Strategic Uncertainty and Equilibrium Selection in Stable Matching Mechanisms: Experimental Evidence

Marco Castillo* Ahrash Dianat**

Abstract

We present experimental evidence on the interplay between strategic uncertainty and equilibrium selection in stable matching mechanisms. In particular, we apply a version of risk-dominance to compare the riskiness of "truncation" against other strategies that secure against remaining unmatched. By keeping subjects' ordinal preferences fixed while changing their cardinal representation, our experimental treatments vary the risk-dominant prediction. We find that both truth-telling and truncation are played more often when they are risk-dominant. In both treatments, however, truncation strategies are played more often in later rounds of the experiment. Our results also shed light on several open questions in market design.

JEL codes: C72, C78, D47
Keywords: stable matching; equilibrium selection; risk-dominance

* Department of Economics, Texas A&M University, marco.castillo@tamu.edu
** Department of Economics, University of Essex, a.dianat@essex.ac.uk

This research was supported by the International Foundation for Research in Experimental Economics (IFREE).

1 Introduction

Strategic uncertainty plays an important role in games with multiple equilibria. The central insight is as follows: when there are multiple rationalizable actions, players often face a tension between profitability and safety. In 2×2 coordination games, the concept of *risk-dominance* was developed to capture the intuition that one equilibrium may appear more or less risky than another equilibrium (Harsanyi and Selten, 1988). Lab experiments have documented the predictive power of risk-dominance in simple settings (Cooper et al., 1990; Van Huyck et al., 1990). More recently, variations of risk-dominance have been fruitfully applied to other domains. In the infinitely repeated prisoners' dilemma, for instance, there is experimental evidence that cooperation is more likely to be sustained when it is both an equilibrium action and a risk-dominant action (Bó and Fréchette, 2011).

In this paper, we apply a version of risk-dominance to stable matching mechanisms. In a stable matching mechanism, participants in a two-sided market report rank-order lists of their preferences over match partners to a central authority. The central authority then uses the reported preferences to calculate the final matching. Crucially, the final matching is stable with respect to the reported preferences.[1] This environment induces a *preference-revelation game* in which the strategy space is the set of all possible ordinal preference lists.

In particular, we investigate how varying the "riskiness" of preference misrepresentation affects selection among stable matchings. By automating the side of the market that has a dominant strategy, we are able to model the strategic environment as a coordination game. Specifically, this coordination game has at least two symmetric and Pareto-ranked equilibria in pure strategies: an equilibrium in "truncation" strategies (i.e., removing less preferred match partners from the tail end of a preference list) and an equilibrium in "permutation" strategies (i.e., switching the order of match partners in a preference list). Although the truncation equilibrium yields a higher payoff, the presence of strategic uncertainty makes truncation strategies less appealing. Intuitively, truncation generates a trade-off between the likelihood of matching and the quality of match partner (conditional on matching). Thus, an agent who plays a truncation strategy opens herself up to the possibility of remaining unmatched for some profile of other agents' preference reports.[2] This insight allows us to apply a version of risk-dominance to compare the riskiness of truncation against other strategies that secure against remaining unmatched.

[1] A matching is said to be stable if no agent prefers remaining unmatched to her current allocation and no pair of agents both prefer each other to their current allocations.

[2] Even after removing strategic uncertainty, there is also the possibility of remaining unmatched due to "over-truncation" (i.e., playing the wrong kind of truncation strategy). However, over-truncation is not possible in our experimental set-up. In a related paper, Castillo and Dianat (2016) present evidence that laboratory subjects are less likely to truncate their preferences when the possibility of over-truncation exists.

We then study this coordination game in the lab, using a simple setting that allows us to isolate the strategic features of interest. Each experimental market consists of four participants: two firms (computer roles) and two workers (subject roles). Subjects play 20 rounds of the preference-revelation game induced by the firm-optimal stable mechanism, with random and anonymous re-pairing across rounds.[3] We use an ordinal constellation of preferences such that each subject has two stable match partners.[4] However, these ordinal preferences can be represented by different cardinal utilities. Indeed, we have full freedom in choosing the payoff difference between two ordered alternatives. In our experiment, we choose cardinal representations such that our treatments vary whether the criterion of risk-dominance selects truth-telling or truncation. When remaining unmatched is particularly costly (to be precisely defined later), then truth-telling is risk-dominant. When an agent has a strong intensity of preference for her first choice partner (to be precisely defined later), then truncation is risk-dominant.

We now preview our main results. Overall, we find that truth-telling is the modal strategy. However, we observe several notable patterns in the experimental data. First, both truth-telling and truncation are played more often when they are risk-dominant. This result is robust to whether the treatment effect is measured at the subject or the session level. Second, in both treatments, truncation strategies are played more often in later rounds of the experiment. Since truncation is a necessary component of all payoff-dominant equilibria, this suggests that the salience of payoff-dominance as a selection criterion increases with subject experience.

It is crucial to note that, under the conditions imposed in our experiment, the set of Nash equilibrium outcomes is identical to the set of stable outcomes. Thus, selecting among the Nash equilibria of the preference-revelation game is equivalent to selecting among the stable matchings of the underlying two-sided matching market. This framing illustrates the welfare implications of equilibrium selection. When attention is confined to stable outcomes, the interests of the two sides of the market are opposed in a fundamental sense: the best stable matching for one side of the market is the worst stable matching for the other side of the market.[5] This makes equilibrium selection an important and relevant consideration for policymakers, who may have reasons to favor the welfare of one side of the market over another when designing matching institutions. An example of this is provided by the history of the National Resident Matching Program (NRMP), the entry-level labor market

[3]Under truth-telling, the firm-optimal stable mechanism generates the firm-optimal stable matching. Thus, firms (computers) have a dominant strategy of truth-telling and workers (subjects) have incentives to misrepresent their preferences to influence the final outcome.

[4]That is, each worker can be matched to either firm at a stable matching.

[5]This is a consequence of the fact that the set of stable matchings has a lattice structure.

for American physicians. In May 1997, the NRMP unanimously voted to alter the algorithm that was being used over concerns that the original design unduly favored hospitals at the expense of students.[6]

The argument for using laboratory experiments in this context is compelling. While data on participants' submitted rank-order lists may be available in field settings, participants' true preferences are unobserved. Under the assumption of truthful preference reporting, matching clearinghouses used in practice implement an extremal stable outcome (i.e., the most preferred stable outcome for one side of the market). But if agents strategically misrepresent their preferences in the field, then it is less clear which stable matching is implemented or even whether the final matching is stable.[7] By allowing us to control for subjects' preferences and other market features, the laboratory setting is ideally suited for answering questions related to equilibrium selection.

Our experiment also sheds light on several open questions in market design. First, our results provide support for the empirical relevance of truncation strategies and highlight the conditions that foster truncation behavior. In particular, our results suggest that truncation should occur more often when individuals have a strong intensity of preference for top-ranked alternatives or when individuals have previous experience with the particular mechanism that is being used. Ideally, market designers could combine this insight with institution-specific details to predict which stable matching is more likely to be implemented when there are multiple candidates for consideration.

Second, our experiment can speak to the literature on "core convergence" in matching markets.[8] Although our experimental markets have two disjoint stable matchings with respect to the true (i.e., induced) preferences, our experimental data reveal that 72% of markets have a unique stable matching with respect to the reported preferences. Thus, our results underscore the fact that preference misrepresentation can contribute to a (false) perception that matching markets have a singleton core. This suggests a note of caution against interpreting reported preferences as true preferences when conducting welfare analysis.

Our paper naturally bridges several different strands of literature. First, there is a growing body of work on the performance of matching mechanisms in the lab.[9] The current paper

[6]Specifically, the NRMP switched from a version of the hospital-proposing deferred acceptance algorithm to a version of the student-proposing deferred acceptance algorithm.

[7]Recent studies by Rees-Jones (2016) and Hassidim et al. (2015) provide evidence of preference misrepresentation in the field.

[8]"Core convergence" refers to the theoretical and empirical result that the set of stable matchings shrinks as the size of the market increases.

[9]See, for instance, Ding and Schotter (2016), Chen and Sönmez (2006), Fragiadakis and Troyan (2015), Castillo and Dianat (2016), Echenique et al. (2016), Featherstone and Mayefsky (2015), Featherstone and Niederle (2014), Harrison and McCabe (1989), Pais and Pintér (2008), and Klijn et al. (2013).

is most closely related to Castillo and Dianat (2016), which is an experimental investigation of truncation behavior in what approximates a decision-theoretic setting. To that end, Castillo and Dianat (2016) employ a restricted strategy space that yields a unique equilibrium outcome. The current design, on the other hand, introduces multiple Pareto-ranked equilibria and allows us to better understand the conditions that favor the implementation of one equilibrium over another. There are also experimental studies of matching mechanisms that investigate whether intensity of preference has implications for strategic behavior. For instance, Echenique et al. (2016) report that the cardinal representation of subjects' preferences has a significant effect on the stability of final outcomes, with instability more likely to arise in the presence of weak incentives. However, Klijn et al. (2013) find that subject behavior is fairly robust to changes in cardinal preferences in the Gale-Shapley mechanism but not the Boston mechanism.

Second, there is a large experimental literature on behavior in coordination games.[10] Several of these studies focus on stag hunt games in an attempt to evaluate the merits of competing equilibrium selection principles. Our results are largely consistent with this literature, suggesting that equilibrium selection arguments may have significant generality and explanatory power across environments. In particular, our finding that the salience of payoff-dominance increases with subject experience mirrors that of Rankin et al. (2000), who show that subjects learn to gravitate toward payoff-dominance in a sequence of stag hunt games where cosmetic details (e.g., action labels, player labels, payoffs) are randomly perturbed.

Finally, there is a literature that attempts to extend the concept of risk-dominance to other strategic environments.[11] In particular, our work can be viewed as methodologically related to Bó and Fréchette (2011), which applies a version of risk-dominance in the context of the infinitely-repeated prisoner's dilemma. In their construction, risk-dominance remains a pairwise comparison among Nash equilibria: instead of applying the concept to the entire strategy space, they focus on a simplified version of the game that consists of two (focal) equilibrium strategies.

The rest of the paper is organized as follows. Section 2 introduces the necessary theoretical background, Section 3 presents our experimental design, Section 4 presents our experimental results, Section 5 discusses broader implications for market design, and Section 6 concludes.

