On the Selection of Arbitrators

By Geoffroy de Clippel, Kfir Eliaz and Brian Knight *

A key feature of arbitration is the possibility for conflicting parties to participate in the selection of the arbitrator, the individual who will rule the case. We analyze this problem of the selection of arbitrators from the perspective of implementation theory. In particular, theoretical analyses document problems with veto-rank, a simultaneous procedure that is commonly used in practice, and develop a new sequential procedure, shortlisting, with better properties. Experimental results are consistent with the theoretical predictions, highlighting both the disadvantages associated with the veto-rank procedure and the advantages associated with the shortlising procedure.

Implementation theory studies the design of institutions and procedures for collective decision-making. It aims to find ways of incentivizing participants to select "desirable" outcomes. What is deemed "desirable" varies across situations, and is represented by a social choice rule (SCR) that maps the participants' preferences to subsets of feasible outcomes. When applied to concrete economic environments, this theory helps address a number of important questions. Do prevalent procedures implement the intended SCR? Are there alternative mechanisms? Are there acceptable variants of the SCR that are implementable? How do alternative mechanisms perform when tested with participants facing real stakes? These questions have been studied in a wide variety of contexts including auctions, the provision of public goods, kidney exchange, school choice and choice of medical residency (see the studies surveyed in Kagel (1995), Chen (2008), Roth (2002, 2007) and Kagel and Levin (2011)).

We contribute to this literature by applying implementation theory to a rich class of situations in which individuals must agree on a collective decision, and where monetary transfers are not available. This class includes elections of public officials, committee decisions, selection of committee members, selection of juries for a trial, selection of judges for an appellate court, etc.

In this paper, we focus on a specific problem within this general class: the selection of arbitrators. For several reasons, this problem is both tractable and

* de Clippel: Brown University, Department of Economics, 64 Waterman Street, Providence, RI 02912, declippel@brown.edu. Eliaz: Tel Aviv University and the University of Michigan, Departments of Economics, Tel Aviv, Israel, and Ann Arbor, MI, kfire@post.tau.ac.il. Knight: Brown University, Department of Economics, 64 Waterman Street, Providence, RI 02912, brian_knight@brown.edu. We wish to thank Eli Zvuluni of Possible Worlds Inc. for programming the experiment, Melis Kartal and Mark Bernard for running the experiment, CESS at NYU and especially Caroline Madden for invaluable administrative help, Samuel Mencoff, Pantelis Solomon, Ee Cheng Ong and especially Neil Thakral for exceptional research support. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

interesting from an implementation perspective. First, contrary to problems involving committees or a large number of voters, arbitrator selection involves only two parties. Second, contrary to jury selection, which involves the selection of a panel of individuals, the final outcome involves the selection of a single individual, the arbitrator. Third, selecting an arbitrator is a case where the assumption of complete information, which underlies many theoretical models, is reasonable. Indeed, most disputes resolved through arbitration occur between parties that have a long-term relationship (e.g., unions and management). In addition, the arbitration agencies provide both parties with the same information about the potential arbitrators. Fourth, it is reasonable to assume that the parties do not necessarily have completely opposed rankings of all arbitrators. This is because arbitrators differ in their fees, their expertise, their past rulings and their delays in reaching a decision.

In addition to being tractable, the problem of selecting an arbitrator is of practical relevance. Arbitration is the most common procedure for resolving disputes without resorting to costly litigation. Having a role in choosing who will rule the case is often cited by participants as one of its main attractive features. Indeed parties dislike facing the risk of being subject to a judge who is not qualified for the case or who is perceived as biased. Hence, the relative appeal of arbitration agencies depends on their ability to assign arbitrators to cases in a way that best reflects the preferences of both parties.

This paper evaluates selection mechanisms based upon two factors: (*i*) a "theoretical" criterion - every equilibrium induced by the mechanism has normatively appealing properties (which we describe shortly), and (*ii*) an "empirical" criterion - when the mechanism is actually carried out with real incentives, it is likely to generate outcomes that satisfy the desired properties. We compare a commonly used simultaneous mechanism and a simple sequential mechanism that we developed and is not currently used in practice. We argue that the sequential mechanism is superior to the simultaneous mechanism according to both criteria.

Our analysis proceeds as follows. We first consider the commonly used procedure for assigning arbitrators, *the Veto-Rank mechanism* (VR).[1] Under this mechanism, two parties receive a list of n (an odd number) potential arbitrators. Each party independently vetoes or removes $\frac{n-1}{2}$ names from the list, and ranks the remaining $\frac{n+1}{2}$ candidates. The selected arbitrator is one with the minimal sum of ranks among candidates who have not been vetoed (ties are resolved via a lottery).

The veto-rank mechanism is appealing *if participants are truthful*, i.e., if they veto their bottom $\frac{n-1}{2}$ candidates and rank the remaining ones truthfully. Specifically, the resulting SCR satisfies two appealing properties: the appointed arbitrator is Pareto efficient and Pareto dominates both parties' median choices (a 'minimal satisfaction' test). However, truth-telling is not always a Nash equilib-

[1]The Supplementary Appendix contains a list of major arbitration agencies that use the veto-rank mechanism to select arbitrators.

rium, hence, participants may strictly gain by deviating from truthful behavior.[2] Therefore, actual outcomes may end up violating the above appealing properties. We argue that these concerns apply to *all* simultaneous mechanisms, not just VR. Indeed, Proposition 1 establishes that there is no simultaneous mechanism that Nash implements a SCR that selects Pareto efficient outcomes, which Pareto dominates the parties' median choices. In particular, the SCR derived from truthtelling in VR is *not* Nash implementable.

Given the potential problems with simultaneous mechanisms, we turn our attention to sequential mechanisms. As shown in the implementation literature, more SCRs can be implemented using extensive-form mechanisms and the subgame-perfect equilibrium notion (see Abreu and Sen (1990)). However, Proposition 2 establishes that the SCR induced by truthful reporting in VR − or any selection of it − is *not* subgame-perfect implementable. This result suggests that combining vetoes with utilitarian-like criteria of minimizing the sum of ranks cannot be implemented by any mechanism - simultaneous or sequential - using Nash or subgame-perfect Nash equilibrium.

In light of these negative results, we consider sequential procedures of perfect information that satisfy two desiderata: (*i*) backwards induction leads to a Pareto efficient outcome, which Pareto dominates both parties' median choices, and (*ii*) there are as few stages as possible so that backwards induction is relatively "simple" to execute (see Binmore et. al., 2002). Proposition 3 establishes that only one procedure - referred to as Shortlisting (SL) - satisfies both criteria: One party starts the game by selecting $\frac{n+1}{2}$ candidates, and the second party then selects the arbitrator out of that shortlist.

The relative performance of VR and SL is then measured in a controlled lab experiment for several preference profiles. Results document that non-truthful behavior occurs under VR in a majority of cases, a significant proportion of which is driven by some strategic motives. Moreover, *SL, which is not used in practice, outperforms the commonly used VR mechanism.*

The paper unfolds as follows. After discussing the related literature, section 2 contains theoretical results (proofs are relegated to the appendix). The experimental design and data analysis are available in Section 3. The concluding section summarizes our findings.

Related Literature

The most closely related paper is Bloom and Cavanagh (1986a), who analyze the selection of arbitrators using data on arbitration cases from the New Jersey Public Employment Relations Commission (PERC) during 1980. Data are based upon the simultaneous veto-rank scheme described in the Introduction (with $n = 7$). Their analysis first examines the degree of overlap between rankings in order to shed light on the similarity of preferences. They show some, but not complete,

[2]If an arbitrator is commonly known among parties to be unqualified for the case, for example, why waste a veto on him if one believes that the other party will veto him?

overlap in rankings, and, under the assumption of sincere rankings, conclude that there is some, but not complete, overlap in preferences. We reach the same conclusion (see online Appendix), but without assuming that parties are truthful in their reports.

Their second analysis uses rankings and characteristics of arbitrators to measure the degree to which certain characteristics are valued by the different parties. They find, for example, that employers rank economists more highly than unions do. Under an assumption of sincere rankings, one can conclude that employers have a relative taste for economists and that unions have a distaste for economists. The assumption of sincere rankings is debatable though, and indeed we present theoretical and experimental evidence that it does not hold. Bloom and Cavanagh try to address this issue by fitting their model under the weaker assumption strategic players always rank their most preferred alternative first but may strategize on other dimensions of their report. They observe that their preference parameter estimates do not vary much when using only the first choice data, and conclude from it that there is no evidence of strategic play. A key limitation of this test involves the breakdown of the assumption that strategic players always rank their most preferred alternative first. It is straightforward to generate counter-examples to this: if the union vetoes the first choice of the employer, the employer may choose to not rank their most preferred alternative first as this is "wasting" the first ranking. Our experiment, presented in Section 3, confirms that a substantial fraction of players do not rank first their most preferred alternative when it is not viable, in the sense of being the worst for their opponents.

