow m_2} \right\} = \Pr \left\{ \text{True} \left(\widetilde{P}_{-w}^{m_1 \leftrightarrow m_2} \right) \right\} = \Pr \left\{ \text{True} \left(\widetilde{P}_{-w} \right) \right\} = \Pr \left\{ \widetilde{P}_{-w} \right\}$$

where the inner equality comes from the fact that all preferences that rank \emptyset last are equally likely in our environment and the outer equalities come from the fact that $\text{True}(\cdot)$ is a one-to-one mapping of true preferences to submitted preferences. We have shown that the reported preferences are $\{m, m'\}$ -symmetric under the assumed strategy restrictions, but since m_1 , m_2 and w' were not special, we further conclude that the reported preferences are M-symmetric under the assumed strategy restrictions. Thus, by Theorem 3.2 and Remark 3.2 from Ehlers (2008), we conclude that non-truncations are stochastically dominated by truncations for w . Also note that the theorem handles the case of mixed strategies. Let a state of the world be defined by the pure strategies that are submitted by each agent. Conditional on a state of the world, the proof proceeds as above. Taking the expectation over states of the world establishes the theorem for mixed strategies.

Theorem 2. *In the uncorrelated market, under M-Proposing Priority, any Bayes-Nash equilibrium in label-independent, weakly undominated strategies has any $m \in M$ reporting all members of W as acceptable and any $w \in W$ playing a truncation.*

PROOF:

The weakly undominated requirement means that each $m \in M$ must report all $w \in W$ as acceptable. Reporting all $w \in W$ as acceptable coupled with all agents playing label-independent strategies combine through Lemma 4 to tell us that at equilibrium, each $w \in W$ must be playing a truncation strategy.

Remark. *The proofs used from here until Theorem 3 borrow heavily from Roth and Rothblum (1999).*

Lemma 5. *Consider, P , w' , and m , and let $v \in (W \setminus \{w, w'\}) \cup \{\emptyset\}$. Denote the match of M when the submitted preferences are P under M-Proposing Priority as $\text{MPP}[P](m)$. Then,*

$$(8) \quad \begin{aligned} \text{MPP}[P](m) = v &\Leftrightarrow \text{MPP}[P^{w \leftrightarrow w'}](m) = v \\ \text{MPP}[P](m) = w &\Leftrightarrow \text{MPP}[P^{w \leftrightarrow w'}](m) = w' \end{aligned}$$

PROOF:

MPP does not give special treatment to any given label.

Corollary (Lemma 5). *Also true is then*

$$(9) \quad \begin{aligned} \text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}](m) = v &\Leftrightarrow \text{MPP}[P_m, P_{-m}^{w \leftrightarrow w'}](m) = v \\ \text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}](m) = w &\Leftrightarrow \text{MPP}[P_m, P_{-m}^{w \leftrightarrow w'}](m) = w' \end{aligned}$$

PROOF:

Applying the $w \leftrightarrow w'$ interchange operator to $(P_m^{w \leftrightarrow w'}, P_{-m})$ yields $(P_m, P_{-m}^{w \leftrightarrow w'})$.

Lemma 6. *Let $w' \prec_m w$. Then, $(\text{MPP}[P](m) = w') \Leftrightarrow (\text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}](m) = w')$.*

PROOF:

Switching w' and w in a submitted ordering means that w' is proposed to in an earlier round. If it was available in the later round, it will still be available in the earlier round, and no one else will be proposing to it in that round.

Corollary (Lemma 6). *Let $w' \prec_m w$. Then,*

$$(10) \quad \left(\text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}](m) = w \right) \Leftrightarrow (\text{MPP}[P](m) = w)$$

PROOF:

Again, if we can get w by moving it down on the list, then we can certainly get it by moving it up further.

Lemma 7. *In a one-to-one, two-sided matching market with W -symmetric reported preferences, then for the members of M , all strategies but truth-telling are weakly dominated.*

PROOF:

Consider the outcome for some $m \in M$ for whom $w \prec_m w'$ when he submits a preference that truthfully ranks w and w' , $\text{MPP}[P](m)$, and when he submits a preference that switches w and w' , $\text{MPP}[P_m^{w \leftrightarrow w'}, P_{-m}](m)$. Lemma 6 and its corollary tell us what cases are impossible, and Lemma 5 and its corollary allow us to determine what happens when everyone else's preferences move from P_{-m} to $P_{-m}^{w \leftrightarrow w'}$. The following table summarizes what is possible, while the next table tells us what lottery m can expect when he truthfully orders w and w' and when he switches their ordering, given that everyone else's preferences are either P_{-m} or $P_{-m}^{w \leftrightarrow w'}$ with equal probability, by the assumption of W -symmetry. Note that we denote the lottery where x and y are drawn with equal probability by $\frac{1}{2}x \oplus \frac{1}{2}y$.