[10] For a survey of this literature, see Camerer (2003).

[11] Other closely-related solution concepts have also been introduced for more general finite games. Morris et al. (1995) define a concept called *p-dominance* in which candidate action pairs are compared against all other actions (not just equilibrium actions). Kojima (2006) defines a concept called *u-dominance* that is well-suited for games with strategic complementarities.

2 Theoretical Background

There are two finite and disjoint sets of agents of equal size: a set F of firms and a set W of workers. The preferences of worker $w \in W$ are represented by the von Neumann-Morgenstern utility function $u_w : F \cup \{w\} \to \mathbb{R}_+$, where $u_w(f) > 0$ is the utility she derives from matching with firm $f \in F$ and $u_w(w) = 0$ is the utility she derives from remaining single.[12] We assume that each function u_w is one-to-one, such that it induces a strict preference ordering P_w on the set F. We will refer to P_w as the *preference list* of worker w. The preferences of the firms are defined similarly. We let $u = (u_i)_{i \in F \cup W}$ denote the profile of agents' utility functions and $P = (P_i)_{i \in F \cup W}$ denote the profile of agents' preference lists.

A *matching market* is a triple (F, W, u). A matching is a function $\mu : F \cup W \to F \cup W$ such that:

1. for any $f \in F$, $\mu(f) \in W \cup \{f\}$
2. for any $w \in W$, $\mu(w) \in F \cup \{w\}$
3. for any $f \in F$, $w \in W$, $\mu(f) = w$ if and only if $\mu(w) = f$

A pair of agents (f, w) is said to *block* a matching μ if they are not matched to one another at μ but they prefer each other to their assignments at μ (i.e., $u_w(f) > u_w(\mu(w))$ and $u_f(w) > u_f(\mu(f))$). A matching μ is *stable* if it is not blocked by any pair of agents. A firm f and a worker w are said to be *achievable* for each other in a matching market (F, W, u) if they are matched to each other at some stable matching. A stable matching is called *firm-optimal* (*worker-pessimal*) if each firm is matched to her most preferred achievable worker (each worker is matched to her least preferred achievable firm). Similarly, a stable matching is called *worker-optimal* (*firm-pessimal*) if each worker is matched to her most preferred achievable firm (each firm is matched to her least preferred achievable worker). We denote the firm-optimal stable matching by μ_F and the worker-optimal stable matching by μ_W. In other words, for each firm $f \in F$, each worker $w \in W$, and each stable matching μ, we have that $u_f(\mu_F(f)) \geq u_f(\mu(f)) \geq u_f(\mu_W(f))$ and $u_w(\mu_W(w)) \geq u_w(\mu(w)) \geq u_w(\mu_F(w))$.

Let $\mathcal{M}$ denote the set of all possible matchings, $\mathcal{Q}$ denote the set of all possible preference profiles, and $\mathcal{Q}_i$ denote the set of all possible preference lists for agent $i \in F \cup W$. Let μ, Q, and Q_i denote arbitrary elements of the sets $\mathcal{M}$, $\mathcal{Q}$, and $\mathcal{Q}_i$, respectively. A mechanism is a function $\phi : \mathcal{Q} \to \mathcal{M}$ that assigns a matching to each preference profile. A mechanism ϕ

[12] Although the classical results we present are usually framed in terms of ordinal preferences, the solution concept of risk-dominance is inherently cardinal. Thus, we assume cardinal preferences throughout the analysis.

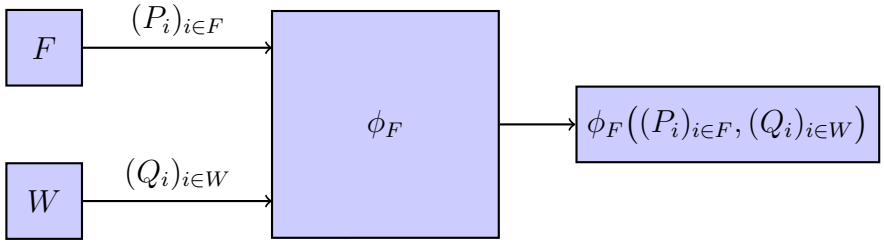


Figure 1: The constrained-preference revelation game induced by the firm-optimal stable mechanism.

that for each preference profile Q produces a matching $\phi(Q)$ that is stable with respect to Q is called a stable mechanism. If $\phi(Q)$ is the firm-optimal stable matching with respect to Q, then ϕ is called the firm-optimal stable mechanism. We denote the firm-optimal stable mechanism by ϕ_F.

The firm-optimal stable mechanism can be modeled as a non-cooperative game in which the strategy space is the set of all possible ordinal preference lists. In this preference-revelation game, it is well-known that the firms have a dominant strategy of truth-telling (Dubins and Freedman, 1981). In markets with more than one stable matching, however, at least one worker will have an incentive to misrepresent her preferences to improve her match outcome (Gale and Sotomayor, 1985). Our goal is to characterize the different equilibria that can arise in this environment. To simplify our analysis, we define the *constrained preference-revelation game* induced by the firm-optimal stable mechanism:

Definition 1. *Consider a matching market (F, W, u) in which $P = (P_i)_{i\in F\cup W}$ denotes the profile of agents' true preference lists and $Q = (Q_i)_{i\in F\cup W}$ denotes the profile of agents' reported preference lists. The* ***constrained preference-revelation game*** *induced by the firm-optimal stable mechanism ϕ_F is the preference-revelation game in which the firms are constrained to truth-telling. That is, for any profile of workers' reports $(Q_i)_{i\in W}$, the constrained preference-revelation game produces the matching $\phi_F\big((P_i)_{i\in F},(Q_i)_{i\in W}\big)$.*

Figure 1 shows an illustration of the constrained preference-revelation game, where the firms are constrained to play their dominant strategy of truth-telling while the workers are free to report any preference ordering.[13]

We will find it useful to define two types of misrepresentation strategies for the workers:

[13]The assumption that firms play their dominant strategy in the firm-optimal stable mechanism is not entirely innocuous. While the firm-optimal stable mechanism is strategy-proof for the firms, Ashlagi and Gonczarowski (2016) show that it is not obviously strategy-proof in the sense of Li (2016). Furthermore, empirical studies by Rees-Jones (2016) and Hassidim et al. (2015) find that a small fraction of participants fail to play their dominant strategy in strategy-proof matching mechanisms.

Definition 2. *A **truncation** of a preference list P_w containing k firms is a list Q_w containing $k' < k$ firms such that the k' elements of Q_w are the first k' elements of P_w, in the same order.*[14]

Definition 3. *A **permutation** of a preference list P_w is a list $Q_w \neq P_w$ that is not a truncation of P_w.*[15]

In other words, a truncation involves misrepresenting preferences by removing match partners from the tail end of a preference list, while a permutation involves misrepresenting preferences by switching the order of match partners in a preference list (regardless of the length of the list).

We now state and prove some basic results that are relevant to our experimental design. Throughout, we let μ_F and μ_W denote the firm-optimal and worker-optimal stable matchings with respect to the true preferences P.

Proposition 1. *Consider a matching market (F, W, u) in which all agents have more than one achievable partner. In the constrained preference-revelation game induced by the firm-optimal stable mechanism ϕ_F, there is a payoff-dominant equilibrium in which all workers play truncation strategies.*

Proof. Let T denote the profile of reported preferences in which each worker w truncates her preference list by removing all firms ranked below $\mu_W(w)$. By Theorem 4.17 of Roth and Sotomayor (1992), T is a Nash equilibrium and it produces the matching μ_W. Suppose another Nash equilibrium Q payoff-dominates T. Let μ denote the matching that is produced by Q. Since Q is a Nash equilibrium, we know by Theorem 4.16 of Roth and Sotomayor (1992) that the matching μ is also stable with respect to the true preferences P. Since Q payoff-dominates T, we know that $u_w(\mu(w)) > u_w(\mu_W(w))$ for all $w \in W$. We have arrived at a contradiction, since μ_W is the W-optimal stable matching with respect to P. Thus, there is no other Nash equilibrium that payoff-dominates T. We conclude that T is payoff-dominant. □

We will refer to the equilibrium in which all workers play truncation strategies as the "symmetric" truncation equilibrium. However, one worker's truncation decision creates positive spillovers for other workers in the market (Ashlagi and Klijn, 2012; Coles and Shorrer, 2014). This implies that asymmetric equilibria also exist in which a subset of workers plays

[14]This definition is taken from Roth and Rothblum (1999). However, it has been slightly modified such that truthful preference revelation is no longer an "edge case" of a truncation strategy.

[15]The term "dropping strategy" is often used to refer to the act of removing a match partner from the middle of a preference list (rather than from the tail end of a preference list). According to our definitions, a dropping strategy would be classified as a permutation.

truncation strategies and the remaining workers report their true preferences, effectively free-riding on others' truncation behavior.

We now construct a "symmetric" permutation equilibrium in which all workers play permutation strategies.

Proposition 2. *Consider a matching market (F, W, u) in which all agents have more than one achievable partner. In the constrained preference-revelation game induced by the firm-optimal stable mechanism ϕ_F, there is a payoff-dominated equilibrium in which all workers play permutation strategies.*

Proof. Let Q denote the profile of reported preferences in which each worker w reports a preference list Q_w that ranks $\mu_F(w)$ in the first position (regardless of the length of the list). Each preference list Q_w is clearly a permutation since $\mu_F(w)$ is not at the head of any worker's true preference list.[16] It is straightforward to see that Q produces the matching μ_F.