In an unpublished working paper, Bloom and Cavanagh (1986b) theoretically analyze the VR mechanism and show that it has non-truthful and inefficient equilibria. They also show that if the parties held uniform priors over all the possible strict rankings of arbitrators, then being truthful is an efficient Bayesian Nash equilibrium in both mechanisms.[3] Our focus, however, is on the implementation-theoretic view of arbitrator selection. In particular, we show that a large class of SCRs with appealing properties is impossible to implement, while alternative SCRs are implementable by "natural" mechanisms.

More generally, the present paper is related to a literature on matching, where economists have identified market failures and proposed new mechanisms that solve these failures. Several of these mechanisms, similarly to the veto-rank scheme used in selecting arbitrators, involve participants submitting rank-ordered preferences. Examples include mechanisms for matching residents to hospitals and students to elementary schools (see Roth (1984, 2007), Abdulkadiroglu, Pathak, and Roth (2005), and Abdulkadiroglu et al. (2005)). This literature has focused on implementing strategy-proof mechanisms using variants of the Gale-Shapley

[3]One complication that arises when analyzing Bayesian Nash equilibria, especially in the veto-rank game, is that one needs to make assumptions about each player's belief about his opponent's preferences over *lotteries*. This concern, however, is not discussed in their paper.

deferred acceptance algorithm or the top-trading cycle mechanism. In the context of the selection of arbitrators, strategy-proofness leads to a dictatorial result and we show that Nash implementation of desirable SCRs is impossible. Therefore, we turn to sequential mechanisms and subgame perfection.

Given our focus on whether participant ranks and vetoes are sincere or strategic, this paper is also related to a literature on strategic voting, which can take many forms. In an experimental setting with three candidates and plurality rule, Forsythe, Myerson, Rietz, and Weber (1993 and 1996) find substantial evidence that voters are strategic in the sense of not voting for their most preferred candidate when this candidate has little chance of winning. Focusing on the case of bundled elections, Degan and Merlo (2007) find little evidence that voters are strategic in the sense that they might account for the fact that policy outcomes may depend upon both the Congress and the President. In a model with incomplete information, Kawai and Watanabe (2013) estimate that a large fraction of voters in Japanese elections are strategic in the sense of conditioning on the state of the world where they are pivotal.

I. Theoretical Motivation

Two parties, $i = 1, 2$, face a finite set $\mathcal{A}$ of $n \geq 4$ candidates that an agency proposes as potential arbitrators. We assume that n is odd, as this is the scenario favored by arbitration agencies and studied in our experimental analysis (all the results in this section can be extended to the case where n is even). $\mathcal{P}$ denotes the set of all possible strict preference relations $\succ$ on $\mathcal{A}$. Most disputes resolved through arbitration occur between parties that have a long-term relationship (e.g., unions and managements). In addition, arbitration agencies provide both parties with the same detailed resumés of the potential arbitrators. Hence it is not unreasonable to assume that the parties' ordinal preferences are commonly known among them (put differently, we consider implementation under "complete information").

DEFINITION 1: *A* social choice rule *(SCR) is a correspondence* $f : \mathcal{P} \times \mathcal{P} \to \mathcal{A}$ *such that* $f(\succ_1, \succ_2)$ *is a non-empty subset of* $\mathcal{A}$*, for each* $(\succ_1, \succ_2) \in \mathcal{P} \times \mathcal{P}$.

DEFINITION 2: *A SCR* f *is* weakly implementable *if there exists a* mechanism $(\mathcal{S}_1, \mathcal{S}_2, \mu)$*, where* $\mathcal{S}_i$ *is* i*'s strategy set and* $\mu : \mathcal{S}_1 \times \mathcal{S}_2 \to \mathcal{A}$ *is the outcome function, such that, for each* $(\succ_1, \succ_2) \in \mathcal{P} \times \mathcal{P}$*, the set of pure-strategy Nash equilibrium outcomes associated to the strategic-form game* $(\mathcal{S}_1, \mathcal{S}_2, \mu, \succ_1, \succ_2)$ *is non-empty and a subset of* $f(\succ_1, \succ_2)$.

Notice that the veto-rank procedure discussed in the Introduction does not qualify as a mechanism in this sense, because the outcome function delivers a lottery in some circumstances. Considering lotteries, and thinking about how parties behave when facing such uncertainty, leads us to consider risk preferences. Let $\mathcal{U}$ be the set of strict Bernoulli functions (the defining ingredient of von

Neumann-Morgenstern preferences). A typical element u of $\mathcal{U}$ is thus simply a function $u : \mathcal{A} \to \mathbb{R}$, with $u(a) \neq u(a')$ whenever $a \neq a'$, and preferences between lotteries over $\mathcal{A}$ are derived by computing expected utility with respect to u. It is less plausible to think that there is complete information regarding these Bernoulli functions, but our analysis is robust against that assumption in that our sole objective when considering lotteries is to show that strong negative results hold *even if* there was complete information in that regard.

DEFINITION 3: *A* random social choice function *(RSCF) is a function* $\psi : \mathcal{U} \times \mathcal{U} \to \Delta(\mathcal{A})$ *that associates a lottery to each pair of strict Bernoulli functions.*

DEFINITION 4: *The RSCF* ψ *is* implementable *if there exists a* random mechanism $(\mathcal{S}_1, \mathcal{S}_2, \mu)$*, where* $\mathcal{S}_i$ *is* i*'s strategy set and* $\mu : \mathcal{S}_1 \times \mathcal{S}_2 \to \Delta(\mathcal{A})$ *is the outcome function, such that, for each* $(u_1, u_2) \in \mathcal{U} \times \mathcal{U}$*, any pure-strategy Nash equilibrium outcomes associated to the strategic-form game* $(\mathcal{S}_1, \mathcal{S}_2, \mu, u_1, u_2)$ *coincides with* $\psi(u_1, u_2)$.[4]

PROCEDURE 1: (*Veto-Rank*) (VR) The veto-rank procedure provides an example of random mechanism. Both parties ($i = 1, 2$) simultaneously choose a pair $(\mathcal{V}_i, r_i)$, where $\mathcal{V}_i$ is a set of vetoed options that contains $\frac{n-1}{2}$ elements from $\mathcal{A}$, and r_i is a scoring rule that assigns to every element in $\mathcal{A}\backslash\mathcal{V}_i$ an integer from zero to $n - k - 1$ such that no two elements are assigned the same score. From the set $\mathcal{A}\backslash(\mathcal{V}_1 \cup \mathcal{V}_2)$, the outcome is selected by maximizing the sum of scores, $r_1(\cdot) + r_2(\cdot)$, with ties being broken via a uniform lottery.

For each $a \in \mathcal{A}$ and each $u \in \mathcal{U}$, let $\sigma(a, u) = \#\{a' \in \mathcal{A} \mid u(a') < u(a)\}$. The veto-rank procedure is played truthfully if, for each $(u_1, u_2) \in \mathcal{U} \times \mathcal{U}$ and both $i = 1, 2$, the set $\mathcal{V}_i$ contains the $\frac{n-1}{2}$ worst elements according to u_i, and $r_i(a) = \sigma(a, u_i) - \frac{n-1}{2}$, for each element $a \in \mathcal{A} \setminus \mathcal{V}_i$. This generates the following natural RSCFs. For each $(u_1, u_2) \in \mathcal{U} \times \mathcal{U}$, $\psi_{VR}(u_1, u_2)$ will denote the uniform lottery defined over

$$\arg \max_{a \in X(u_1, u_2)} (\sigma(a, u_1) + \sigma(a, u_2))$$

where

$$X(u_1, u_2) = \{a \in \mathcal{A} | \sigma(a, u_i) \geq \frac{n-1}{2}, \text{ for } i = 1, 2\}.$$

The support of ψ_{VR} also defines a natural SCR: for each $(\succ_1, \succ_2)$,

$$f_{VR}(\succ_1, \succ_2) := support(\psi_{VR}(u_1, u_2)),$$

where u_i is any[5] strict Bernoulli function that is consistent with $\succ_i$ over $\mathcal{A}$.

[4] Implementable RSCFs are thus invariant to affine transformations of u_1 and u_2.

[5] Notice indeed that ψ_{VR} varies only with the ordinal information encoded in the Bernoulli functions.

We believe that the main reason why arbitration agencies aim to implement f_{VR} is that all the outcomes that emerge with positive probabilities satisfy the following two properties. A RSCF ψ is *Pareto efficient* if, for each (u_1, u_2) and each x in the support of $\psi(u_1, u_2)$, it is impossible to find $a \in \mathcal{A}$ such that $u_i(a) > u_i(x)$ for both $i \in \{1, 2\}$. It passes the *minimal satisfaction test* (MST) if $\sigma(x, u_i) \geq \frac{n-1}{2}$ for each $i \in \{1, 2\}$, each $u \in \mathcal{U} \times \mathcal{U}$, and each x in the support of $\psi(u_1, u_2)$. Similar definitions also apply to SCRs. The SCR f_{VR} and the RSCF ψ_{VR} are both Pareto efficient and both pass the minimal satisfaction test. However, the VR procedure need not lead to desirable outcomes.

