American Economic Review 2017, 107(11): 3257–3287
https://doi.org/10.1257/aer.20160425

Obviously Strategy-Proof Mechanisms[†]

By Shengwu Li*

A strategy is obviously dominant if, for any deviation, at any information set where both strategies first diverge, the best outcome under the deviation is no better than the worst outcome under the dominant strategy. A mechanism is obviously strategy-proof (OSP) if it has an equilibrium in obviously dominant strategies. This has a behavioral interpretation: a strategy is obviously dominant if and only if a cognitively limited agent can recognize it as weakly dominant. It also has a classical interpretation: a choice rule is OSP-implementable if and only if it can be carried out by a social planner under a particular regime of partial commitment. (*JEL* D11, D44, D82)

Dominant-strategy mechanisms are often said to be desirable. They reduce participation costs and cognitive costs, by making it easy for agents to decide what to do.[1] They protect agents from strategic errors.[2] Dominant-strategy mechanisms prevent waste from rent-seeking espionage, since spying on other players yields no strategic advantage. Moreover, the resulting outcome does not depend sensitively on each agent's higher-order beliefs.

These benefits largely depend on agents *understanding* that the mechanism has an equilibrium in dominant strategies, i.e., that it is strategy-proof (SP). Only then can they conclude that they need not attempt to discover their opponents' strategies or to game the system.[3]

However, some strategy-proof mechanisms are simpler for real people to understand than others. For instance, choosing when to quit in an ascending clock auction

* Harvard Society of Fellows, 78 Mt. Auburn Street, Cambridge, MA 02138 (email: shengwu_li@fas.harvard.edu). This paper was accepted to the *AER* under the guidance of Roland Bénabou, Coeditor. I thank especially my advisors, Paul Milgrom and Muriel Niederle. I thank Nick Arnosti, Douglas Bernheim, Gabriel Carroll, Paul J. Healy, Matthew Jackson, Fuhito Kojima, Roger Myerson, Michael Ostrovsky, Matthew Rabin, Alvin Roth, Ilya Segal, and four anonymous referees for their invaluable advice. I thank Paul J. Healy for his generosity in allowing my use of the Ohio State University Experimental Economics Laboratory, and Luyao Zhang for her help in running the experiments. This work was supported by the Kohlhagen Fellowship Fund, through a grant to the Stanford Institute for Economic Policy Research. The author declares that he has no relevant or material financial interests that relate to the research described in this paper.

[†] Go to https://doi.org/10.1257/aer.20160425 to visit the article page for additional materials and author disclosure statement.

[1] Vickrey (1961, p.22) writes that, in second-price auctions: "Each bidder can confine his efforts and attention to an appraisal of the value the article would have in his own hands, at a considerable saving in mental strain and possibly in out-of-pocket expense."

[2] For instance, school choice mechanisms that lack dominant strategies may harm parents who do not strategize well (Pathak and Sönmez 2008).

[3] Policymakers could announce that a mechanism is strategy-proof, but that may not be enough. If agents do not understand the mechanism well, then they may be justifiably skeptical of such declarations. For instance, Google's advertising materials for the Generalized Second-Price auction appeared to imply that it was strategy-proof, when in fact it was not (Edelman, Ostrovsky, and Schwarz 2007).

is the same as choosing a bid in a second-price sealed-bid auction (Vickrey 1961). The two formats are strategically equivalent; they have the same reduced normal form.[4] Nonetheless, laboratory subjects are substantially more likely to play the dominant strategy under a clock auction than under sealed bids (Kagel, Harstad, and Levin 1987). Theorists have also expressed this intuition:

> *Some other possible advantages of dynamic auctions over static auctions are difficult to model explicitly within standard economics or game-theory frameworks. For example… it is generally held that the English auction is simpler for real-world bidders to understand than the sealed-bid second-price auction, leading the English auction to perform more closely to theory* (Ausubel 2004, p. 1469).

In this paper, I model what it means for a mechanism to be *obviously* strategy-proof. This approach invokes no new primitives. Thus, it identifies a set of mechanisms as simple to understand, while remaining as parsimonious as standard game theory.

A strategy S_i is *obviously dominant* if, for any deviating strategy S_i', starting from any earliest information set where S_i and S_i' diverge, the best possible outcome from S_i' is no better than the worst possible outcome from S_i. A mechanism is *obviously strategy-proof* (OSP) if it has an equilibrium in obviously dominant strategies. By construction, OSP depends on the extensive game form, so two games with the same normal form may differ on this criterion. Obvious dominance implies weak dominance, so OSP implies SP.