We argue that the profile of reported preferences Q constitutes a Nash equilibrium.[17] To see this, suppose that Q is not a Nash equilibrium. Then, there exists some worker w who can deviate and report a preference list Q'_w, which leads to a new profile of reported preferences $Q' = (Q_{-w}, Q'_w)$ and a new matching μ' such that $u_w(\mu'(w)) > u_w(\mu_F(w))$. Let $f = \mu'(w)$. Then firm f must have been matched to a worker she prefers to w at μ_F, otherwise (f, w) would have blocked the matching μ_F under the true preferences P. But now firm f and worker $\mu_F(f)$ block the matching μ' under the reported preferences Q', which is a contradiction. Therefore, Q is a Nash equilibrium. Furthermore, Q is payoff-dominated by the truncation equilibrium constructed in Proposition 1. □

Although all workers prefer the truncation equilibrium to the permutation equilibrium, truncation behavior introduces the possibility of remaining unmatched for some profiles of other agents' reported preferences. This exposure to the worst possible outcome is not present for the permutation strategy that we identify. To see this more clearly, it is instructive to consider the steps of the firm-proposing deferred acceptance algorithm.[18] A consequential truncation (i.e., a truncation that affects the final outcome) requires a worker to reject a proposal from an achievable firm. This rejection frees the firm to make other proposals, which may cause a chain of further rejections. If the truncating worker does not receive new proposals from this rejection chain, then she will remain single. The permutation strategy

[16] If $\mu_F(w)$ were at the head of any worker's true preference list, then this contradicts the assumption that all workers have more than one achievable partner.

[17] The proof of this claim closely follows the proof of Theorem 4.15 in Roth and Sotomayor (1992). The only difference is that we allow the preference lists in Q to be of any length.

[18] The firm-proposing deferred acceptance algorithm is a procedure that generates the firm-optimal stable matching for any preference profile.

$$P_{f_1} = w_1, w_2 \quad P_{w_1} = f_2, f_1$$
$$P_{f_2} = w_2, w_1 \quad P_{w_2} = f_1, f_2$$

Table 1: The ordinal preferences used in the experiment. The firm-optimal stable matching is shown in red and the worker-optimal stable matching is shown in blue.

	Truth	*Truncate*	*Permute*
Truth	v_2, v_2	v_1, v_1	v_2, v_2
Truncate	v_1, v_1	v_1, v_1	v_3, v_2
Permute	v_2, v_2	v_2, v_3	v_2, v_2

Table 2: Normal-form representation of the constrained preference-revelation game ($v_1 > v_2 > v_3$). The Nash equilibrium that implements the firm-optimal stable matching is shown in red. The Nash equilibria that implement the worker-optimal stable matching are shown in blue.

in Proposition 2 is inherently not consequential. It produces the same outcome as truth-telling: it merely prioritizes the least preferred achievable firm by elevating her to the top of a worker's preference ordering.

3 Experimental Design

Each experimental market consists of four participants: two firms (computer roles) and two workers (subject roles). The firms are automated to play their dominant strategy of truth-telling, while the workers are free to report any preference ordering. In each session, subjects play 20 rounds of the constrained preference-revelation game induced by the firm-optimal stable mechanism.[19] Subjects are randomly and anonymously re-paired at the start of each round.

Table 1 shows the ordinal preferences used across all 20 rounds of the experiment. With this constellation of preferences, there are two disjoint stable matchings (i.e., each agent has two achievable match partners). For convenience, we let v_1, v_2 and v_3 denote an agent's utilities from matching with her most preferred partner, her least preferred partner, and remaining single, respectively. Table 2 depicts the normal-form representation of the constrained preference-revelation game.[20] It should be noted that *Permute* combines two pure strategies that are strategically equivalent.[21]

[19] In the experiment, the firm-proposing deferred acceptance algorithm is used to illustrate to subjects how reported preferences map to final outcomes.

[20] When firms are unconstrained, there exist other equilibria in which firms play dominated strategies.

[21] For instance, consider the situation facing worker w_1 with preference list $P_{w_1} = f_2, f_1$. Both permutation strategies ($Q_{w_1} = f_1, f_2$ and $Q'_{w_1} = f_1$) yield the same outcome for all preference reports by the other player. More generally, the equivalence of different permutation strategies need not hold.

We now apply our notion of risk-dominance to the constrained preference-revelation game. The original Harsanyi and Selten (1988) criterion of risk-dominance concerns the pairwise comparison of Nash equilibria and is intended to capture the intuition of one equilibrium being more or less "risky" than another equilibrium. In their formulation, constructed for 2×2 games, equilibrium A risk-dominates equilibrium B if A has the larger Nash product (i.e., if the product of the two players' unilateral deviation losses are larger when moving from A to B than when moving from B to A). However, generalizing the concept of risk-dominance presents complications.[22] In this paper, we use an alternative characterization of risk-dominance in terms of the basins of attraction of different strategies.[23] We will say that a strategy is risk-dominant if it has the largest basin of attraction. It should be noted that, although we will speak of risk-dominant *strategies* rather than risk-dominant *strategy profiles*, our notion coincides with the Harsanyi and Selten (1988) criterion for 2×2 games and has the additional advantage of straightforwardly generalizing to any finite normal-form game.

Assume that all players play truth-telling with probability σ_{TT}, truncation with probability σ_{TR}, and permutation with probability σ_P. Then, player i's expected payoffs from the three strategies are as follows:

$$u_{TT} = v_2 + \sigma_{TR}(v_1 - v_2),$$

$$u_{TR} = v_3 + (\sigma_{TT} + \sigma_{TR})(v_1 - v_3),$$

$$u_P = v_2.$$

It is easy to see that a player is indifferent between truth-telling and truncation when

$$\sigma_{TR} = 1 - \sigma_{TT}\Big(\frac{v_1 - v_3}{v_2 - v_3}\Big),$$

which yields the basins of attraction shown in Figure 2.[24] Denote the basins of attraction of truth-telling and truncation by $\mathcal{B}_{TT}$ and $\mathcal{B}_{TR}$, respectively. Then, the Lebesgue measures of

[22]For instance, the binary relation imposed by risk-dominance can fail to be transitive. Morris et al. (1995) provide an example of a 3×3 game with three strict Nash equilibria in which the risk-dominance relationship is cyclical.

[23]For player i, the basin of attraction of strategy a_i is the set of beliefs σ_{-i} on the other players' strategies such that playing a_i is a best-response.

[24]Since permutation is a weakly dominated strategy, it can never be a strict best-response.

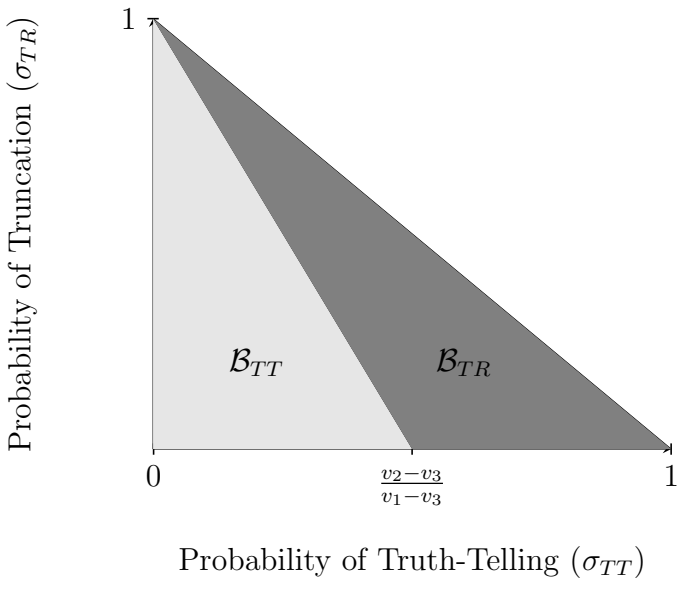


Figure 2: Basins of attraction of truth-telling ($\mathcal{B}_{TT}$) and truncation ($\mathcal{B}_{TR}$).

the two sets are as follows:[25]

$$m(\mathcal{B}_{TT}) = \frac{1}{2}\Big(\frac{v_2 - v_3}{v_1 - v_3}\Big)$$

and

$$m(\mathcal{B}_{TR}) = \frac{1}{2}\Big(1 - \frac{v_2 - v_3}{v_1 - v_3}\Big).$$

We will say that a strategy is risk-dominant if it has the larger basin of attraction. Truth-telling has the larger basin of attraction when $m(\mathcal{B}_{TT}) > m(\mathcal{B}_{TR})$, which yields $v_2 - v_3 > v_1 - v_2$. On the other hand, truncation has the larger basin of attraction when $m(\mathcal{B}_{TR}) > m(\mathcal{B}_{TT})$, which yields $v_1 - v_2 > v_2 - v_3$. Our characterization of risk-dominance neatly captures both the logic and the danger of playing a truncation strategy. When remaining unmatched is particularly costly (i.e., $v_2 - v_3 > v_1 - v_2$), then truth-telling is risk-dominant. When an agent has a strong intensity of preference for her first choice partner (i.e., $v_1 - v_2 > v_2 - v_3$), then truncation is risk-dominant.

For all agents, however, truth-telling constitutes the unique *protective* strategy in stable matching mechanisms (Barberà and Dutta, 1995).[26] This is because truth-telling is the only strategy that accomplishes the following two objectives: (1) it secures against the worst possible outcome (i.e., remaining single) and (2) it leads to the best possible outcome for some profile of other agents' preference reports.

[25]Let $A, B, C \in \mathbb{R}^2$. Then, the Lebesgue measure of the triangle with vertices A, B, C is given by $\frac{1}{2}|(A - C) \wedge (B - C)|$, where $|(a, b) \wedge (c, d)| = |ad - bc|$.

[26]A protective strategy is a refinement of a maxmin strategy. Notice that while both $Truth$ and $Permute$ are maxmin strategies, $Truth$ weakly dominates $Permute$.

	Treatment	
	TT ($v_2 = 15$)	TR ($v_2 = 5$)
Truth-Telling	protective and risk dominant	protective
Truncation	payoff dominant	payoff dominant and risk dominant

Table 3: Our experimental treatments vary the risk-dominant prediction.
TT: truth-telling is risk-dominant. TR: truncation is risk-dominant.

We review the key strategic features of this game:[27]

1. $(Truth, Truth)$ is not an equilibrium.
2. There are two symmetric equilibria: $(Truncate, Truncate)$ and $(Permute, Permute)$.
3. There are asymmetric equilibria in which one player truncates her preferences and the other player reports her true preferences.
4. Any equilibrium involving truncation strategies is payoff-dominant.
5. If $v_2 - v_3 > v_1 - v_2$, then truth is risk-dominant.
6. If $v_1 - v_2 > v_2 - v_3$, then truncation is risk-dominant.
7. $Truth$ is the unique protective strategy.