Preliminary Observations. *The VR procedure has the following properties.*

(a) (non-truthfulness) *Truthtelling is not a Nash equilibrium for some preference profiles, and for every preference profile there is a (undominated) non-truthtelling Nash equilibrium.*

(b) (undesirable equilibria) *There are preference profiles for which the mechanism induces (undominated) Nash equilibrium outcomes not selected by f_{VR}, and which may even be Pareto inefficient.*

(c) (risk of miscoordination) *There are preference profiles for which there exists a pair of (undominated) equilibria, $s = (s_1, s_2)$ and $s' = (s'_1, s'_2)$, such that if both players coordinate on either s or s' the outcome is in f_{VR}, but if one player follows s and another follows s', the outcome is Pareto inefficient and/or violates the MST.*

These preliminary observations raise the questions of whether there exists another normal-form mechanism that implements the RSCF ψ_{VR}, or that weakly implements the SCR f_{VR}. The next proposition establishes a stronger result: any SCR (or RSCF) that satisfies Pareto efficiency and MST is not implementable.

PROPOSITION 1: *The following three statements hold.*

(a) There is no SCR that is weakly implementable, Pareto efficient, and that passes the MST.

(b) There is no RSCF that is implementable, Pareto efficient, and that passes the MST.

(c) In particular, ψ_{VR} is not implementable, and f_{VR} is not weakly implementable.

In view of this negative result, we turn our attention to mechanisms that have more structure, namely dynamic procedures, and investigate implementation in subgame-perfect equilibrium. Even though this implementation notion is much more permissive than Nash implementation (see Moore and Repullo (1988) or Abreu and Sen (1990)), the VR SCR remains impossible to implement. It does not even admit a selection that is subgame-perfect implementable.

DEFINITION 5: *A SCR f is* weakly subgame-perfect implementable *if there exists a dynamic mechanism such that, for each $(\succ_1, \succ_2) \in \mathcal{P} \times \mathcal{P}$, the set of pure-strategy subgame-perfect Nash equilibrium outcomes of the extensive-form game is non-empty and a subset of $f(\succ_1, \succ_2)$.*

PROPOSITION 2: *f_{VR} is not weakly subgame-perfect implementable.*

While implementing the Veto-Rank social choice rule is clearly out of reach, considering dynamic mechanisms makes it possible to guarantee Pareto efficiency and minimal satisfaction. There are in fact multiple SCRs with these properties that are weakly subgame-perfect implementable. This leaves us the possibility to add requirements. We add two desiderata to make the analysis more relevant in practice. First, we restrict attention to dynamic mechanisms of *perfect information*, meaning that both individuals know all previous moves when making decisions. A subgame-perfect Nash equilibrium can thus be computed simply by backward induction. Preferences being strict, backward induction always leads to a unique outcome, in which case weak and full subgame-perfect implementation coincide, and the risk of miscoordination is eliminated. Second, even though backward induction does simplify the computation of subgame-perfect Nash equilibria, it is well-documented that expecting participants to carry out backward induction may be unrealistic when the game involves multiple stages (see e.g. Binmore et al. (2002) and Levitt, List and Sadoff (2011)). Also, the epistemic conditions underlying backward induction become more restrictive as the game becomes longer. There are thus reasons to focus on short dynamic mechanisms. Adding this *behavioral constraint* leads to a natural question. Which SCRs meet the MST, are Pareto efficient, implementable by backward induction, and are such that it is impossible to find a shorter dynamic mechanism of perfect information whose backward induction outcome systematically meet these properties?

Note that dynamic mechanisms of perfect information must specify which individual assumes the role of the first-mover, which may have an impact on the backward induction outcome. In light of this, we introduce a notion of "role-robust" implementation, which means that outcomes attained via backward induction fall within the SCR regardless of which individual assumes the role of the first-mover, and that all elements of the SCR can be attained by backward induction by assigning some individual to the role of the first-mover. With only two individuals, SCRs that are role-robust implementable can thus select at most two elements for each preference pair. A role-robust implementable SCR naturally leads to an RSCF by tossing a coin to randomly select an element of the SCR. This associated RSCF is clearly implementable by backward induction, via the extensive-form where chance decides in a first move who will assume the role of the first player.

DEFINITION 6: *A SCR f is* role-robust implementable by backward induction *if there exists a two-player extensive-form mechanism of perfect information such that, for each $(\succ_1, \succ_2) \in \mathcal{P} \times \mathcal{P}$, $f(\succ_1, \succ_2)$ coincides with the union of the*

two subgame perfect equilibrium outcomes associated with the two extensive-form games obtained when assigning either the first or the second party to the role of the first player.

We are now ready to provide a sharp answer to our question, as there is a unique SCR that is Pareto efficient, passes the MST, and is role-robust implementable by backward induction via a two-stage mechanism. In addition, it is role-robust implementable via a simple, intuitive shortlisting mechanism.

PROPOSITION 3: *There exists a unique SCR f^* that is Pareto efficient, passes the MST, and is role-robust implementable by backward induction via a two-stage mechanism:*

$$f^*(\succ) = \bigcup_{i\in\{1,2\}} \max_{\succ_i}\{a \in \mathcal{A} \mid \#\{b \in \mathcal{A} \mid a \succ_{j\neq i} b\} \geq \frac{n-1}{2}\}.$$

In addition, f^* is implementable via the following two-stage mechanism:

PROCEDURE 2: *(Shortlisting)* (SL) The party that has been selected to be the first mover chooses a subset containing $\frac{n+1}{2}$ elements of $\mathcal{A}$, and the other party subsequently picks an arbitrator out of that subset.

While having a short dynamic game makes counterfactual reasoning simpler, and should make it more likely that participants' choices are consistent with backward induction, games with fewer rounds may be complex in other dimensions.[6] In the case of SL, one may perhaps fear at first that finding an optimal strategy is relatively difficult for the first-mover as he faces many options to choose from. This should not be a concern though, as his optimal strategy is easy to derive. First, player 1 finds his most preferred alternative in the set of the top $(n+1)/2$ elements for player 2. Player 1 then proposes his most preferred element from this set along with the bottom $(n-1)/2$ elements for player 2.

II. Empirical Analysis

The Preliminary Observations in the previous section highlighted a number of theoretical concerns with using the VR mechanism. There are two important assumptions underlying these results. First, the preferences of the two parties should not be strictly opposed, as otherwise truthtelling would be a Nash equilibrium. Second, parties must behave strategically. If parties naïvely delete worse options and truthfully report their ranking for the remaining arbitrators, then VR would attain desirable outcomes.

To obtain some indirect empirical evidence in support of these two implicit assumptions, we conducted two tests using real-world arbitration cases from the

[6] In SL, for instance, backward induction amounts to a single-person decision problem under the fairly weak epistemic conditions that the first-mover is rational and believes that the second-mover is also rational.

New Jersey Public Employment Relations Commission, which employed VR during the years 1985 to 1996. Full details of both tests can be found in the online appendix. The first test examined the assumption that preferences are not strictly opposed. If preferences are strictly opposed, truthful behavior, while not a unique Nash equilibrium, is a focal equilibrium, and, moreover, under any equilibrium, there should be no overlap in vetoes. In the data, we show that there is a significant degree of overlap in both the rankings and vetoes submitted by the union and the employer, suggesting that preferences are not strictly opposed.

The second test provided suggestive evidence for non-truthful strategic behavior. Our data contains 249 instances in which the same employer had the *same* two arbitrators in his choice set in two different arbitration cases, and *neither* arbitrator was selected in these two cases, *nor* in any case during the period between them. Under the assumption that an employer's relative ranking of an arbitrator can change only as a result of direct experience with that arbitrator, a truthful employer should treat the two arbitrators in the same way in both cases. In roughly-one third of the 249 observations, however, an employer reverses his ranking of the two arbitrators.

Testing the VR in the controlled environment of the lab would allow us to obtain direct evidence on participants' behavior and also on the performance of this mechanism. Since SL is not currently being used, lab experiments are the only way to obtain evidence on actual behavior in this mechanism and to compare its observed performance both with the theoretical predictions and also with the observed behavior in VR.

A. Design

The experiments were conducted at NYU's Center for Experimental Social Science. A total of 158 subjects from the undergraduate student population participated.

In each treatment, an even number of subjects was presented with a set of five alternatives, $\mathcal{A} = \{a, b, c, d, e\}$, and were randomly matched to play one of the mechanisms on this set of options. Each treatment consisted of 40 rounds. In every round subjects were randomly re-matched. Each of the rounds was divided into four "blocks" of ten rounds. In each of these blocks, subjects had the same preference relation over the five options, but these preferences changed from one block to another (i.e., in total there are four distinct preference profiles). Preferences over $\mathcal{A}$ are induced by assigning each of the options a distinct monetary value in the set $\{\$1.00, \$0.75, \$0.50, \$0.25, \$0.00\}$.

As shown in Table 1, the first profile, Pf_1, consists of completely opposed rankings. The second profile, Pf_2, represents partial conflict of interest involving only the top two options. This is a case where in the VR mechanism, truthtelling does not form a Nash equilibrium, and where there is a risk of bad outcome due to miscoordination (see proof of Preliminary Observations a) and c) in the previous section). The third profile, Pf_3 displays a similar partial conflict of interest at the

Table 1—Four Preference Profiles Tested in the Experiment.