This definition distinguishes ascending auctions and second-price sealed-bid auctions. Ascending auctions are obviously strategy-proof. Suppose you value the object at \$10. If the current price is below \$10, then the best possible outcome from quitting now is no better than the worst possible outcome from staying in the auction (and quitting at \$10). If the price is above \$10, then the best possible outcome from staying in the auction is no better than the worst possible outcome from quitting now.

Second-price sealed-bid auctions are strategy-proof, but not obviously strategy-proof. Consider the strategies "bid \$10" and "bid \$11." The earliest information set where these diverge is the point where you submit your bid. If you bid \$11, you might win the object at some price strictly below \$10. If you bid \$10, you might not win the object. The best possible outcome from deviating is better than the worst possible outcome from truth-telling. This captures an intuition expressed by experimental economists:

> *The idea that bidding modestly in excess of x only increases the chance of winning the auction when you don't want to win is far from obvious from the sealed bid procedure* (Kagel, Harstad, and Levin 1987, p. 1299).

I produce two characterization theorems, which suggest two interpretations of obvious strategy-proofness. The first interpretation is behavioral: obviously dominant

[4] This equivalence assumes that we restrict attention to cut-off strategies in ascending auctions.

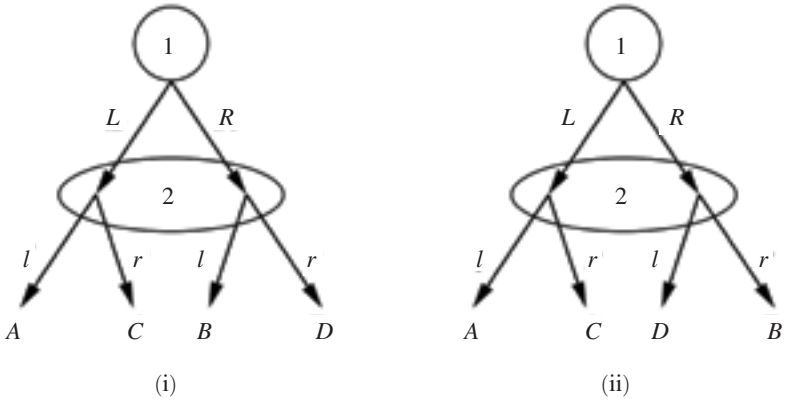

FIGURE 1. "SIMILAR" MECHANISMS FROM 1'S PERSPECTIVE

strategies are those that can be recognized as dominant by a cognitively limited agent. The second interpretation is classical: obviously strategy-proof mechanisms are those that can be carried out by a social planner with only partial commitment power.

First, I model an agent who has a simplified mental representation of the world: instead of understanding every detail of every game, his understanding is limited by a coarse partition on the space of all games. I show that a strategy S_i is obviously dominant if and only if such an agent can recognize S_i as weakly dominant.

Consider the mechanisms in Figure 1. Suppose Agent 1 has preferences: $A \succ B \succ C \succ D$. In mechanism (i), it is a weakly dominant strategy for 1 to play L. Both mechanisms are intuitively similar, but it is not a weakly dominant strategy for Agent 1 to play L in mechanism (ii).

In order for Agent 1 to recognize that it is weakly dominant to play L in mechanism (i), he must use contingent reasoning. That is, he must think through hypothetical scenarios: "If Agent 2 plays l, then I should play L, since I prefer A to B. If Agent 2 plays r, then I should play L, since I prefer C to D. Therefore, I should play L, no matter what Agent 2 plays." Notice that the quoted inferences are valid in (i), but not valid in (ii).

Suppose Agent 1 is unable to engage in contingent reasoning. That is, he knows that playing L might lead to A or C, and playing R might lead to B or D. However, he does not understand how, state-by-state, the outcomes after playing L are related to the outcomes after playing R. Then it is as though he cannot distinguish (i) and (ii).

This idea can be made formal and general. I define an equivalence relation on the space of mechanisms: the *experience* of agent i at history h records the information sets where i was called to play, and the actions that i took, in chronological order.[5] Two mechanisms G and G' are *i-indistinguishable* if there is a bijection from i's information sets and actions in G, onto i's information sets and actions in G', such that

(i) G can produce for i some experience if and only if G' can produce for i the image (under the bijection) of that experience;

[5] An experience is a standard concept in the theory of extensive games; experiences are sometimes used to define perfect recall.