Table 3 summarizes our experimental design. In our experiment, we fix the payoff from matching with the most preferred partner ($v_1 = 20$) and from the outside option of remaining single ($v_3 = 0$). Our treatments vary the risk-dominant prediction by manipulating the payoff from matching with the least preferred partner (v_2). For $v_2 \in (10, 20)$, truth-telling is risk-dominant (TT treatment). For $v_2 \in (0, 10)$, truncation is risk-dominant (TR treatment). In our experiment, we set $v_2 = 15$ in the TT treatment and $v_2 = 5$ in the TR treatment.

To secure comprehension, subjects are required to walk through a demonstration of the deferred acceptance algorithm with a hypothetical set of reported preferences and to correctly answer a series of questions. In each round of the experiment, subjects observe the preferences of all market participants and they are then asked to report a preference ordering. At the end of a round, subjects receive feedback about the identity of their match partner (i.e., FIRM A, FIRM B, unmatched) and their payoff in that particular round. The experimental

[27]Recall that these features are not a function of the ordinal preferences we are using in the experiment. Rather, these features exist in any matching market where agents have multiple achievable match partners (i.e., where there are disjoint stable matchings).

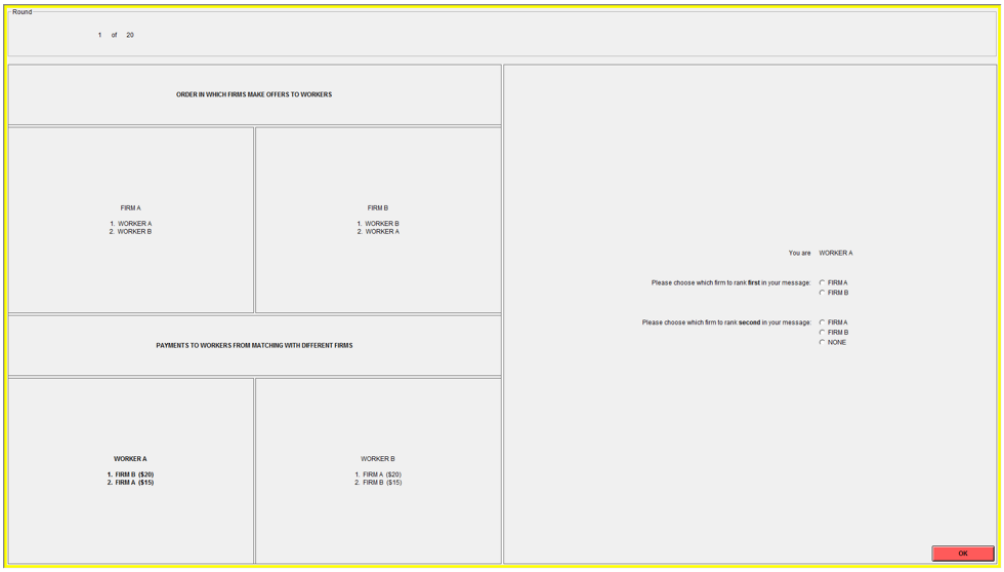


Figure 3: The experimental interface for the TT treatment.

interface for the TT treatment is shown in Figure 3 and the experimental instructions for the TT treatment are included at the end of the paper.[28]

4 Experimental Results

The experimental sessions were run at the Experimental Social Science Laboratory (ESSL) at UC Irvine. A total of 120 subjects participated in the experiment (TT treatment: 64 subjects, TR treatment: 56 subjects). Each experimental session lasted approximately one hour. One of the 20 experimental rounds was randomly selected for subject payment. Average subject earnings were $21.75 (including a $7 show-up payment). The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007). We now present our main results.

Result 1. *Both truth-telling and truncation are played more often when they are risk-dominant.*

	Round 1		Round 20	
Strategy	TT	TR	TT	TR
Truth-Telling	39 (61%)	39 (70%)	41 (64%)	26 (46%)
Truncation	3 (5%)	4 (7%)	16 (25%)	30 (54%)
Permutation	22 (34%)	13 (23%)	7 (11%)	0 (0%)
Total	64 (100%)	56 (100%)	64 (100%)	56 (100%)

Table 4: Round 1 and Round 20 Choices by Treatment.
TT: truth-telling is risk-dominant. TR: truncation is risk-dominant.

By round

We first demonstrate that this treatment effect is not present initially, but rather emerges with subject experience. Table 4 shows subject behavior in the first and last rounds of the experiment. In Round 1, truncation is rare in both treatments (TT: 5%, TR: 7%). In addition, a majority of subjects report their true preferences in both treatments (TT: 61%, TR: 70%). Neither of these treatment differences are statistically significant at conventional levels (truth-telling: two-sided t-test, $p = 0.32$; truncation: two-sided t-test, $p = 0.57$). In aggregate, we also find that initial behavior does not vary significantly across treatments (Fisher's exact test, $p = 0.39$).

However, there is a significant treatment effect at the end of the session (Fisher's exact test, $p < 0.01$). The observed treatment effect is consistent with subjects using risk-dominance as a selection criterion. In Round 20, subjects are more than twice as likely to play a truncation strategy when truncation is risk-dominant (TT: 25%, TR: 54%). At the same time, subjects are more likely to report their true preferences when truth-telling is risk-dominant (TT: 64%, TR: 46%). Both of these treatment differences are statistically significant at conventional levels (truth-telling: two-sided t-test, $p = 0.05$; truncation: two-sided t-test, $p < 0.01$).

By session

An alternative way to measure a treatment effect is to consider experimental sessions (i.e., cohorts) as independent units of observation and rank the sessions by their average frequencies of different strategies. This procedure is shown in Table 5. The three TT sessions have higher average truth-telling rates than the three TR sessions, while the three TR sessions have higher average truncation rates than the three TT sessions. Using a non-parametric

[28]To reduce experimenter demand effects, the terminology of preferences is never used in the experiment. A subject's true preference list is referred to as a "list of payments" and a subject's reported preference list is referred to as a "message."

Treatment	Frequency of Truth-Telling	Rank Score
TT	0.62	6
TT	0.60	5
TT	0.57	4
TR	0.52	3
TR	0.51	2
TR	0.46	1
Treatment	Frequency of Truncation	Rank Score
TR	0.47	6
TR	0.41	5
TR	0.36	4
TT	0.25	3
TT	0.19	2
TT	0.16	1
Treatment	Frequency of Permutation	Rank Score
TT	0.25	6
TT	0.22	5
TT	0.15	4
TR	0.12	3
TR	0.08	2
TR	0.08	1

Table 5: Ranking average frequencies of different strategies by session. TT: truth-telling is risk-dominant. TR: truncation is risk-dominant.

rank-sum test, we can reject the null hypothesis of no treatment difference for all three cases ($p = 0.05$). Truth-telling and truncation are both played significantly more often when they are the risk-dominant prediction.

By subject

Each subject submits 20 rank-order lists during the course of the experiment (one in each experimental round). From this data, we can calculate a truth-telling, truncation, and permutation rate for each individual subject. The average subject-level truth-telling rates are 0.59 and 0.49 for the TT and TR treatments, respectively (two-sided t-test, $p = 0.11$). The average subject-level truncation rates are 0.20 and 0.42 for the TT and TR treatments, respectively (two-sided t-test, $p < 0.01$).

Figure 4 shows the empirical cumulative distribution functions (CDFs) of subject-level behavior. When comparing the truncation CDFs, it is apparent that the first-order stochastic dominance relationship is consistent with the prediction of risk-dominance. In other words, the distribution of truncation rates from the TR treatment first-order stochastically dominates the distribution of truncation rates from the TT treatment (middle graph). We can also reject the null hypothesis that subject-level truncation rates have the same distribution function across treatments ($p < 0.01$, assessed with a Kolmogorov-Smirnov test). However, the first-order stochastic dominance relationship does not hold when comparing the truth-telling CDFs. Further, we fail to reject the null hypothesis that subject-level truth-telling rates from the two treatments come from the same theoretical distribution ($p = 0.10$, assessed with a Kolmogorov-Smirnov test).

Finally, we can use the CDFs to investigate the incidence of "purists" who always play a particular strategy. We find that very few subjects play the same strategy across all rounds of the experiment. Across both treatments, only 7% (8/120) of subjects consistently report their true preference list, 3% (3/120) of subjects consistently truncate their preference list, and 2% (2/120) of subjects consistently permute their preference list.

Result 2. *In both treatments, truncation strategies are played more often in later rounds of the experiment.*

We now investigate the effect of learning on subject behavior. Figure 5 shows the average frequencies of different strategies across rounds of the experiment. In the TT treatment, only 5% of subjects play a truncation strategy in Round 1 while 25% of subjects play a truncation strategy in Round 20. In the TR treatment, the corresponding numbers are 7% in Round 1 and 54% in Round 20. For both treatments, we reject the null hypothesis that