Pf1		**Pf2**		**Pf3**		**Pf4**		
Pl. 1	Pl. 2	Pl. 1	Pl. 2	Pl. 1	Pl. 2	Pl. 1	Pl. 2	Payment
a	e	a	b	a	c	a	e	\$1.00
b	d	b	a	b	b	b	c	\$0.75
c	c	c	c	c	a	c	a	\$0.50
d	b	d	d	d	d	d	b	\$0.25
e	a	e	e	e	e	e	d	\$0

top, but this time with the addition of a focal compromise (b). The fourth profile, Pf_4, captures cases where the veto-rank mechanism admits (undominated) Nash equilibria whose outcome do not belong to the veto-rank SCR (see Preliminary Observation b) in the previous section).

There were two treatments, one for VR and one for SL. There were 70 participants in the first treatment and 88 in the second. For each mechanism and each preference profile, we have characterized the set of pure-strategy equilibria.[7] For each treatment we ran four sessions, where in each session the four induced preference profiles appear in a different order. The four orders were: $Pf_1 - Pf_2 - Pf_3 - Pf_4$, $Pf_4 - Pf_3 - Pf_2 - Pf_1$, $Pf_1 - Pf_3 - Pf_2 - Pf_4$, and $Pf_4 - Pf_2 - Pf_3 - Pf_1$. Hence, each profile was played (by a different group of subjects) at two different stages in the experiment: an "early" stage (the first ten rounds for Pf_1 and Pf_4 and the second block of ten rounds for Pf_2 and Pf_3) and a "late" stage (the last ten rounds for Pf_1 and Pf_4 and the third block of ten rounds for Pf_2 and Pf_3). This allows us to examine whether there was a learning "spillover" from one profile to another.

Subjects were paid the sum of their earnings across the 40 rounds in addition to a show-up fee of \$10. The Supplementary Appendix contains the instructions provided to participants. After the subjects read these instructions, they were presented with a short quiz, which is also included in the Appendix, testing their understanding of the game. When the subjects finished answering the quiz, they were presented with the correct answers.

B. Strategic behavior and outcomes in VR

As explained in Section 2, the veto-rank mechanism delivers appealing outcomes when participants are truthful, with both participants vetoing their bottom two options and ranking the remaining three in accordance to their preferences. Yet there are theoretical reasons to believe that participants would not be truthful, and strategize instead. Do participants in the VR procedure tend to be truthful?

[7] It is straightforward to verify whether a pair of strategies constitute an equilibrium in the veto-rank, and equilibrium strategies for the shortlisting scheme were described in the previous section.

RESULT 1: *The majority of participants in the VR treatment are not truthful. Those who do not play truthfully appear to follow some strategic motives instead of playing randomly.*

SUPPORT: As shown in Table 2, a minority of participants play truthfully. These percentages constitute upper bounds on the fraction of "naïve" participants who played the truthful strategy non-strategically. This is because truthful behavior may be a best response against the other party's strategy (both parties being truthful is even a Nash equilibrium in Pf1).

TABLE 2—PERCENTAGE OF SUBJECTS WHO PLAYED TRUTHFULLY

	Pf1	**Pf2**	**Pf3**	**Pf4**
Truthful	50%	43%	31%	26%

In light of Table 2 we proceed to analyze the behavior of non-truthful participants and understand whether their behavior reflects strategic reasoning. We begin with a test which is based on the idea that the choices of a subject who engages in strategic reasoning take into account the preferences of his opponent. This suggests that if a subject is *not* strategizing (e.g., a subject who just picks his strategy at random) then how he treats his k-th ranked option (whether he vetoes it or how he ranks it in his strategy) should not be affected by his opponent's preferences. Given that the number of possible rankings is large, we focus here on the distribution of vetoes. Further, given the small number of cases in which a player vetoes his most-preferred alternative, we combine all such cases into one. This yields six possible non-truthful vetoes.[8] We then test the null hypothesis that the distribution of such vetoes is the same across the preference profiles of the other player. The Pearson chi-square statistic associated with this test is 679 (with a p-value less than 0.001), and we can thus strongly reject the null hypothesis.

Given this suggestive evidence on strategic reasoning, we turn to examine whether the behavior of non-truthful participants is consistent with Nash equilibrium. Our data reveals that a large fraction of participants selected strategies that are part of some Nash equilibrium. However, the existence of many Nash equilibrium strategies (due to thick best response correspondences) may make statistical tests useless. For instance, all observed individual strategies are part of some Nash equilibrium in Pf2 and Pf3, but 58 (54) strategies out of 60 satisfy this property for Pf 2 (Pf3). The only preference pair where a truly tight test is available is Pf1, where only 6 out of 60 individual strategies (10%) are part

[8]For player 1, these are $\{b, c\}$, $\{b, d\}$, $\{b, e\}$, $\{c, d\}$, $\{c, e\}$, and $\{a, x\}$, where x can be either b, c, d, or e. For player 2, we use the same set after re-labeling the options so that the first-choice for player 2 is always a, the second choice is b, the third choice is c, the fourth choice is d, and the least preferred alternative is e. Given that the labels do change for player 2, we have also run this test for only player 1, and the results are similar in nature.

of a Nash equilibrium. By contrast, 80% of observed individual strategies are compatible with Nash for that preference profile, and 47% conditional on being non-truthful. These numbers become even more striking when considering sessions where Pf1 is played in the later part of the experiment (allowing subjects to grow accustomed with the rules of the game): 90% among all observations, and 74% conditional on being non-truthful.

In addition to looking at individual behavior we also computed for each preference profile the percentage of *matched pairs* whose joint actions is a non-truthful Nash equilibrium (Pf1 is the only profile with a truthful equilibrium). The following table compares these percentages with the probability of drawing a Nash pair at random.

Table 3—Percentage of action pairs that are Nash equilibrium.

	Pf1	**Pf2**	**Pf3**	**Pf4**
Overall	40%	35%	19%	37%
Late, Last 5	53%	34%	25%	44%
Random	1%	16%	10%	4%

Note: "Overall" means average over rounds in which the preference pair was played; Late, Last 5" means average over rounds 36-40 for Pf1 and Pf4 and rounds 26-30 for Pf2 and Pf3.

For each profile, the differences between the percentage of observed (non-truthful) Nash pairs and the probability of randomly drawing a Nash pair are statistically significant at the 1% level. Coordination on non-truthful Nash pairs is highest in Pf1 (where it exceeds 50% in rounds $36-40$) and lowest in Pf3. Still, as was shown in Section 2, the VR mechanism may result in undesirable outcomes even if players always coordinated on a Nash equilibrium.

Participants could also be strategizing without necessarily coordinating on a Nash equilibrium. To investigate this possibility we adopt the non-equilibrium framework of k-level reasoning (see the survey in Crawford, Costa-Gomes and Iriberri (2013)). The natural candidate for level zero (non-strategic) behavior is being truthful. Level 1 would then constitute a best response against truthful behavior. Table 4 depicts the percentages of Level 1 choices in the data.

Table 4—Percentage of Non-Truthful Subjects Who Played Level 1.

	Pf1	**Pf2**	**Pf3**	**Pf4**
Level 1 Among Non-Truthful	63%	69%	28%	47%

As shown in Table 4, a significant proportion of non-truthful strategies are consistent with level 1 behavior, which suggests that non-truthful subjects are behaving strategically rather than randomly (the p-value associated with getting a larger percentage under the assumption that non-truthful subjects play randomly

is less than 0.001 for each of the four preference profiles).[9] While at least 97% of observed choices can be explained by levels 0, 1 and 2 in all four preference profiles, adding depths of reasoning larger than 1 is not as informative as may seem because each strategy admits many best responses.[10] □

Truthful behavior is a sufficient condition to obtain "desirable" outcomes (in terms of efficiency and MST) in the veto-rank mechanism if both participants play truthfully. With 43% of participants playing truthfully under Pf2, as reported in Table 2, then on average only 18% of the matched pairs have both participants play truthfully, and these frequencies are even lower under Pf3 (10%) and Pf4 (7%). However, truthful behavior is not a necessary condition to achieve efficiency and MST. To evaluate the performance of VR, we now turn to investigate the observed outcomes. Even though the individual behavior of subjects does not exactly match theoretical predictions, the analysis of outcomes does confirm the insight derived from the theory.

RESULT 2: *Observed outcomes for the Veto-Rank procedure are often inefficient and/or fail the MST.*

SUPPORT: Table 5 displays for each preference pair the percentage of observed outcomes that are inefficient or fail the MST. As a benchmark, we indicate the likelihood of such outcomes if subjects were to play randomly. Preferences being perfectly opposed in Pf1, passing the MST is a stringent test: only c qualifies, and 80% of observed outcomes would fail the test if participants were to play randomly. By contrast, all outcomes are Pareto efficient. For Pf2, Pareto inefficiency will occur 60% of the time if participants play randomly, as only a and b are Pareto efficient. Both pass the MST, as does c. MST and Pareto efficiency coincide in Pf3, both ruling out d and e. Finally, in Pf4, Pareto efficiency narrows the set of outcomes to $\{a, c, e\}$, while the MST further rules out e.