(ii) An experience might result in some outcome in G if and only if its image might result in that same outcome in G'.

With this relation, we can partition the set of all mechanisms into equivalence classes. For instance, the mechanisms in Figure 1 are 1-indistinguishable.

Our agent knows the experiences that a mechanism might generate, and the resulting outcomes. Thus, at each of his information sets, for each continuation strategy, he knows all the possible outcomes that might result from that strategy. However, this does not nail down every detail of the mechanism. In particular, he does not know the possible outcomes *contingent on some unobserved event*.

For instance, in a second-price auction, our agent knows that for each bid, he will either win the object at a price weakly less than that bid, or not win (and pay zero). However, he does not know that if bidding \$11 would cause him to win the object at \$8, then bidding \$10 would also cause him to win the object at \$8. Since the agent is unable to make these inferences about the second-price auction, he is unable to correctly identify his dominant strategy.

In general, what strategies can our agent recognize as weakly dominant, when he cannot tell apart i-indistinguishable mechanisms? These are all (and only) the obviously dominant strategies. The first characterization theorem states: a strategy S_i is obviously dominant in G if and only if it is weakly dominant in every G' that is i-indistinguishable from G.

The second characterization theorem for OSP relates to the problem of mechanism design under partial commitment. In mechanism design, we usually assume that the Planner can commit to every detail of a mechanism, including the events that an individual agent does not directly observe. For instance, in a sealed-bid auction, we assume that the Planner can commit to the function from all bid profiles to allocations and payments, even though each agent only directly observes his own bid. Sometimes this assumption is too strong. If agents cannot individually verify the details of a mechanism, the Planner may be unable to commit to it.

Mechanism design under partial commitment is a pressing problem. Auctions run by central brokers over the internet account for billions of dollars of economic activity (Edelman, Ostrovsky, and Schwarz 2007). In such settings, bidders may be unable to verify that the other bidders exist, let alone what actions they have taken. As another example, some wireless spectrum auctions use computationally demanding techniques to solve complex assignment problems, and the auctioneer may not be permitted to publicly disclose all the bids. In these settings, individual bidders may find it difficult and costly to verify the output of the auctioneer's algorithm (Milgrom and Segal 2015).

For the second characterization theorem, I consider a "metagame" where the Planner privately communicates with agents, and eventually decides on an outcome. The Planner chooses one agent, and sends a private message, along with a set of acceptable replies. That agent chooses a reply, which the Planner observes. The Planner can then either repeat this process (possibly with a different agent) or announce an outcome and end the game.

The Planner has partial commitment power: for each agent, she can commit to use only a subset of her available strategies. However, the subset she promises to Agent i must be measurable with respect to i's observations in the game. That is, if

the Planner plays a strategy not in that subset, then there exists some agent strategy profile such that Agent i detects (with certainty) that the Planner has deviated. We call this a *bilateral commitment*.

Suppose we require that each agent i's strategy be optimal, for any strategies of the other agents, and for any Planner strategy compatible with the commitment made to i. What choice rules can be implemented in this metagame?

The second characterization theorem states: a choice rule can be supported by bilateral commitments if and only if that choice rule is OSP-implementable. Consequently, in addition to formalizing a notion of cognitive simplicity, OSP also captures the set of choice rules that can be carried out with only bilateral commitments.

After defining and characterizing OSP, I apply this concept to several mechanism design environments.

For the first application, I consider binary allocation problems. In this environment, there is a set of agents N with continuous single-dimensional types $\theta_i \in [\underline{\theta}_i, \overline{\theta}_i]$. An *allocation* y is a subset of N. An *allocation rule* f_y is a function from type profiles to allocations. We augment this with a *transfer rule* f_t, which specifies money transfers for each agent. Each agent has utility equal to his type if he is in the allocation, plus his net transfer:

$$u_i(\theta_i, y, t) = 1_{i \in y}\theta_i + t_i. \tag{1}$$

Binary allocation problems encompass several canonical settings. They include private-value auctions with unit demand. They include procurement auctions with unit supply; not being in the allocation is "winning the contract," and the bidder's type is his cost of provision. They also include binary public good problems; the feasible allocations are N and the empty set.