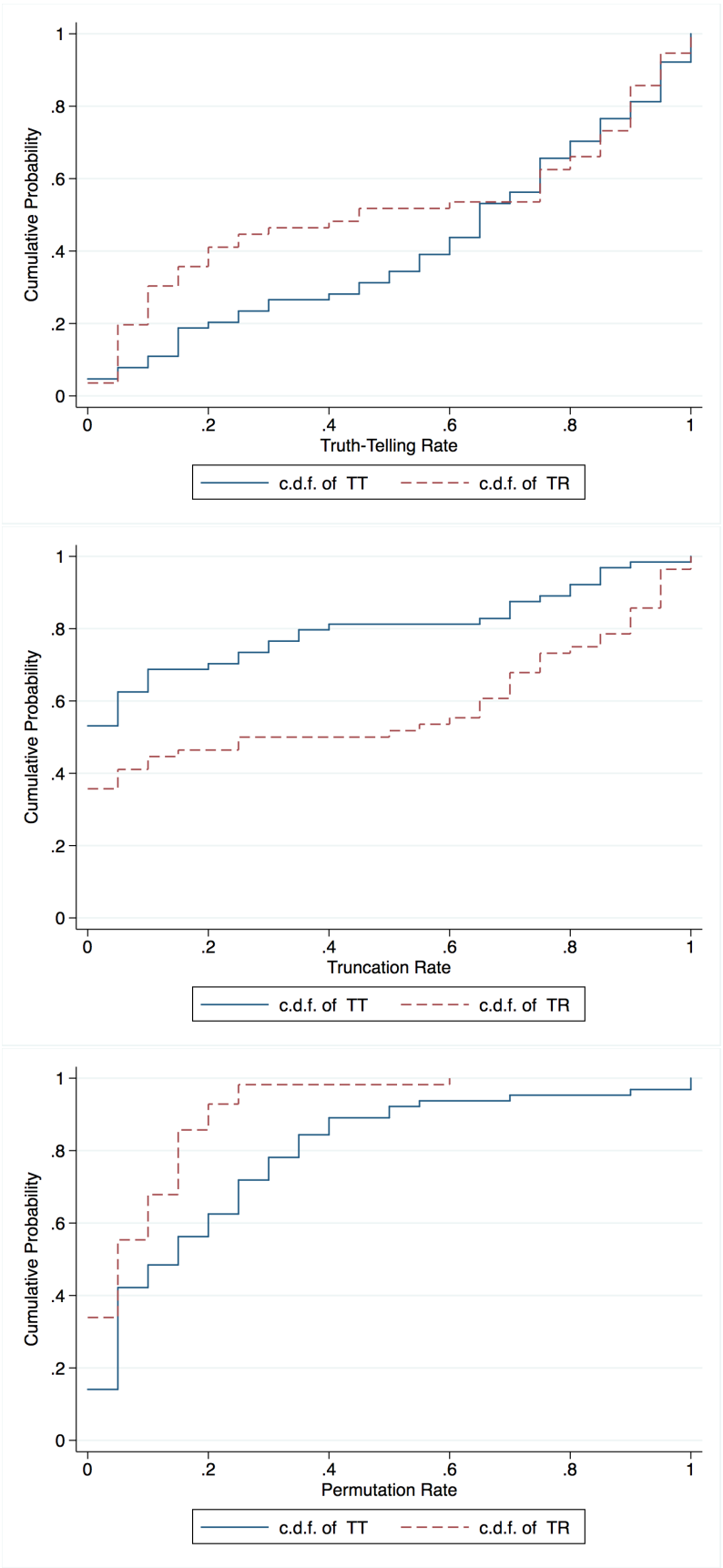


Figure 4: Empirical CDFs of subject-level behavior.
TT: truth-telling is risk-dominant. TR: truncation is risk-dominant.

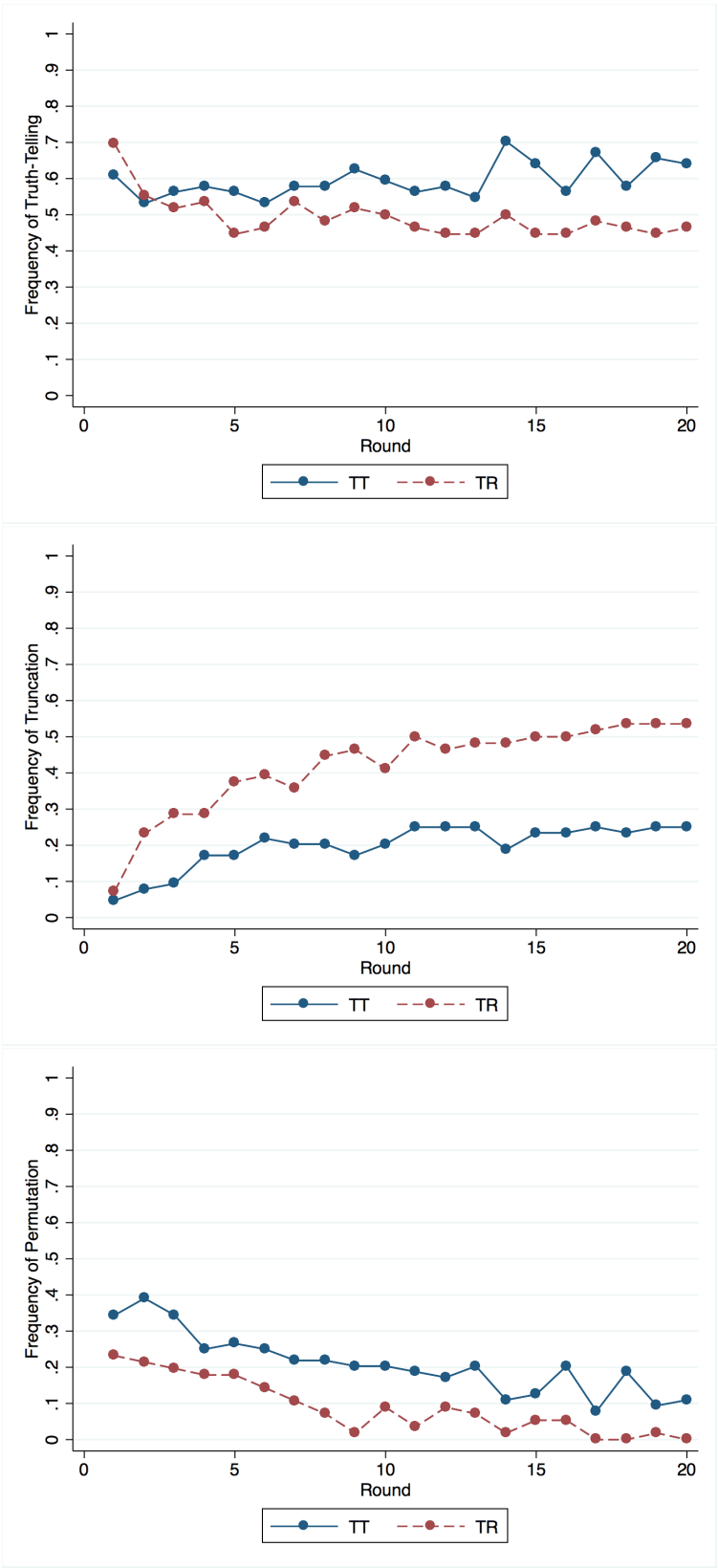


Figure 5: Average frequencies of different strategies across rounds of the experiment. TT: truth-telling is risk-dominant. TR: truncation is risk-dominant.

truncation rates across the initial 10 rounds are the same as truncation rates across the final 10 rounds (TT: $\chi^2(1) = 13.8$, $p < 0.01$; TR: $\chi^2(1) = 34.5$, $p < 0.01$). Since this truncation time trend is present in both treatments, where risk-dominance yields different predictions but truncation remains a necessary component of all payoff-dominant equilibria, it suggests that payoff-dominance has increasing salience as a selection criterion in later rounds of the experiment.[29]

We also see that truth-telling rates are remarkably consistent across the experiment. In the TT treatment, the average frequency of truthful reporting ranges from a minimum of 53% (in Round 2) to a maximum of 70% (in Round 14). In the TR treatment, the minimum and maximum frequencies are 45% (in several rounds) and 70% (in Round 1). As before, we test for whether truth-telling rates across the initial 10 rounds are significantly different from truth-telling rates across the final 10 rounds. For the TT treatment, we fail to reject the null hypothesis of no time trend ($\chi^2(1) = 2.03$, $p = 0.155$). For the TR treatment, we can reject the null hypothesis of no time trend ($\chi^2(1) = 4.63$, $p = 0.031$). However, the latter result is attributable to the high level of truth-telling in Round 1 and disappears when Round 1 data is removed from the analysis ($\chi^2(1) = 2.17$, $p = 0.140$).

With respect to permutation strategies, we find a decrease in the TT treatment from 34% in Round 1 to 11% in Round 20. In the TR treatment, although 23% of subjects report a permuted preference list in Round 1, there are no permutations in Round 20. Again, for both treatments, we reject the null hypothesis that permutation rates across the initial 10 rounds are the same as permutation rates across the final 10 rounds (TT: $\chi^2(1) = 28.9$, $p < 0.01$; TR: $\chi^2(1) = 41.2$, $p < 0.01$). The decay in preference-list permutation, which constitutes a weakly dominated strategy, is consistent with an increase in subjects' strategic sophistication over the course of the experiment.

Result 3. *We find limited evidence for protective behavior as a solution concept.*

Overall, truth-telling is the modal strategy in both treatments. A majority (59%) of submitted rank-order lists in the TT treatment and a plurality (49%) of submitted rank-order lists in the TR treatment correspond to subjects' true preferences. The prevalence of truthful behavior is consistent with the hypothesis that subjects use protective strategies. However, since truth-telling is the unique protective strategy in both treatments, theories in which subjects always use protective strategies would predict the absence of a treatment effect. As documented earlier, there is a significant treatment effect at the session level. In the

[29] In the context of stag hunt games, Rankin et al. (2000) also find that laboratory subjects focus on payoff-dominance rather than other solution concepts.

Final Outcome	Treatment TT	TR
Firm-Optimal	409 (64%)	193 (34%)
Worker-Optimal	191 (30%)	339 (61%)
Unstable	40 (6%)	28 (5%)
Total	640 (100%)	560 (100%)

Table 6: Empirical distribution of final outcomes.
TT: truth-telling is risk-dominant. TR: truncation is risk-dominant.

aggregate, we also find a statistically significant difference between the average truth-telling rate across treatments (two-sided t-test, $p < 0.01$). The higher rate of truth-telling in the TT treatment suggests that protective behavior fails to capture important aspects of the data.

Result 4. *The worker-optimal (firm-optimal) stable matching is more likely to be implemented when truncation (truth-telling) is risk-dominant.*

Table 6 presents data on final outcomes. First, unstable outcomes are rare in both treatments.[30] Second, it is clear that risk-dominance plays a role in selecting among stable outcomes. When truth-telling is risk-dominant, the firm-optimal stable matching is roughly twice as likely to be implemented (TT: 64%, TR: 34%). On the other hand, when truncation is risk-dominant, the worker-optimal stable matching is roughly twice as likely to be implemented (TR: 61%, TT: 30%). These treatment differences are statistically significant at conventional levels (two-sided t-test, $p < 0.01$ for both cases).

Result 5. *Subject behavior is consistent with the predictions of quantal response equilibrium.*

We now investigate how well subject behavior can be described by a theory of *quantal response equilibrium (QRE)*, a statistical model of equilibrium for normal-form games developed by McKelvey and Palfrey (1995). Unlike Nash equilibrium, in which players use best-response strategies with probability one, QRE assumes that players make decisions using a probabilistic choice model. Thus, QRE allows for every strategy to be played with non-zero probability, with the key feature that strategies yielding higher expected payoffs are played with higher probabilities. Since QRE is a more forgiving theory than Nash equilibrium, it has become a popular tool for explaining data from lab experiments (e.g., auctions,

[30]There is a unique action profile that yields an unstable matching: a subject-pair where one subject plays a truncation strategy and the other subject plays a permutation strategy.