Notice how, for each of the four preference pairs we tested, one of the two criteria turns out to be more restrictive. Thus the frequency of observed outcomes that violate Pareto efficiency or the MST is simply the maximum of the percentage of observed outcomes that violate either criterion.

As evident from the table, a significant proportion of realized outcomes either violate the MST or are Pareto inefficient. We included Pf1 in our experiment because strategic behavior has more robust implications when preferences are perfectly opposed (as in zero-sum games). The fact that 27% of outcomes violate the MST for Pf1 may thus seem surprising, even compared to the large 80%

[9] We conjecture that the percentage drop for Pf3 may be attributable to fairness concerns. While Table 4 is built under the assumption that subjects care only about their own monetary payoff, we observe that 61% of subjects who are neither truthful nor level 1 ranked b above their top choice. This behavior is natural for subjects who value the fact that b strikes a compromise between the two other Pareto efficient options, a and c.

[10] Between 75 and 90% (depending on the preference profile) of strategies belong to one of these three levels. By contrast, only 10 and 25% of all strategies qualify as level 1.

Table 5—Percentage of Outcomes in VR that Are Inefficient or Fail MST.

		Pf1	Pf2	Pf3	Pf4
% Inefficient	Observed Outcomes	0%	9%	12%	19%
	Random Play	0%	60%	40%	40%
% Failing MST	Observed Outcomes	27%	3%	12%	21%
	Random Play	80%	40%	40%	60%

Random Play benchmark. This is likely due to mistakes and subjects' experimentation to get better accustomed with the strategic features of the new game they face. To test this hypothesis, we exploit the fact that different groups of subjects faced preference pairs at different times in the experiment. As explained in Section 3.1, roughly half of the subjects played Pf1 early in their session (Rounds 1-10), while the other half played it late in their session (Rounds 31-40). Similarly, each of the other three preference pairs was played by roughly half of the subjects at a relatively early stage of their session, while the other half played it at a relatively later stage (see footnote 11). Table 5a refines Table 5 by showing how the proportions of outcomes that are inefficient or violate the MST spread over earlier and later stages. The Random Play benchmark remains unchanged, and is thus omitted.

		Pf1	Pf2	Pf3	Pf4
% Inefficient	Early	0%	6%	10%	17%
	Late	0%	12%	13%	21%
% Failing MST	Early	41%	2%	10%	19%
	Late	11%	3%	13%	24%

Table 5a. Inefficiency/Violation of MST in VR as a Function of how Early or Late the Preference Pair Was Played in the Experimental Session[11]

Note that the frequency of outcomes other than c decreases dramatically when Pf1 is played later in the session (this difference is statistically significant at the 1% level). This suggests that participants have a better understanding of the strategic features of the VR mechanism as they gain experience by playing it with other preference pairs. We therefore turn to examine whether the instances of Pareto inefficiency and MST violations under Pf2-4 are due to subjects' lack of experience with the mechanism. The theory suggests that this is not the case since (*i*) a better understanding of the strategic features of the game will not help resolve miscoordination, and (*ii*) outcomes need not be desirable under VR even when participants play a Nash equilibrium (in which case they would fully understand the strategic features of the game, and have rational expectations).

[11]Early = Rounds 1-10 for Pf1 and Pf4, or Rounds 11-20 for Pf2 and Pf3; Late = Rounds 31-40 for Pf1 and Pf4, or Rounds 21-30 for Pf2 and Pf3.

As shown in Table 5a, the percentage of undesirable outcomes for Pf2-4 actually rises when VR is played at a later stage.

One can also address the question of experience and learning by exploiting the fact that subjects played the VR mechanism with a same preference pair for ten rounds in a row. Percentages from Table 5 can then be decomposed based on whether the preference pair was played over the first half or the second half in the block of ten rounds. A table analogue to Table 5a is provided in the Appendix. Results are qualitatively the same as for the Early-Late analysis: the percentage of MST violations under Pf1 significantly decreases at later rounds in a block of ten, while percentages of both inefficiency and MST violations remain constant or increase at later rounds under Pf2-4. □

To summarize, we find a significant degree of non-truthful behavior under VR, which may be explained by strategic considerations and which leads to poor outcomes in many cases. We next examine whether this mechanism is outperformed by SL, a sequential mechanism that is not currently used in practice.

C. Comparing the Performance of SL and VR

The theoretical analysis of Section 2 predicts that SL dominates VR according to the criteria of Pareto efficiency and MST. We now turn to examine the extent to which this prediction is consistent with our experimental data.

RESULT 3: *SL outperforms VR according to both Pareto efficiency and the MST. The difference is statistically significant for all preference pairs except Pf3.*

SUPPORT: Table 6 presents the percentage of matches whose outcome failed the efficiency criterion or the MST, as a function of the preference profile.

TABLE 6—PERCENTAGE OF OUTCOMES IN SL THAT ARE INEFFICIENT OR FAIL MST.

		Pf1	**Pf2**	**Pf3**	**Pf4**
Inefficient	Observed Outcomes	0%	3%	11%	7%
	Random Play	0%	60%	40%	40%
Failing MST	Observed Outcomes	18%	1%	11%	10%
	Random Play	80%	40%	40%	60%

A comparison of Tables 5 and 6 shows that the percentages of outcomes that are inefficient or violate the MST (see second paragraph in the support of Result 2) are systematically lower in SL, but only marginally so with respect to Pf3. To examine the statistical significance of these differences, we tested whether outcomes are more likely to be either inefficient or to fail MST under VR, relative to SL. The differences are statistically significant with $p \leq 0.01$ for preference pairs except Pf3, for which the p value is 0.567.

In the discussion of Result 2, we observed that subjects may gain experience with a mechanism by playing it with other preference pairs. Given that arbitration participants are oftentimes professionals (e.g. lawyers) who are experienced with the selection process, it is interesting to check how the percentages of inefficiency and MST violations change depending on whether the preference pair is played earlier or later in the experiment.

		Pf1	**Pf2**	**Pf3**	**Pf4**
% Inefficient	Early	0%	3%	10%	9%
	Late	0%	3%	12%	5%
% Failing MST	Early	21%	0%	10%	16%
	Late	13%	2%	12%	6%

Table 6a. Inefficiency/Violation of MST in SL as a Function of how Early or Late the Preference Profile Was Played in the Experimental Session[12]

As was the case in VR, one would expect the prevalence of c to increase as participants become better accustomed with the procedure. As seen in Table 6a, the data confirms this intuition.[13] However, contrary to VR, one would expect the outcomes to become only more desirable if a significant change occurs when the preference pair is played later in the experiment. This is because equilibrium behavior in the SL mechanism (as captured by subgame perfection) leads to desirable outcomes (while Nash equilibrium outcomes in the VR mechanism need not be, as shown in Section 2). Pf4 is the only preference among Pf2-4 for which a statistically significant change occurs. Outcomes indeed become only more desirable, and SL becomes only more appealing than VR, when Pf4 is played later in the experimental session. SL outperforms VR in Pf2 independently of the stage at which it is played. □

The relative underperformance of SL in Pf3, compared to the theoretical benchmark, appears to be a robust feature, unrelated to inexperience. If anything, it becomes only more prevalent when the preference pair is played at a later stage, as seen in Table 6a. The next subsection argues that this feature of the data is consistent with an existing theory of social preferences.

Finally, we provide evidence on payoffs under the two mechanisms. As shown in Table 7, when summed across the two agents, payoffs are higher under SL under all three preference profiles.[15] Moreover, while the differences are not statistically significant for Pf3, the differences are statistically significant with $p = 0.001$ for Pf2 and $p = 0.055$ for Pf4. This provides further evidence that SL outperforms VR.

[12]See footnote 11.

[13]The vast majority (20/24) of non-c outcomes occur when Pf1 is played in the last ten rounds are due to the same four subjects who systematically depart from equilibrium when playing the role of both the first and the second mover (that is picking an option which is suboptimal for them, at the benefit of the opponent when they are second-movers, and picking a suboptimal shortlist - including more advantageous options to the opponent - when being first-movers). Either these four subjects did not understand the preference structure at all, or they have very strong altruistic (not intention-based) preferences.

[15]We do not present results here for Pf1 given the zero-sum nature of this treatment.

Table 7—Average aggregate payoffs under the two mechanisms.

	Pf2	Pf3	Pf4
VR	1.6693	1.3593	1.2640
SL	1.7210	1.3886	1.3017
Difference	0.0517	0.0293	0.0377

14

D. Intention-Based Reciprocity

An unexpected outcome of our experiment is that SL does not perform as well as anticipated with Pf3: 11% of observed outcomes are both inefficient and/or violate the MST (the two criteria coincide for Pf3). This percentage is only slightly smaller than for VR, and large in view of the 40% Random Play benchmark. Moreover, SL does not yield statistically significant differences in payoffs, when compared to VR.

It is also important to note that inefficiency/MST violations in Pf3 are hardly attributable to mistakes. Given that efficiency/MST rules out only two options in that preference pair, the second mover has the opportunity to pick an efficient option whatever the shortlist offered by the first-mover. Thus inefficiency occurs only when the second-mover decides to pick an option which is inferior both for him and the other party than an alternative in the shortlist.