Mechanism design theory has extensively investigated SP-implementation in this environment. The function f_y is SP-implementable if and only if f_y is monotone in each agent's type (Spence 1974; Mirrlees 1971; Myerson 1981). If f_y is SP-implementable, then the required transfer rule f_t is essentially unique (Green and Laffont 1977; Holmström 1979).

What are analogues of these canonical results, if we require OSP-implementation rather than SP-implementation? Are ascending clock auctions special, or are there other OSP mechanisms in this environment?

I prove the following theorem: every mechanism that OSP-implements an allocation rule is "essentially" a *personal-clock auction*, which is a new generalization of ascending auctions. Moreover, this is a full characterization of OSP mechanisms: for any personal-clock auction, there exists some allocation rule that it OSP-implements.

These results imply that when we desire OSP-implementation in a binary allocation problem, we need not search the space of all extensive game forms. Without loss of generality, we can focus our attention on the class of personal-clock auctions.[6]

[6] Of course, if we do not impose the additional structure of a binary allocation problem, then there exist OSP mechanisms that are not personal-clock auctions. This paper contains several examples.

As a second application, I produce an impossibility result for a classic matching algorithm: with three or more agents, there does not exist a mechanism that OSP-implements Top Trading Cycles (Shapley and Scarf 1974).

I conduct a laboratory experiment to test the theory. In the experiment, I compare three pairs of mechanisms. In each pair, both mechanisms implement the same choice rule. One mechanism is obviously strategy-proof. The other mechanism is strategy-proof, but not obviously strategy-proof. Standard theory predicts that both mechanisms result in dominant strategy play, and have identical outcomes. Instead, subjects play the dominant strategy at significantly higher rates under the OSP mechanism, compared to the mechanism that is just SP. This effect occurs for all three pairs of mechanisms, and persists even after playing each mechanism five times with feedback.

The rest of the paper proceeds in the usual order. Section I reviews the literature. Section II provides formal definitions and characterizations. Section III covers applications. Section IV reports the laboratory experiment. Section V concludes. Proofs omitted from the main text are in the online Appendix.

I. Related Literature

It is widely acknowledged that ascending auctions are simpler for real bidders than second-price sealed-bid auctions (Ausubel 2004). Laboratory experiments have investigated and corroborated this claim (Kagel, Harstad, and Levin 1987; Kagel and Levin 1993). More generally, laboratory subjects find it difficult to reason state-by-state about hypothetical scenarios (Charness and Levin 2009; Esponda and Vespa 2014; Ngangoue and Weizsäcker 2015). This mental process, often called "contingent reasoning," has received little formal treatment in economic theory.[7]

There is also a strand of literature, including Vickrey's seminal paper, that observes that sealed-bid auctions raise problems of commitment (Vickrey 1961; Rothkopf and Harstad 1995; Cramton 1998). For instance, it may be difficult to prevent shill bidding without third-party verification. Rothkopf, Teisberg, and Kahn (1990) argue that "robustness in the face of cheating and of fear of cheating is important in determining auction form."

This paper formalizes and unifies these two strands of thought. It shows that mechanisms that do not require contingent reasoning are identical to mechanisms that can be run under bilateral commitment.

For combinatorial auctions, Vickrey-Clarke-Groves mechanisms can be strategically complex and computationally infeasible. Consequently, there has been substantial interest in designing "simple" mechanisms that perform well, such as deferred-acceptance clock auctions (Milgrom and Segal 2015; Gkatzelis, Markakis, and Roughgarden 2017) and posted-price mechanisms (Bartal, Gonen, and Nisan 2003; Feldman, Gravin, and Lucier 2014; Dütting et al. 2016). These have the (previously unmodeled) advantage of being obviously strategy-proof.

OSP is distinct from O-solvability, a solution concept used in the literature on decentralized learning (Friedman and Shenker 1996; Friedman 2002). Strategy S_i

[7] In subsequent work, Esponda and Vespa (2016) investigate axioms governing contingent reasoning in single-agent decision problems.

overwhelms S_i' if the worst possible outcome from S_i is strictly better than the best possible outcome from S_i'. O-solvability calls for the iterated deletion of overwhelmed strategies. One difference between the two concepts is that O-solvability is for normal form games, whereas OSP invokes a notion of an "earliest point of departure," which is only defined in the extensive form. O-solvability is too strong for our current purposes, because almost no games studied in mechanism design are O-solvable.