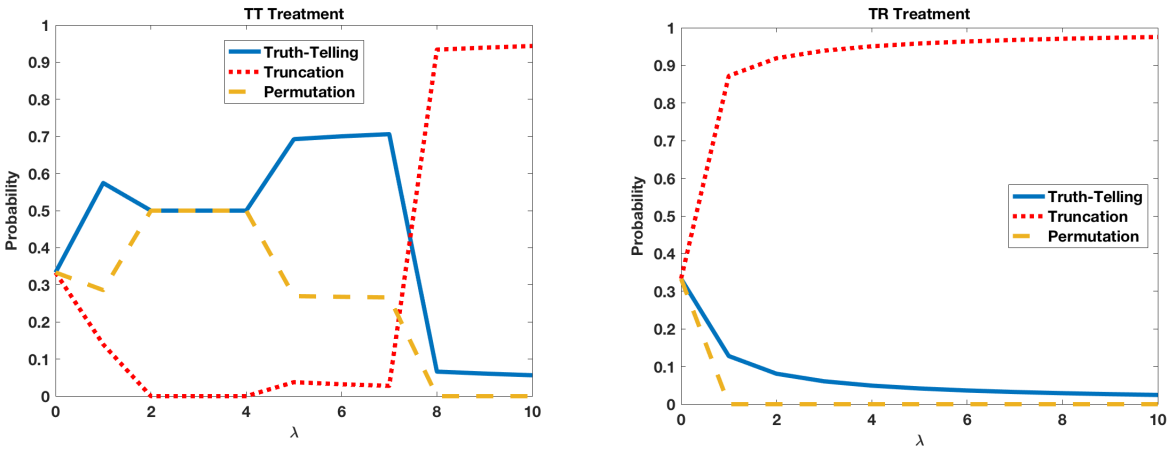


Figure 6: Logit equilibrium correspondence as a function of the rationality parameter λ. TT: truth-telling is risk-dominant. TR: truncation is risk-dominant.

bargaining, coordination games).[31] In our context, the QRE approach is of particular interest due to the subtlety of the strategic environment. In particular, the mapping from strategy profiles (i.e., reported preference lists) to outcomes (i.e., matchings) is less transparent than in a standard coordination game and thus subjects may not be immediately aware that the incentives they face mimic a standard coordination game.

A common parametric specification is the logistic quantal response function, where the error terms by which agents perceive expected payoffs have an extreme value distribution. In a logit equilibrium, the probability with which player i plays pure strategy $j \in \{1, 2, ..., J_i\}$ is given by

$$\pi_{ij} = \frac{e^{\lambda x_{ij}}}{\sum_{k=1}^{J_i} e^{\lambda x_{ik}}}$$

where x_{ij} is the expected payoff to player i from strategy j when all other players follow π_{ij}. The free parameter λ captures the level of rationality. As $\lambda \to 0$, the choice of strategy becomes purely random and the unique logit equilibrium involves playing each pure strategy with equal probability (i.e., $\pi_{ij} = \frac{1}{J_i}$ for all i, j). As $\lambda \to \infty$, the strategy with the highest expected payoff is chosen for sure and all limit points of logit equilibria are Nash equilibria.

However, tracing the graph of the logit equilibrium correspondence starting from the unique solution at $\lambda = 0$ yields a unique selection among the set of Nash equilibria. In this way, QRE can act as an equilibrium selection device. Figure 6 graphs the logit equilibrium correspondence for both treatments. At $\lambda = 0$, the logit equilibrium involves playing each

[31] For a recent survey of QRE methods and applications to economics, see Goeree et al. (2016).

	Treatment	
	TT	TR
λ	0.11	0.13
	(0.11)	(0.03)
Observations	64	56

Table 7: Maximum likelihood estimates of λ in the logit quantal response equilibrium model. Standard errors are shown in parentheses.
TT: truth-telling is risk-dominant. TR: truncation is risk-dominant

	TT		TR	
Strategy	Data	QRE	Data	QRE
Truth	0.64	0.40	0.46	0.33
Truncation	0.25	0.27	0.54	0.55
Permutation	0.11	0.33	0.00	0.12

Table 8: Comparison of Round 20 experimental data and QRE prediction ($\lambda = 0.12$).
TT: truth-telling is risk-dominant. TR: truncation is risk-dominant.

pure strategy with equal probability.[32] As $\lambda \to \infty$, it is apparent that the limiting logit equilibrium is the payoff-dominant equilibrium in truncation strategies.

We use maximum likelihood estimation to fit the free parameter λ to the experimental data. We estimate λ separately for each treatment using Round 20 data, when subjects have the most experience with the preference-revelation game. The maximum likelihood estimates are shown in Table 7.[33] Setting $\lambda = 0.12$, we conduct a comparative statics exercise to measure how changes in the treatment variable v_2 (i.e., the cardinal payoff from matching with the least preferred partner) affect the logit equilibrium. Figure 7 shows the logit equilibrium correspondence as a function of v_2. For both treatments, the Round 20 averages from the experimental data are also plotted (in crosshatch marks).[34] The exact figures can also be found in Table 8, which reproduces the Round 20 experimental data alongside the logit QRE prediction for $\lambda = 0.12$. With regard to truncation, the data matches the QRE prediction almost perfectly in both treatments. However, in the data, there is also *more* truth-telling and *less* permutation compared to the QRE prediction.

[32]For this analysis, we combine both permutation strategies. Hence, the probability of playing a particular strategy is $\frac{1}{3}$.

[33]Details of the maximum likelihood procedure are provided in the appendix.

[34]Recall that $v_2 = 15$ for the TT treatment and $v_2 = 5$ for the TR treatment.

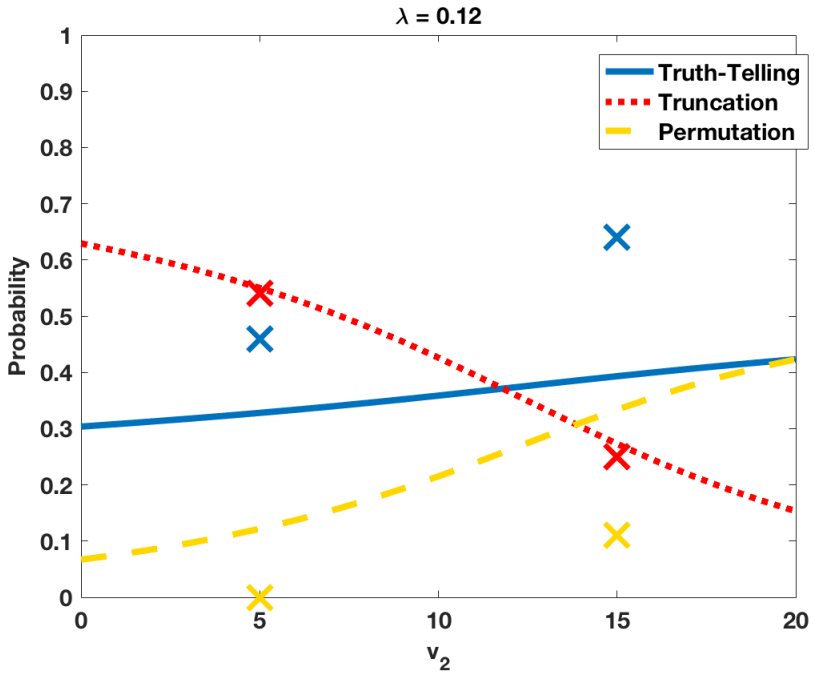


Figure 7: Logit equilibrium correspondence as a function of the treatment variable v_2.

5 Implications for Market Design

Our experiment can help shed light on several open questions in market design. In a seminal paper, Roth and Peranson (1999) conduct computational experiments using NRMP submitted rank-order lists from 1987 and 1993-1996. They find that only 0.1% of applicants would have received a different match from the applicant-proposing and hospital-proposing versions of the deferred acceptance algorithm. Assuming that the NRMP rank-order lists accurately reflect participants' true preferences, this exercise suggests that the applicant-optimal and hospital-optimal stable matchings coincide.[35] This result has profound implications for market design: if there is a unique stable matching in these environments, then there is no incentive for market participants to behave strategically.

However, the Roth and Peranson (1999) result depends crucially on the assumption of truthful preference reporting. By credibly documenting the use of non-truth-telling strategies, our experiment suggests a degree of caution in this regard. In our experimental markets, there are two disjoint stable matchings with respect to the true (i.e., induced) preferences. In our experimental data, however, 72% of markets have a unique stable matching with respect

[35]There is now a growing literature on "core convergence" in matching models (e.g. Immorlica and Mahdian, 2005; Kojima and Pathak, 2009; Lee, 2016). Under certain conditions, these papers show that the set of stable matchings shrinks as the size of the market increases.

Average Number of Applicants	2010	2012	2014	2016
Interviewed	85	89	96	94
Ranked	66	66	77	80

Table 9: NRMP Program Director Survey Reports (across all medical specialties).

to the reported preferences (TT treatment: 67%, TR treatment: 77%). In other words, the (mistaken) assumption of truthful preference reporting can make it appear as though a matching market has a unique stable matching when in fact there is a multiplicity of stable matchings. It is also noteworthy how robust this perception can be when confronted with reported preference data. In our experiment, any strategy profile in which at least one subject in a pair deviates from truth-telling is sufficient to produce a unique stable matching with respect to the reported preferences. Thus, a large range of reporting behavior is consistent with the empirical regularity of a small core span.

In addition, previous theoretical work on strategic behavior in matching markets has largely focused on truncation strategies (Coles and Shorrer, 2014; Roth and Rothblum, 1999). There are two common arguments for this emphasis. First, truncation can be profitably implemented even with incomplete information about other agents' true preferences or strategic uncertainty about other agents' reported preferences (Coles and Shorrer, 2014; Roth and Rothblum, 1999). Second, it is sufficient to restrict attention to truncation when considering the space of profitable misrepresentation strategies (Roth and Vate, 1991). In other words, any agent who can improve her match partner by deviating from truth-telling can do so by submitting a truncation of her true preferences.