As is the case in many applications of mechanism design, our theoretical analysis from Section 2 assumed that parties care only about their own monetary payoff. The surprising relative underperformance of SL with Pf3 can be explained in perspective of recent developments on social preferences.

RESULT 4: *The relative underperformance of SL with Pf3 is attributable to individual behavior consistent with intention-based reciprocity.*

Support: Recall Pf3: the two participants have opposed preferences over the top three elements (a, b, and c), and rank the other two at the bottom. Suppose, to fix ideas, that Party 1 is the first-mover. Clearly, he can get his top choice (a) by offering a shortlist that consists of Party 2's bottom three alternatives ($\{a, d, e\}$). While Party 2 is then expected to pick a out of that shortlist, all the inefficient outcomes (except for one case) occur from Party 2 picking the dominated d or e out of the shortlist $\{a, d, e\}$ (or, likewise, Party 1 picking the dominated d or e out of the shortlist c, d, e when Party 2 is the first-mover).

To see how this behavior relates to the literature on other-regarding preferences, note first that SL may be viewed as a variation of the ultimatum game: instead of offering a single efficient pair of payoffs (a split of some monetary amount), the first-mover proposes a *set* of payoff pairs - some of which may be inefficient and dominated by another pair in the set; since the second-mover must pick one pair, the analog of refusing an offer in the ultimatum game (which destroys surplus) is to choose an inefficient payoff pair. Such destructive behavior is inconsistent

with standard models of implementation. However, it is consistent with the more recent literature on other-regarding preferences. To see this, one must compare behavior in SL with Pf2 versus the closely related Pf3.

Under Pf2, participants have opposed preferences over the top two elements a and b and rank the other three at the bottom. Backward induction with selfish preferences then induces almost the same behavior as with Pf3: the first-mover proposes his top choice along with two dominated options and the second-mover picks the first-mover's top choice. However, remarkably, the vast majority of subjects conforms with backward induction in Pf2, contrary to what we observed in Pf3. We argue that this difference is consistent indeed with the well-documented phenomenon of *intention-based reciprocity*.

According to the theory of intention-based reciprocity,[16] whether a player's action is likely to trigger negative reciprocity depends not only on that action's consequences, but also on the players' intentions, as measured by the consequences of the other actions that were available to him. In Pf2, the first-mover has essentially two alternatives - propose his top pick or propose the other player's top pick. Consequently, the responder does not view a proposal of the first-mover's top choice as "greedy" or "unfair". However, in Pf3 the proposer had the option of also proposing the compromise outcome b, which both players rank second-best. Offering the shortlist $\{a, d, e\}$ may now appear as greedy or unfair, and trigger retaliation, as noted above. There is also evidence that first-movers accounted for the potential of retaliation. In particular, in Pf3 a significant proportion of first movers (42%) also departed from their optimal selfish strategy by including b in the shortlist.[17] □

III. Concluding remarks

This paper takes an implementation-theoretic approach to the problem of selecting a public good, namely an arbitrator, to two parties with symmetric information. First, we establish that in order to have a mechanism with "socially desirable" properties, one must consider sequential mechanisms and focus on SCRs that choose efficient outcomes that are at least as good as the median outcome for both players. Second, we show that there is only one such SCR, which is implementable by the shortest sequential mechanism, i.e., one with only two stages.

[16] Numerous experimental studies emphasize the key role of intentions in behavior (see Goranson and Berkowitz (1966), Greenberg and Frisch (1972), Falk, Fehr and Fischbacher (2008), Brandts and Sola (2001), Offerman (2002) and McCabe, Rigdon and Smith (2003)).

[17] Notice how behavior in SL with Pf2 versus Pf3 is closely related to Falk et al.'s (2003) observation that identical offers in an ultimatum game generate systematically different rejection rates depending on the proposers' other options. In their experiment, subjects play four versions of the ultimatum game where a proposer must submit one of two options for approval by the responder. One option in all four treatments allocates 80% of a monetary amount to the proposer and leaves the remainder to the responder. The second option varies across treatments. The rejection rate of the 80/20 split was 44% when the second option available to the proposer is a 50/50 split, 27% when the second option is a 20/80 split, 18% when the other option is a 80/20 split (i.e., no choice but to propose 80% for himself), and 9% when the second option is a 100/0 split. SL with Pf2 is analogous to the case where the other option available to the proposer is a 20/80 split, while SL with Pf3 is analogous to the 50/50 split treatment.

Finally, we conducted a series of laboratory experiments where we tested our proposed two-stage SL mechanism and the commonly used simultaneous VR procedure.

Our experimental analysis yielded three key results. First, a large fraction of participants followed strategic behavior, suggesting that the VR procedure may suffer from the deficiencies outlined in the theoretical section. Second, the SL procedure - *which is not used in practice* - outperforms the VR mechanism in terms of two criteria: Pareto efficiency and "the minimal satisfaction test". Third, fairness concerns seem to affect the behavior of some participants, whose decisions may be reconciled with a theory of intentions-based reciprocity.

While our results are presented in the context of arbitrator selection, they potentially may be extended to other situations in which a collective of individuals with symmetric information need to agree on a public good (i.e., an outcome that affects the payoffs of the participants). Examples may include hiring decisions, choosing a set of employees to promote, selecting jury members and deciding on the composition of some committee. Our paper suggests that it may be valuable to study these situations from an implementation-theoretic approach: start by identifying "reasonable" SCRs for the problem at hand; ask whether prevalent procedures implement in theory any of these SCRs; study whether participants in such mechanisms tend to behave according to theory; explore alternative mechanisms that "perform well" both theoretically and behaviorally.

In related work, we also investigate two other sequential mechanisms. One such mechanism, *Alternate Strikes*, is currently used in some arbitration cases. Under this mechanism, both parties alternatively remove a name from the list of potential arbitrators, and the final remaining option is chosen to be the arbitrator. A second sequential mechanism under invesigation, *Voting by Alternating Offers and Vetoes* (first proposed by Anbarci (1993)), is not used in practice. Under this procedure, players take turns in proposing arbitrators. If a proposed arbitrator is rejected by the other party, that arbitrator is removed from the list and the rejecting party then proposes a name from the remaining list. The procedure continues until a proposal is accepted or only one name remains. While these mechanisms are not two-step mechanisms and thus rely even more strongly on backwards induction, it will be interesting to compare their performance with the two key mechanisms, VR and SL, considered here.

Our paper also suggests a new direction in which the implementation literature needs to develop, one in which *behavioral* concerns are taken into account when designing a mechanism. Our SL procedure was derived by taking into account that individuals find it difficult (and therefore are more prone to mistakes) to perform backwards induction for more than two steps. The experimental results illustrated that individuals' preferences may be affected by the mechanism itself. In our study, participants were more likely to be affected by fairness and reciprocity concerns when the mechanism was sequential rather than simultaneous. This suggests that participants' preferences may be *endogenously* determined by the

choice of mechanism. One possible implication of this is that both a mechanism and the agents' preferences may need to be derived as a fixed-point: given a mechanism M, the set of possible preferences is $P(M)$, and given $P(M)$ the mechanism M implements a SCR defined over $P(M)$. Exploring this new direction for mechanism design is left for future research (for recent works in this direction, see Bowles and Hwang (2008) and Bierbrauer and Netzer (2012)).

A natural question that arises from our paper is why the SL procedure is not used in practice, given that it appears to outperform VR. One hypothesis is that the SL procedure creates an asymmetry between the two parties (one moves before another), while, under VR, the two parties have symmetric roles and have exactly the same set of available actions. However, the two parties can be made symmetric ex-ante in the SL procedure by tossing a fair coin to determine who will move first. Indeed this is how the identity of the first mover is determined in the Alternate Strikes procedure mentioned above, which is another example of an asymmetric sequential procedure that is used in practice. An alternative hypothesis may be that the agencies involved believe that the parties are in fact truthful in their reports. Finally, it may be the case that the agencies involved have simply not considered SL and its potential advantages. Of course, we have no way to establish the true reason and can only speculate.

References

Abdulkadiroglu, Atila., Parag A. Pathak, and Alvin E. Roth. 2005. "The New York City High School." *American Economic Review (Papers and Proceedings)* 95 (2): 364-67.

Abdulkadiroglu, Atila., Parag A. Pathak, Alvin E. Roth and Tayfun Sönmez. 2005. "The Boston Public School." *American Economic Review (Papers and Proceedings)* 95 (2): 368-71.

Anbarci, Nejat. 1993. "Noncooperative Foundations of the Area Monotonic Solution." *Quarterly Journal of Economics* 108 (1): 245-58.

Bierbrauer, Felix and Nick Netzer. 2012. "Mechanism Design and Intentions." University of Zurich Working Paper.

Binmore, Ken, John McCarthy, Giovanni Ponti, Larry Samuelson, and Avner Shaked. 2002. "A Backward Induction Experiment." *Journal of Economic Theory* 104 (1): 48-88.

Bloom, David E., and Christopher L. Cavanagh. 1986a. "An Analysis of the Selection of Arbitrators." *American Economic Review* 76 (3): 408-22.