There is a small literature that builds on this paper. Ashlagi and Gonczarowski (2015) and Troyan (2016) study the OSP-implementation of matching algorithms; deferred acceptance and top trading cycles require additional assumptions about preferences in order to be OSP-implemented. Pycia and Troyan (2016) characterize OSP mechanisms in settings with no transfers and "rich" preferences, and propose a stronger solution concept (strong obvious strategy-proofness). Bade and Gonczarowski (2016) characterize OSP mechanisms for a variety of social choice problems. Zhang and Levin (2017) provide decision-theoretic foundations for obvious dominance. Li (2017) defines obvious ex post equilibrium for settings with interdependent values.

II. Definition and Characterization

The planner operates in an *environment* consisting of

(i) a set of agents, $N \equiv \{1, \ldots, n\}$;

(ii) a set of outcomes, X;

(iii) a set of type profiles, $\Theta_N \equiv \prod_{i \in N} \Theta_i$;

(iv) a utility function for each agent, $u_i : X \times \Theta_i \to \mathbb{R}$.

A mechanism is an *extensive game form with consequences in* X.[8] This is an extensive game form where each terminal history z results in some outcome $g(z) \in X$. We restrict attention to game forms with perfect recall[9] and finite depth.[10] The full definition is familiar to most readers, so we relegate it to the Appendix. The term $\mathcal{G}$ denotes the set of all such game forms, with representative element G. Useful notation is compiled in Table 1.

[8] By adopting this paradigm, we assume that the mechanism describes the *entire* strategic interaction between the agents. For instance, after the mechanism has concluded, agent 1 cannot send money to agent 2, and agent 2 cannot throw a brick through agent 1's window. Such postgame moves often imply that agents do not have dominant strategies, let alone obviously dominant strategies. Savage (1954) notes that "the use of modest little worlds, tailored to particular contexts, is often a simplification, the advantage of which is justified by a considerable body of mathematical experience with related ideas."

[9] G has *perfect recall* if for any information set I_i, for any two histories h and h' in I_i, $\psi_i(h) = \psi_i(h')$, where $\psi(h)$ is the experience of agent i at history h (see Definition 8).

[10] That is, for each game form, there exists some number k such that no history has more than k predecessors. This restriction is not essential to the main results, but simplifies the metagame in Theorem 2.

TABLE 1—NOTATION FOR EXTENSIVE GAME FORMS

Name	Notation	Representative element
Histories	H	h
Precedence relation over histories	$\prec$	
Immediate successors of h	$\sigma(h)$	
Initial history	$h_\emptyset$	
Terminal histories	Z	z
Outcome resulting from z	$g(z)$	
Player (agent or chance) called to play at h	$P(h)$	
Information sets for agent i	$\mathcal{I}_i$	I_i
Actions available at I_i	$A(I_i)$	
Most recent action at h	$\mathcal{A}(h)$	
Probability measure for chance moves	δ_c	
Realization of chance moves	d_c	

We write $I_i \prec I_i'$ if there exist histories $h \in I_i$ and $h' \in I_i'$ such that $h \prec h'$. We write $I_i \prec h$ if there exists $h' \in I_i$ such that $h' \prec h$. Precedence relation $h \prec I_i$ is defined symmetrically. We use $\preceq$ to denote the corresponding weak order.

A *strategy* $S_i(\cdot)$ chooses an action at every information set for agent i, $S_i(I_i) \in A(I_i)$. A *strategy profile* $S_N = (S_i)_{i \in N}$ specifies a strategy for each agent.

A *type-strategy* $\mathbf{S}_i(\cdot)$ specifies a strategy for every type of agent i, where $\mathbf{S}_i(\theta_i)$ denotes the strategy assigned to type θ_i. A *type-strategy profile* $\mathbf{S}_N = (\mathbf{S}_i)_{i \in N}$ specifies a type-strategy for each agent.

Let $z^G(h, S_N, \delta_c)$ be the lottery over terminal histories that results in game form G when we start from h and play proceeds according to (S_N, δ_c). Here, $z^G(h, S_N, d_c)$ is the result of one realization of the chance moves under δ_c. We sometimes write this as $z^G(h, S_i, S_{-i}, d_c)$.

Let $u_i^G(h, S_i, S_{-i}, d_c, \theta_i) \equiv u_i(g(z^G(h, S_i, S_{-i}, d_c)), \theta_i)$. This is the utility to agent i in game form G, when we start at history h, play proceeds according to (S_i, S_{-i}, d_c), and the resulting outcome is evaluated according to preferences θ_i.