Denote the probability of being in Case B by p_B , etc. We want to express the probability of getting certain outcomes when m truthfully orders w and w' and when he doesn't in terms of these p_i s. We will then compare these probability distributions and show that the truthful strategy first-order stochastically dominates the switching strategy. Note that it is only in Cases B, C, and D that truthful ordering and switching given different

TABLE A.1—TABLE OF CASES FOR THE LEMMA.

		Lie: $\text{MPP}[P_m^{w \leftrightarrow w'}, P_m](m)$		
		$= v \notin \{w, w'\}$	$= w$	$= w'$
Truth: $\text{MPP}[P](m)$	$= u \notin \{w, w'\}$	Case A	Impossible	Case B
	$= w$	Case C	Case D	Case E
	$= w'$	Impossible	Impossible	Case F

TABLE A.2—PAYOFFS FOR EACH CASE.

	Truth	Lie
Case A	$\frac{1}{2}u \oplus \frac{1}{2}v$	$\frac{1}{2}u \oplus \frac{1}{2}v$
Case B	$\frac{1}{2}u \oplus \frac{1}{2}w$	$\frac{1}{2}u \oplus \frac{1}{2}w'$
Case C	$\frac{1}{2}v \oplus \frac{1}{2}w$	$\frac{1}{2}v \oplus \frac{1}{2}w'$
Case D	w	w'
Case E	$\frac{1}{2}w \oplus \frac{1}{2}w'$	$\frac{1}{2}w \oplus \frac{1}{2}w'$
Case F	$\frac{1}{2}w \oplus \frac{1}{2}w'$	$\frac{1}{2}w \oplus \frac{1}{2}w'$

stochastic outcomes. Under a preference with true ordering of w and w' , the probability distribution of the outcome conditional on being in these interesting cases is

$$(11) \quad \Pr \{v | \text{Truth and Case B, C, or D}\} = \begin{cases} \frac{p_B + p_C}{2 \cdot (p_B + p_C + p_D)} & v \notin \{w, w'\} \\ \frac{2 \cdot p_D + p_B + p_C}{2 \cdot (p_B + p_C + p_D)} & v = w \\ 0 & v = w' \end{cases}$$

And under the same ordering where w and w' are switched, the distribution of the outcome is

$$(12) \quad \Pr \{v | \text{Lie and Case B, C, or D}\} = \begin{cases} \frac{p_B + p_C}{2 \cdot (p_B + p_C + p_D)} & v \notin \{w, w'\} \\ 0 & v = w \\ \frac{2 \cdot p_D + p_B + p_C}{2 \cdot (p_B + p_C + p_D)} & v = w' \end{cases}$$

Taking the unconditional difference, remembering that it is only non-zero in Cases B, C,

and D , we find

$$(13) \quad \Pr \{v|\text{Truth}\} - \Pr \{v|\text{Lie}\} = \begin{cases} 0 & v \notin \{w, w'\} \\ p_D + \frac{1}{2} \cdot (p_B + p_C) & v = w \\ -(p_D + \frac{1}{2} \cdot (p_B + p_C)) & v = w' \end{cases}$$

Hence, by lying about the ordering of w and w' , m takes weight off of w and moves it to w' without changing the weighting of the other alternatives. Hence, truthfully ordering w and w' must weakly stochastically dominate reversing their order, if preference reports are $\{w, w'\}$ -symmetric. Since, by assumption, reported preferences are W -symmetric, m cannot gain by switching the order of any two W s. We also know, by Lemma 3 that under M -Proposing Priority, not ranking all acceptable W s is weakly dominated. Ranking unacceptable W s is also weakly dominated, since matches are permanent in each round. Hence, truth-telling is the only strategy for m that is not weakly dominated.

Theorem 3. *In a one-to-one, two-sided matching market with W -symmetric true preferences, if all members of W report the same number of M s acceptable, then under any Bayes-Nash equilibrium in label-independent, weakly undominated strategies, the members of M must be truth-telling.*

PROOF:

It is weakly dominated not to rank all acceptable match partners. Ranking an unacceptable match partner acceptable wont help, since matches are made permanent in every round of the M -Proposing Priority mechanism. If we assume the same number of M s are declared acceptable in a label-independent way by all W s, then we know that the reported preferences are W -symmetric. Hence, by Lemma 7, truth-telling is weakly dominant.