Bloom, David E., and Christopher L. Cavanagh. 1986b. "An Analysis of Alternative Mechanisms for Selecting Arbitrators." Harvard Institute of Economic Research Discussion Paper 1224.

Bowles, Samuel and Sung-Ha Hwang. 2008. "Social Preferences and Public Economics: Mechanism Design When Social Preferences Depend on Incentives." *Journal of Public Economics* 92 (8-9): 1811-20.

Brandts, Jordi and Carles Solà. 2001. "Reference Points and Negative Reciprocity in Simple Sequential Games." *Games and Economic Behavior* 36 (2): 138-57.

Charness, Gary and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics* 117 (3): 817-69.

Chen, Yan. 2008. "Incentive-Compatible Mechanisms for Pure Public Goods: A Survey of Experimental Research." In *The Handbook of Experimental Economics Results.* Edited by Charles Plott and Vernon L. Smith, 625-43. Amsterdam: Elsevier Press.

Crawford, V. P., Miguel A. Costa-Gomes, and Nagore Iriberri. 2013. "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications." *Journal of Economic Literature* 51 (1): 5-62.

Degan, Arianna, and Antonio Merlo. 2009. "Do Voters Vote Ideologically?" *Journal of Economic Theory* 144 (5): 1868-94.

Falk, Armin and Urs Fischbacher. 2006. "A Theory of Reciprocity." *Games and Economic Behavior* 54 (2): 293-315.

Falk, Armin, Fehr, Ernst and Urs Fischbacher. 2008. "Testing Theories of Fairness - Intentions Matter." *Games and Economic Behavior* 62 (1): 287-303.

Forsythe, Robert, Roger B. Myerson, Thomas A. Rietz, and Robert J. Weber. 1993. "An Experiment on Coordination in Multi-Candidate Elections: The Importance of Polls and Election Histories." *Social Choice and Welfare* 10 (3): 223-47.

Forsythe, Robert, Roger B. Myerson, Thomas A. Rietz, and Robert J. Weber. 1996. An Experimental Study of Voting Rules and Polls in Three-Candidate Elections. *International Journal of Game Theory* 25 (3): 355-83.

Goranson, Richard E. and Leonard Berkowitz. 1966. "Reciprocity and Responsibility Reactions to Prior Help." *Journal of Personality and Social Psychology* 3 (2): 227-32.

Greenberg, Martin S. and David M. Frisch. 1972. "Effect of Intentionality on Willingness to Reciprocate a Favor." *Journal of Experimental Social Psychology* 8 (2): 99-111.

Hurwicz, Leonid and David Schmeidler. 1978. "Construction of Outcome Functions Guaranteeing Existence and Pareto Optimality of Nash Equilibria." *Econometrica* 46 (6): 1447-74.

Kagel, John H. 1995. "Auctions: A Survey of Experimental Research." In Handbook of Experimental Economics. Vol. 1, edited by John H. Kagel and Alvin E. Roth, 501-85. Princeton, NJ: Princeton University Press.

Kagel, John H., and Dan Levin, 2011. Auctions: A Survey of Experimental Research, 1995-2010." in *Handbook of Experimental Economics.* Vol. 2, edited by John H. Kagel and Alvin E. Roth, *forthcoming* Princeton, NJ: Princeton University Press.

Kawai, Kei and Yasutora Watanabe. 2013. "Inferring Strategic Voting." *American Economic Review* 103 (2): 624-62.

Levitt, Steven D., John A. List, and Sally E. Sadoff. 2011. "Checkmate: Exploring Backward Induction Among Chess Players." *American Economic Review* 101 (2): 975–90.

McCabe, K., Rigdon, M. and V. Smith, 2003. Positive Reciprocity and Intentions in Trust Games. *Journal of Economic Behavior and Organization* **52**, 267–275.

Offerman, Theo. 2002. "Hurting Hurts More Than Helping Helps: The Role of the Self-Serving Bias." *European Economic Review* 46 (8): 1423-37.

Roth, Alvin E. 1984. "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory." *Journal of Political Economy* 92 (6): 991-1016.

Roth, Alvin E. 2002. "The Economist as Engineer: Game Theory, Experimental Economics and Computation as Tools of Design Economics." *Econometrica* 70 (4): 1341-78.

Roth, Alvin E. 2007. "The Art of Designing Markets." *Harvard Business Review* October Issue: 118-26.

Mathematical Appendix

Proof of Preliminary Observations.

We provide an argument for $n = 5$, the case studied in the experimental section, but it easily generalizes to any n.

Proof of a. Let $\mathcal{A} = \{a, b, c, d, e\}$ and (u_1, u_2) generate the following rankings: $a \succ_1 b \succ_1 c \succ_1 d \succ_1 e$ and $b \succ_2 a \succ_2 c \succ_2 d \succ_2 e$ ($Pf2$ in our experiment). Note that reporting truthfully (i.e., vetoing d and e and giving a score of 2 to the top ranked element, a score of 1 to the second-best element and a score of 0 to the remaining element) is not a Nash equilibrium of the veto-rank procedure. If players followed this naïve strategy, they would end up in a tie, where either a or b is randomly chosen. If, on the other hand, player 1 would veto b instead of d, then a would be chosen uniquely, which he prefers.

Observe that any strategy where a player does not veto his most-preferred option, and ranks the remaining three options truthfully cannot be weakly dominated. We prove this by contradiction for player 1 (a similar reasoning applies to player 2), while assuming without loss of generality that $a \succ_1 b \succ_1 c \succ_1 d \succ_1 e$ (otherwise options can simply be relabeled). Consider thus a strategy s where player 1 vetoes options x and y, which are both different from a, ranks the other three alternatives truthfully, and suppose, contrary to what we want to prove, that it is weakly dominated by an alternative strategy s'. If the second player vetoes the two options other than a, x and y, and ranks x above y above a (resp.

y above x above a), then the outcome under s' is worse than under s if s' does not veto x (resp. y). Since it is assumed that s' weakly dominates s, it must be that both x and y are vetoed under s'. Since s ranks the other options truthfully, it must be that an option $z \in \mathcal{A} \setminus \{x, y\}$ is ranked above an option $w \in \mathcal{A} \setminus \{x, y\}$ under s', while $w \succ_1 z$. This results in a contradiction though, as the outcome when 1 plays s' is worse than the outcome when he plays s if the second player vetoes both a and x, and ranks z top and w second-best.

We are now ready to prove that the VR mechanism admits a non-truthful undominated Nash equilibrium for any preference pair. Again, we assume without loss of generality that $a \succ_1 b \succ_1 c \succ_1 d \succ_1 e$. We show in different steps that for all of player 2's preferences, one can construct a non-truthful undominated Nash equilibrium, using the result from the previous paragraph to show that strategies are undominated. If a falls below the median option in $\succ_2$, then an example of such strategies would be to have player 1 make a truthful report, with player 2 reporting a truthful ranking, and vetoing both a and an option that falls at or above his median, that is not his top choice, nor his most-preferred option among those that survives player 1's vetoes (if there are two options that satisfy these three properties, then player 2 picks the one that is not vetoed by 1). Suppose now that a does not fall below the median option in $\succ_2$. We conclude the argument by considering three subcases. First, suppose that a is also most-preferred for player 2. The property then holds with a truthful report for player 2, and player 1 reporting a truthful ranking, while vetoing both his second-best option together with an option below the median in $\succ_1$. Second, suppose there exists an option that player 2 ranks above a and that is not below the median for $\succ_1$. The property then holds with a truthful report for player 2, and player 1 reporting a truthful ranking, while vetoing all the options that 2 prefers over a (plus possibly an additional option below the median for $\succ_1$ if one more veto remains free). Third, suppose there are options that player 2 ranks above a, but they all fall below the median for $\succ_1$. The property then holds with a truthful report for player 1, and player 2 reporting a truthful ranking, while vetoing both an element at or above his median which is different from both a and his top pick, and an option that falls below the median in $\succ_2$.

Proof of b. Observe that c, which is Pareto dominated by a and b, is an equilibrium outcome when considering the pair of preferences $(\succ_1, \succ_2)$ from the first paragraph of the proof of (a). On the other hand, one might argue that this Nash equilibrium is less likely to emerge since it involves dominated strategies.

Consider then the Bernoulli functions (v_1, v_2) generating the rankings $a \succ'_1 b \succ'_1 c \succ'_1 d \succ'_1 e$ and $e \succ'_2 c \succ'_2 a \succ'_2 b \succ'_2 d$ ($Pf4$ in our experiment). Then $f_{VR}(\succ'_1, \succ'_2) = \{a\}$. However, there exists a (undominated) Nash equilibrium in which player 2 chooses $\mathcal{V}_2 = \{a, b\}$ and s_2 such that $r_2(e) = 2$, $r_2(c) = 1$ and $r_3(d) = 0$, while player 1 chooses $\mathcal{V}_1 = \{d, e\}$ and r_1 such that $r_1(a) = 2$, $r_1(b) = 1$ and $r_1(c) = 0$. The outcome of this equilibrium is c, which does not belong to $f_{VR}(\succ'_1, \succ'_2) = \{a\}$.