However, the empirical content of truncation strategies remains an open question. Field data are insufficient to address this issue: while centralized matching clearinghouses may provide access to participants' submitted rank-order lists, participants' true preferences are unobserved. Our experiment provides preliminary support for the empirical relevance of truncation strategies and helps highlight the conditions that foster truncation behavior. Our results suggest that truncation should be more likely to arise in settings where participants either have a strong intensity of preference for top-ranked alternatives or have previous experience with the particular mechanism that is being used. Real-world matching protocols often more closely resemble one-shot games for many participants. In the context of the NRMP, however, hospitals usually do have considerable experience based on their participation in the match process in previous years. The NRMP conducts regular surveys of the directors of all hospital programs participating in the residency match.[36] Data from recent

[36]The NRMP Program Director Survey Reports can be found at the following website:

survey reports are shown in Table 9. In recent years, between 15-26% of interviewed applicants have not been included in hospitals' submitted rankings. While not conclusive, the NRMP survey data suggests that truncation strategies may play a role in field settings.

6 Conclusion

In this paper, we investigate the interplay between strategic uncertainty and equilibrium selection in the context of stable matching mechanisms. We report three main findings from a laboratory experiment: (1) truth-telling is the most common strategy, (2) both truth-telling and truncation are played more often when they are risk-dominant, and (3) in both treatments, truncation strategies are played more often in later rounds of the experiment. The final point suggests that the salience of payoff-dominance as a selection criterion increases with subject experience.

There are several promising avenues for further research. As it stands, there is no consensus on how to appropriately generalize concepts such as risk-dominance from 2×2 games to games with either additional players or larger action spaces. Theoretical work can hopefully bridge this gap. Further, the theory of two-sided matching is largely silent on the question of which stable matching will arise in markets with a multiplicity of stable matchings. Our experiment indicates that cardinal payoff differences can play a role in this selection process, but more empirical work is needed before a unified theory of selection among stable matchings can be developed.

http://www.nrmp.org/report-archives

References

Itai Ashlagi and Yannai A. Gonczarowski. Stable matching mechanisms are not obviously strategy-proof. *Working Paper*, 2016.

Itai Ashlagi and Flip Klijn. Manipulability in matching markets: conflict and coincidence of interests. *Social Choice and Welfare*, 39(1):23–33, 2012.

Salvador Barberà and Bhaskar Dutta. Protective behavior in matching models. *Games and Economic Behavior*, 8(2):281–296, 1995.

Pedro Dal Bó and Guillaume R Fréchette. The evolution of cooperation in infinitely repeated games: Experimental evidence. *The American Economic Review*, 101(1):411–429, 2011.

Colin Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press, 2003.

Marco Castillo and Ahrash Dianat. Truncation strategies in two-sided matching markets: Theory and experiment. *Games and Economic Behavior*, 2016.

Yan Chen and Tayfun Sönmez. School choice: an experimental study. *Journal of Economic Theory*, 127(1):202–231, 2006.

Peter A Coles and Ran I Shorrer. Optimal truncation in matching markets. *Games and Economic Behavior*, 87:591–615, 2014.

Russell W Cooper, Douglas V DeJong, Robert Forsythe, and Thomas W Ross. Selection criteria in coordination games: Some experimental results. *American Economic Review*, 80(1):218–233, 1990.

Tingting Ding and Andrew Schotter. Matching and chatting: An experimental study of the impact of network communication on school-matching mechanisms. *Games and Economic Behavior*, 2016.

Lester E Dubins and David A Freedman. Machiavelli and the gale-shapley algorithm. *American Mathematical Monthly*, 88(7):485–494, 1981.

Federico Echenique, Alistair J Wilson, and Leeat Yariv. Clearinghouses for two-sided matching: An experimental study. *Quantitative Economics*, 2016.

Clayton R Featherstone and Eric Mayefsky. Why do some clearinghouses yield stable outcomes? experimental evidence on out-of-equilibrium truth-telling. *Working Paper*, 2015.

Clayton R Featherstone and Muriel Niederle. School choice mechanisms under incomplete information: An experimental investigation. *Working Paper*, 2014.

Urs Fischbacher. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178, 2007.

Daniel E Fragiadakis and Peter Troyan. Designing mechanisms to make welfare-improving strategies focal. *Working Paper*, 2015.

David Gale and Marilda Sotomayor. Ms. machiavelli and the stable matching problem. *American Mathematical Monthly*, 92(4):261–268, 1985.

Jacob K Goeree, Charles A Holt, and Thomas R Palfrey. *Quantal Response Equilibrium: A Stochastic Theory of Games*. Princeton University Press, 2016.

Glenn W Harrison and Kevin A McCabe. Stability and preference distortion in resource matching: An experimental study of the marriage market. *Research in Experimental Economics*, 8, 1989.

John C Harsanyi and Reinhard Selten. *A general theory of equilibrium selection in games*. MIT Press, 1988.

Avinatan Hassidim, Deborah Marciano-Romm, Assaf Romm, and Ran I Shorrer. "strategic behavior in a strategy-proof environment. *Working Paper*, 2015.

Nicole Immorlica and Mohammad Mahdian. Marriage, honesty, and stability. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 53–62. Society for Industrial and Applied Mathematics, 2005.

Flip Klijn, Joana Pais, and Marc Vorsatz. Preference intensities and risk aversion in school choice: A laboratory experiment. *Experimental Economics*, 16(1):1–22, 2013.

Fuhito Kojima. Risk-dominance and perfect foresight dynamics in n-player games. *Journal of Economic Theory*, 128(1):255–273, 2006.

Fuhito Kojima and Parag A Pathak. Incentives and stability in large two-sided matching markets. *American Economic Review*, 99(3):608–627, 2009.

SangMok Lee. Incentive compatibility of large centralized matching markets. *Working Paper*, 2016.

Shengwu Li. Obviously strategy-proof mechanisms. *Working Paper*, 2016.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Stephen Morris, Rafael Rob, and Hyun Song Shin. p-dominance and belief potential. *Econometrica*, pages 145–157, 1995.

Joanna Pais and Ágnes Pintér. School choice and information: an experimental study on matching mechanisms. *Games and Economic Behavior*, 64(1):303–328, 2008.

Frederick W Rankin, John B Van Huyck, and Raymond C Battalio. Strategic similarity and emergent conventions: Evidence from similar stag hunt games. *Games and Economic Behavior*, 32(2):315–337, 2000.

Alex Rees-Jones. Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. *Working Paper*, 2016.

Alvin E Roth and Elliott Peranson. The redesign of the matching market for american physicians: Some engineering aspects of economic design. *American Economic Review*, 89 (4):748–780, 1999.

Alvin E Roth and Uriel G Rothblum. Truncation strategies in matching marketsin search of advice for participants. *Econometrica*, 67(1):21–43, 1999.

Alvin E Roth and Marilda A Oliveira Sotomayor. *Two-sided matching: A study in game-theoretic modeling and analysis*. Cambridge University Press, 1992.

Alvin E Roth and John H Vande Vate. Incentives in two-sided matching with random stable mechanisms. *Economic theory*, 1(1):31–44, 1991.

John B Van Huyck, Raymond C Battalio, and Richard O Beil. Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review*, 80(1):234–248, 1990.

Appendix: Maximum Likelihood Estimation

We illustrate the procedure for the TR treatment ($v_1 = 20$, $v_2 = 5$, $v_3 = 0$). Since we restrict attention to Round 20 data, there are 56 independent observations. An observation consists of subject i choosing an action $a_i \in \{1, 2, 3\}$ in Round 20 of the experiment, where $a_i = 1$ corresponds to truth-telling, $a_i = 2$ corresponds to truncation, and $a_i = 3$ corresponds to permutation. In the experimental data, the Round 20 choice probabilities are $\sigma_1 = 0.46$, $\sigma_2 = 0.54$, and $\sigma_3 = 0$. Thus, the expected payoffs are as follows:

$$u_1 = 5 + 15\sigma_2 = 13.1$$

$$u_2 = 20(\sigma_1 + \sigma_2) = 20$$

$$u_3 = 5$$

In a logit quantal response equilibrium, each player's choice probabilities must satisfy the following equation:

$$\sigma_j = \frac{e^{\lambda u_j}}{e^{13.1\lambda} + e^{20\lambda} + e^{5\lambda}}$$

for $j = 1, 2, 3$ and $\lambda \geq 0$, where λ is a free parameter capturing players' rationality. The likelihood function for the observed sample $a = (a_1, a_2, \ldots, a_{56})$ is then

$$L(a, \lambda) = \prod_{i=1}^{56} \prod_{j=1}^{3} \Big(\frac{e^{\lambda u_j}}{e^{13.1\lambda} + e^{20\lambda} + e^{5\lambda}} \Big)^{\mathbb{1}_{\{a_i = j\}}},$$

with an associated log-likelihood function given by

$$\ln L(a, \lambda) = \sum_{i=1}^{56} \sum_{j=1}^{3} \mathbb{1}_{\{a_i = j\}} \Big(\lambda u_j - \ln(e^{13.1\lambda} + e^{20\lambda} + e^{5\lambda}) \Big).$$

We use the ml routine in STATA to maximize the log-likelihood function with respect to λ, which yields an estimate of $\widehat{\lambda} = 0.13$ with a standard error of 0.03.

WELCOME!

Please turn off all electronic devices and place them in your bag or under your desk.

Throughout the experiment, please do not talk to anybody else and please remain silent at all times. If you have any questions, raise your hand and the experimenter will come to personally assist you.

Thank you for participating in this experiment. By showing up on time, you have automatically earned a $7 payment. If you follow the instructions carefully and make good decisions, you can earn additional money. The amount of money that you ultimately earn in this experiment depends on your decisions and the decisions of others. This session will last approximately one hour. At the end of the session, you will be paid privately in cash.

The experiment will be run entirely on the computer and all interactions between yourself and others will take place via the computer terminal. There are a total of 20 rounds in this experiment. At the end of the experiment, one of the 20 rounds will be randomly selected and your monetary payment will be determined based on the outcome of that round. Each of the 20 rounds is equally likely to be selected. Thus, it is in your best interest to take each round seriously. Each round is self-contained: your decisions in one round will not affect your opportunities or earnings in another round.