Theorem 4. *In the uncorrelated market, a member of M who plays a label-independent, weakly undominated strategy reports all complete preference lists that declare every member of W acceptable with equal probability. Thus, the interim payoff probability distribution for any strategy played by any W is independent of the strategies being played by the M s, so long as those M s are playing label-independent, weakly undominated strategies.*

PROOF:

In the uncorrelated market, all potential match partners are acceptable. Thus, from Lemma 3, all submitted priorities for M s playing weakly undominated strategies will in-

clude all potential match partners. Thus, when considering weakly undominated strategies we consider only permutations of an M 's true preferences, i.e., all W s are included. Furthermore, all preference permutations are equally likely ex ante. Thus, in order to respect label independence, any weakly undominated strategy played by an M must result in all full orderings of W s being reported with equal ex ante likelihood. Thus, W 's actions are not affected ex ante by the strategy chosen by any M (as long as the strategy is undominated and label-independent).

Lemma 8. *In the correlated market, if each $m \in M$ truthfully reveals, then there is a unique stable match relative to the submitted preferences.*

PROOF:

Define $B(w, A) = \max_{\succ_w} \{m \in A \cup \{\emptyset\} | m \succ_w \emptyset\}$. This is w 's favorite acceptable match out of the set A , or \emptyset if w has no acceptable match in A . Now, without loss of generality, assume that $P_m = (w_1, w_2, w_3, w_4, w_5, \emptyset)$ for each $m \in M$. First, note that any stable match must contain the pair $(w_1, B(w_1, M))$, since otherwise, this would be a blocking pair. Given this, we also know that $(w_2, B(w_2, M \setminus B(w_1, M)))$ must be in any stable match, since otherwise it would be a blocking pair. We can continue this process down the list of agents, leaving us with a unique stable match.

Lemma 9. *If there is a unique stable match relative to the reported preferences when some $w \in W$ truthfully reveals, then that w cannot do better by deviating from truth-telling.*

PROOF:

Assume to the contrary that she could do better, i.e. that when she truthfully reveals, she matches to m and when she reports some non-truthful preference, she matches to $m' \succ_w m$. By uniqueness in the first match, there must be a blocking pair in the second match relative to the originally submitted preferences. Since the second match is stable relative to the altered preferences, and only w altered her preferences, that blocking pair must be of form (w, m'') where $m'' \succ_w m'$ and m'' prefers w to his partner in the second match. Because of this, m'' must have proposed to w and been rejected. Clearly, w could do better by simply not rejecting m' in favor of some other member of M . Then, w has a new preference that is not truthful that yields a match where she is partnered to m'' . At this point, the same logic that got her from m' to m'' can also get her from m'' to some $m''' \succ_w m''$. Continuing in this way, she can always come up with a new lie that outperforms her old one. Since the market is finite, this is a contradiction.

Theorem 5. *In the correlated market, under M -Proposing DA, the unique Bayes-Nash equilibrium in label-independent, weakly undominated strategies entails truth-telling by all agents.*

PROOF:

Weakly undominated strategies requires that the M s truthfully reveal, which by Lemma 8 implies a unique stable match. Lemma 9 then tells us that any member of W can best-respond by truth-telling. Since any member of W has a one-in-five chance to be the first ranked w by the members of M , any deviation from truth-telling can hurt the w . Hence, all members of W must truthfully reveal in an equilibrium where no agent plays weakly undominated strategies.

Theorem 6. *In the correlated market, if all members of M have the same weakly undominated, label-independent revelation strategy, then all members of W best respond by truthfully revealing under M -Proposing Priority.*

PROOF:

Weakly undominated for the M s mean that all women are listed as acceptable. Under the assumptions, a member of W will receive all offers in one round of the algorithm. There is no gain to not truthfully revealing then, as our member of W is facing a static decision problem. Since every W has a one-in-five chance of being the last ranked W by all M s, there is always a positive loss to dropping, and since every W also has a one-in-five chance of receiving all offers in the first round.

Theorem 7. *In the uncorrelated environment, there exist cardinal payoffs such that there exists an equilibrium where all M s and W s truthfully reveal their preferences.*

PROOF:

Consider the payoff vector $(u(w_i))_{i \in \{1,2,3,4,5,\emptyset\}} = (206, 41, 10, 3, 1, 0)$. Even if all other members of M rank w_1 first, a given $m \in M$ does best to also rank w_1 first, since $206/5 > 41$. Similar logic holds for ranking w_2 2nd, etc. Hence all M s truthfully reveal, and by Theorem 6, the W s must as well.