Next, we show that undominated Nash equilibrium outcomes may even be Pareto dominated. Suppose that player 1 has the same preference as in the first paragraph of the proof of (a), but that 2's preference is given by $e \succ_2 b \succ_2 c \succ_2 d \succ_2 a$. It is straightforward to verify that the following strategy profile is a Nash equilibrium: player 1 vetoes $\{e, b\}$ and ranks the remaining options truthfully, while player 2 vetoes $\{a, b\}$ and ranks the remaining options truthfully. The resulting option, c, is Pareto inefficient, and the strategies are not weakly dominated (see second paragraph in the proof of (a)).

Proof of c. The preference profile $(\succ_1, \succ_2)$ described in the first paragraph of the proof of (a) induces a pair of undominated equilibria, s and s', with the following properties. In s, player 1 ranks options truthfully while vetoing b and c and player 2 is truthful. In s', player 2 ranks options truthfully while vetoing a and d, and player 1 is truthful. It follows that (s_1, s_2') induces e which both violates the MST and is Pareto inefficient. ■

Proof of Proposition 1

Part a) follows as a Corollary of Hurwicz and Schmeidler (1978). Indeed, they proved that any SCR that is Pareto efficient and weakly implementable must be dictatorial. Any such SCR will thus fail the MST.

We now pay attention to RSCFs. The proof is made for the case where $\mathcal{A}$ contains five elements - $\mathcal{A} = \{a, b, c, d, e\}$ - but can easily be extended to any number of elements. Consider (u_1, u_2) such that $u_1(a) > u_1(b) > u_1(c) > u_1(d) > u_1(e)$, and u_2 is completely opposite. If ψ passes the MST, then $\psi(u_1, u_2)$ yields c with certainty. Maskin Monotonicity implies that $\psi(u_1', u_2')$ also yields c with certainty, where $u_1'(c) > u_1'(e) > u_1'(a) > u_1'(b) > u_1'(d)$ and $u_2'(e) > u_2'(c) > u_2'(a) > u_2'(b) > u_2'(d)$. Consider (u_1'', u_2'') such that $u_1''(c) > u_1''(a) > u_1''(e) > u_1''(b) > u_1''(d)$, and u_2'' is completely opposite. If ψ passes the minimal satisfaction test, then $\psi(u_1'', u_2'')$ yields e with certainty. Maskin monotonicity then implies that $\psi(u_1', u_2')$ also yields e with certainty, a contradiction. This establishes b).

Statement c) then follows from a) and b), given that ψ_{VR} and f_{VR} are Pareto efficient and satisfy the MST. ■

Proof of Proposition 2

We provide an argument for $n = 5$, where the elements are given by the set $\mathcal{A} = \{a, b, c, d, e\}$, but it is easy to generalize it to any larger integer by adding options at the bottom of preference rankings. Suppose that $f \subseteq f_{VR}$ is subgame-perfect implementable, and consider the following three pairs of preferences:

$$a \succ_1 b \succ_1 c \succ_1 d \succ_1 e \text{ and } c \succ_2 d \succ_2 b \succ_2 a \succ_2 e$$

$$a \succ_1' b \succ_1' c \succ_1' d \succ_1' e \text{ and } c \succ_2' b \succ_2' a \succ_2' d \succ_2' e$$

$$b \succ_1'' a \succ_1'' c \succ_1'' d \succ_1'' e \text{ and } c \succ_2'' b \succ_2'' a \succ_2'' d \succ_2'' e$$

We first argue that $c \notin f(\succ')$. Observe that $f(\succ'')$ is a subset of $f_{VR}(\succ'')$, which is equal to $\{b\}$. Hence $f(\succ'') = \{b\}$. Suppose that $c \in f(\succ')$. Since c is dropped from f when moving from $\succ'$ to $\succ''$, Abreu and Sen's (1992) necessary condition implies that there exist a non-negative integer ℓ, a sequence $(a_k)_{k=0}^{\ell+1}$ in $\mathcal{A}$, and a sequence $(i_k)_{k=0}^{\ell+1}$ in $\{1,2\}$ such that (i) $a_0 = c$, (ii) $a_k \succ'_{i_k} a_{k+1}$, for all $k = 0, \ldots, \ell$, (iii) $a_{\ell+1} \succ''_{i_\ell} a_\ell$, and (iv) a_k is not $\succ''_{i_k}$-maximal in $\mathcal{A}$.[18] Conditions (ii) and (iii) imply a preference reversal for i_ℓ regarding a_ℓ and $a_{\ell+1}$. This is possible only if $i_\ell = 1$, $a_{\ell+1} = b$, and $a_\ell = a$. Hence, $a_k \notin \{d, e\}$, for all k (otherwise one would contradict (ii)). In turn, this implies that $i_0 = 2$, as otherwise $a_1 = d$ or e by (ii), but this contradicts (iv). To avoid this contradiction, one must have $c \notin f(\succ')$, as claimed.

The only option that minimizes the sum of scores for $(\succ_1, \succ_2)$ is c. It also passes the minimal satisfaction test for those preferences. Hence $f_{VR}(\succ)= \{c\}$, and $f(\succ) = \{c\}$ a fortiori. We have established that $c \in f(\succ) \setminus f(\succ')$. Notice though that $\succ_1 = \succ'_1$. The only preference reversals occur for the second party. This contradicts Abreu and Sen's necessary condition, since c is $\succ'_2$-maximal. Hence f is not subgame-perfect implementable. ■

Proof of Proposition 3

It is easy to check that Procedure 2 role-robust implements f^* by backward induction, and that f^* is Pareto efficient and passes the MST. It remains to show that f^* is the only SCR with those properties. Let $\succ \in \mathcal{P} \times \mathcal{P}$, and let x be the element of $f^*(\succ)$ corresponding to $i = 1$. We now define a new ordering $\succ'_1$. First the elements that are preferred to $f^*(\succ)$ according to $\succ_1$ keep the same rank[19] in $\succ'_1$. Notice that the rank of all these elements must be strictly larger than $\frac{n+1}{2}$ in $\succ_2$, by definition of f^*. Then place the other elements ranked strictly larger than $\frac{n+1}{2}$ in $\succ_2$ (if any) in some specific order (say, alphabetically) in the next available spots in $\succ'_1$ (that is, after those elements above $f^*(\succ)$ according to $\succ_1$). The next available spot in $\succ'_1$ must be the $\frac{n+1}{2}$-rank. Place $f^*(\succ)$ there, and then rank the remaining elements in some specific order (say, alphabetically again). Let $\bar{f}$ be a SCR that is role-robust implemented by backward induction via a two-stage mechanism, is Pareto efficient and passes the MST. The MST applied to both players implies that $\bar{f}(\succ'_1, \succ_2) = x$. Notice that the lower contour set of x expands when moving from $\succ'_1$ to $\succ_1$. Hence the backward induction outcome of the two-stage mechanism in $(\succ_1, \succ_2)$ when 1 is the first-mover remains x (the second party's optimal strategy remains unchanged since his preference remains fixed). If y denotes the element of $f^*(\succ)$ associated to $i = 2$, then a similar reasoning implies that y is the backward induction outcome of the two-stage mechanism

[18] The set B in Abreu and Sen's necessary condition must contain the range of the social choice rule. It is easy to check that any selection of f_{VR} has full range, since for each option x there exists a preference profile for which x is the only option selected by f_{VR}. Hence the condition must be satisfied for $B = \mathcal{A}$.

[19] The rank of a top ranked element is 1. The rank of the second element according to the ordering is 2, and so on so forth. The rank is thus equal to n minus the score.

when the second-party is the first-mover. It thus follows that $\bar{f}(\succ) = f^*(\succ)$, as desired. ■

First 5-Last 5 Tables

In the main text we analyzed experience by comparing subjects' behavior when a preference profile was played relatively early versus late in the session. This captures the idea that one may learn about the strategic features of a mechanism by playing it with other preferences in the past. Another form of learning occurs when one plays a same game (that is the same game form with the same preferences) multiple times. This can be tested in our data by comparing subjects' behavior when playing a given preference pair in a given mechanism for the first or the last five rounds.

		Pf1	**Pf2**	**Pf3**	**Pf4**
% Inefficient	First 5	0%	8%	12%	19%
	Last 5	0%	9%	12%	19%
% Failing MST	First 5	31%	2%	12%	21%
	Last 5	22%	3%	12%	22%

Table 5b Inefficiency/Violation of MST in VR in First 5 vs. Last 5 Rounds

		Pf1	**Pf2**	**Pf3**	**Pf4**
% Inefficient	First 5	0%	3%	9%	6%
	Last 5	0%	3%	12%	8%
% Failing MST	First 5	23%	2%	9%	10%
	Last 5	12%	0%	12%	10%

Table 6b Inefficiency/Violation of MST in SL in First 5 vs. Last 5 Rounds

Results are qualitatively comparable to that of Tables 5a and 6a: violations of MST decrease in Pf1 with experience, the overall superiority of SL over VR is robust to experience, and the percentage of bad outcomes in Pf3 (due to intention-based reciprocity) does not decrease with experience.