This experiment is about matching. There are two groups in the experiment: firms and workers. You will be randomly assigned to the role of either WORKER A or WORKER B. Your role will remain the same across all 20 rounds of the experiment. In each round, you will be randomly and anonymously paired with a worker who is assigned to the other role. In other words, if your role is WORKER A then you will be paired with someone in the role of WORKER B in each round (and vice versa).

Your goal in each round is to match with one of the two firms: FIRM A or FIRM B. The roles of the firms are computerized. They are programmed to behave in a certain way, which will be discussed in more detail shortly. You will earn different payments from different matches. The payments corresponding to all possible matches will be shown to you on your screen. There will be four lists on your screen (one for each firm and one for each worker). A firm's list contains two workers and it shows the order in which the firm will be making offers to match with the different workers. A worker's list contains two firms and it shows how much money the worker would earn if matched with the different firms. These lists will remain the same across all 20 rounds of the experiment.

As a worker, if you are matched with the firm in the first position of your list in a given round, you will earn a payment of $20 for that round. If you are matched with the firm in the second position of your list in a given round, you will earn a payment of $15 for that round. Remaining unmatched will result in a payment of $0 for that round.

To determine which firm you are matched with, you will send a message in each round. This message is sent to the computer. It is very important that you understand what a message is since the messages sent by both you and the other worker in your pair determine which firm you are matched with. A message is a ranking of the firms. The message may or may not include all the firms. Thus, although there are two firms you could potentially be matched with, your message can contain either one or two firms.

The way the computer uses the messages to determine which firm you are matched with will be explained below. The computer will go through these steps on its own and you will not observe this process in each round. Instead, you will only see which firm you are matched with and how much money you have earned at the end of each round. However, we will go through these steps so you understand how the matches are calculated.

Before we go through the procedure in detail, we will summarize the main ideas. Essentially, the computer uses the message you submit to decide which firms' offers to accept and reject on your behalf. There are two rules that describe this process.

1. The computer never matches you with a firm that you have not included in your submitted message. This is because, even if that firm makes an offer to match with you, the computer will reject that offer.

2. The computer always matches you with the highest ranked firm (according to your submitted message) that has made you an offer. If you receive an offer from only one firm and that firm is included in your message, then the computer will accept that offer. If you receive offers from both firms and both firms are included in your message, then the computer will accept the offer from the firm that you ranked higher in your message and reject the other offer.

STEP 0: All firms and workers are unmatched. Workers (YOU) send messages to the computer. These messages are a ranking of the firms that the computer will use in the steps below.

NOTE: THE REMAINING STEPS ARE PERFORMED BY THE COMPUTER

STEP 1: Firms propose offers.

Each firm makes an offer to the first worker on its list.

STEP 2: Workers respond to offers.

(a) If a worker receives no offers, then nothing changes. The worker remains unmatched.

(b) If a worker receives one offer, then the computer uses the message of that worker to decide whether or not to accept the offer. For example, suppose that WORKER A receives an offer from FIRM A.

- If WORKER A included FIRM A in its message, then WORKER A is matched with FIRM A.

- If WORKER A did not include FIRM A in its message, then WORKER A is not matched with FIRM A. In this case, WORKER A "rejects" FIRM A.

(c) If a worker receives two offers, then the computer uses the message of that worker to decide which firm to match that worker to. For example, suppose that WORKER A receives offers from both FIRM A and FIRM B.

- If WORKER A included FIRM A in its message but not FIRM B, then WORKER A is matched with FIRM A. WORKER A rejects the offer from FIRM B.

- If WORKER A included FIRM B in its message but not FIRM A, then WORKER A is matched with FIRM B. WORKER A rejects the offer from FIRM A.

- If WORKER A included both FIRM A and FIRM B in its message, then the computer looks at the relative positions of FIRM A and FIRM B in WORKER A's message. If WORKER A's message ranks FIRM A in a higher position than FIRM B, then WORKER A is matched with FIRM A and WORKER A rejects the offer from FIRM B. If WORKER A's message ranks FIRM B in a higher position than FIRM A, then WORKER A is matched with FIRM B and WORKER A rejects the offer from FIRM A.

STEP 3: Unmatched firms propose new offers.

Each firm that is unmatched makes an offer to the second worker on its list.

STEP 4: Workers respond to offers.

(a) If a worker is unmatched, then refer to STEP 2 to determine how the worker decides among offers.

(b) If a worker is currently matched and receives no new offers, then nothing changes. The worker remains matched to whichever firm they were already matched with.

(c) If a worker is currently matched and receives a new offer, then the computer looks at the relative positions of the current match and the new offer in the worker's message. For example, suppose that WORKER A is currently matched with FIRM A and receives a new offer from FIRM B.

- If WORKER A did not include FIRM B in its message, then WORKER A rejects the offer from FIRM B. WORKER A is still matched with FIRM A.

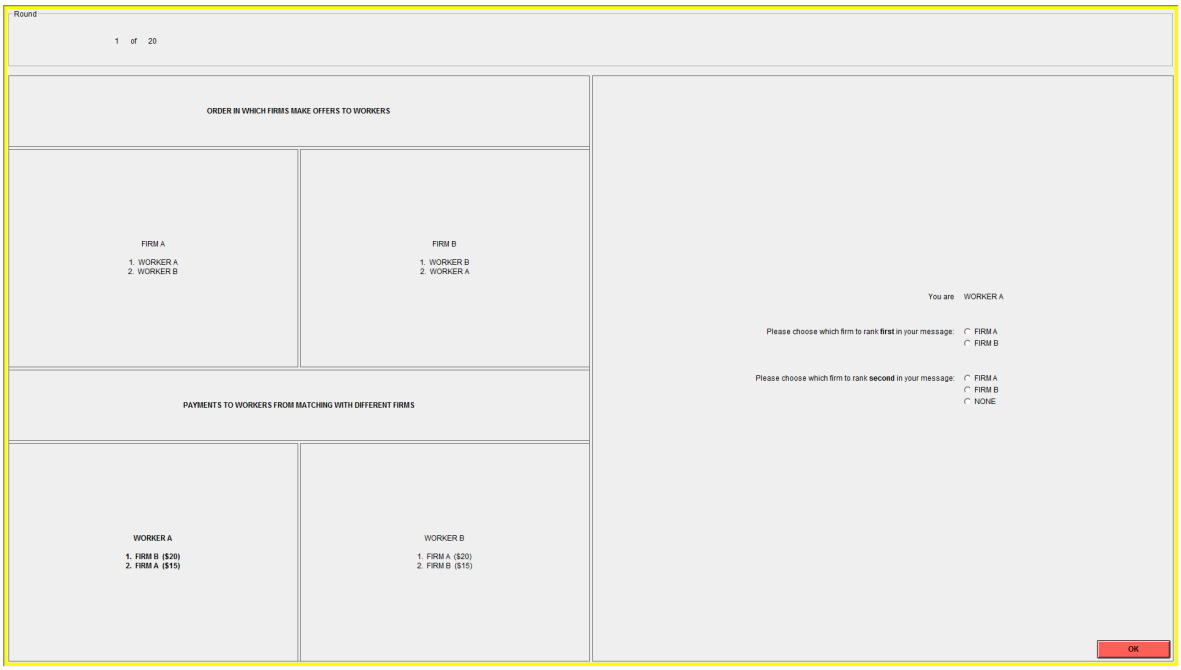
- If WORKER A included FIRM B in its message, then the computer looks at the relative positions of FIRM A and FIRM B in WORKER A's message. If WORKER A's message ranks FIRM A in a higher position than FIRM B, then WORKER A is still matched with FIRM A and WORKER A rejects the offer from FIRM B. If WORKER A's message ranks FIRM B in a higher position than FIRM A, then WORKER A's previous match with FIRM A is broken and WORKER A is now matched with FIRM B.

-
-
-

The procedure continues in this fashion until there are no firms left to make offers. This can happen for two reasons: either both firms are already matched or there is an unmatched firm that has already been rejected by both workers. The final matches for a given round are the matches that are in place when the procedure ends. It is only the final matches that count to determine payments. Matches that are made and then broken do not count for payments.

Note that the computer does not consider your list of payments when deciding which firm to match you with. The computer only uses the rankings from your submitted message (and the submitted message of the other worker you are paired with) to calculate the final matching. Once you are matched to a firm, then your list of payments is used to determine how much money you have earned in that round.

The experimental interface is shown below. The bar at the top of the screen indicates which round the players are currently in. The left hand side of the screen displays the lists for the firms and the workers. Your own list of payments will always be in bold. The right hand side of the screen displays your role in the experiment and asks you to submit a message.



At the beginning of the experiment, there will be a brief demonstration of the procedure that the computer uses to determine the final matchings. You will walk through the steps discussed above to better understand how the messages that are submitted affect which firm you are matched with. Again, keep in mind that you will not have to go through a similar process during the actual experiment. In the experiment, the only action that you will take is to submit a message. The computer will go through these steps on its own to determine which firm you are matched with and it will then report that information to you. The purpose of the example is just to show you in detail the steps the computer is taking to determine the final matchings based on the submitted messages.

To summarize, the order of events in the experiment is as follows:

1. You will go through an example demonstrating the procedure that the computer uses to calculate the final matchings.

2. You are randomly assigned to the role of either WORKER A or WORKER B. Your role will remain the same across all 20 rounds.

3. You learn your payments (as well as the payments of the other worker) for all possible matches. You also learn the order in which the firms will make offers to match with the workers. This information will remain the same across all 20 rounds.

4. You are randomly and anonymously paired with a worker in the other role.

5. You submit a message to the computer which is a ranking of the firms. This ranking can contain either one or two firms.

6. The computer uses the submitted messages to calculate the final matches.

7. The computer reports to you which firm you are matched with and how much money you have earned.

8. You will repeat steps 4-7 a total of 19 times (since there are 20 rounds in the experiment).

9. At the end of the experiment, one of the 20 rounds will be randomly selected and you will be paid your earnings for that round (in addition to the $7 show-up payment). All payments will be made privately and in cash.

If you have any questions at this point, please raise your hand. If not, we will begin the experiment shortly.

Good luck!