DEFINITION 1: *Given G and θ_i, S_i is* ***weakly dominant*** *if* $\forall S_i', \forall S_{-i}$:

$$E_{\delta_c}[u_i^G(h_\emptyset, S_i', S_{-i}, d_c, \theta_i)] \leq E_{\delta_c}[u_i^G(h_\emptyset, S_i, S_{-i}, d_c, \theta_i)]. \tag{2}$$

DEFINITION 2 (Earliest Points of Departure): $I_i \in \alpha(S_i, S_i')$ *if and only if*

(*i*) $S_i(I_i) \neq S_i'(I_i)$;

(*ii*) *There exist* S_{-i}, d_c *such that* $I_i \prec z^G(h_\emptyset, S_i, S_{-i}, d_c)$;

(*iii*) *There exist* S_{-i}, d_c *such that* $I_i \prec z^G(h_\emptyset, S_i', S_{-i}, d_c)$.

The earliest points of departure $\alpha(S_i, S_i')$ are the information sets such that S_i and S_i' choose different actions, that can be on the path of play given S_i and given S_i'. Under perfect recall, this implies that S_i and S_i' chose the same action at every

previous information set, so there is no earlier point of departure. The definition can be extended to deal with mixed strategies,[11] but pure strategies suffice for our current purposes.

DEFINITION 3: *Given G and* θ_i, S_i *is* ***obviously dominant*** *if* $\forall S_i'$, $\forall I_i \in \alpha(S_i, S_i')$:

$$(3) \qquad \sup_{h \in I_i, S_{-i}, d_c} u_i^G(h, S_i', S_{-i}, d_c, \theta_i) \leq \inf_{h \in I_i, S_{-i}, d_c} u_i^G(h, S_i, S_{-i}, d_c, \theta_i).$$

In words, S_i is obviously dominant if, for any deviating strategy S_i', conditional on reaching any earliest point of departure, the best possible outcome under S_i' is no better than the worst possible outcome under S_i.[12]

Compare Definition 1 and Definition 3. Weak dominance is defined using $h_\emptyset$, the history that begins the game. Consequently, if two extensive games have the same normal form, then they have the same weakly dominant strategies. Obvious dominance is defined with histories that are in information sets that are earliest points of departure. Thus, two extensive games with the same normal form may not have the same obviously dominant strategies.[13] Switching to a direct revelation mechanism may not preserve obvious dominance, so the standard revelation principle does not apply.[14]

Weak dominance treats chance moves and other players asymmetrically. Suppose G has a Bayes-Nash equilibrium, and we replace all the agents in $N \backslash 1$ with chance moves drawn from the Bayes-Nash equilibrium distribution. Then agent 1 has a weakly dominant strategy. By contrast, obvious dominance treats chance moves and other players symmetrically.

A *choice rule* is a function $f : \Theta_N \to X$. If we consider stochastic choice rules, then it is a function $f : \Theta_N \to \Delta X$.[15]

A *solution concept* $\mathcal{C}(\cdot)$ is a set-valued function; for each G, it specifies a set of type-strategy profiles $\mathcal{C}(G)$, which may be an empty set.

DEFINITION 4: $(G, \mathbf{S}_N)$ ***C-implements*** *f if*

(i) $\mathbf{S}_N \in \mathcal{C}(G)$;

[11] Three modifications are necessary: first, we change requirement (i) to be that both strategies specify different *probability measures* at I_i. Second, we adapt requirements (ii) and (iii) to hold for *some realization* of the mixed strategies. Finally, we include the recursive requirement, "There does not exist $I_i' \prec I_i$ such that $I_i' \in \alpha(S_i, S_i')$."

[12] Obvious dominance is related to conditional dominance (Shimoji and Watson 1998); S_i conditionally dominates S_i' at I_i if, for any S_{-i} consistent with reaching I_i, the payoff under S_i' is no better than the payoff under S_i.

[13] Two extensive games have the same reduced normal form if and only if they can be made identical using a small set of elementary transformations (Thompson 1952; Elmes and Reny 1994). Which of these transformations does not preserve obvious dominance? Elmes and Reny (1994) propose three such transformations, INT, COA, and ADD, which preserve perfect recall. Brief inspection reveals that obvious dominance is invariant under INT and COA, but varies under ADD.

[14] Glazer and Rubinstein (1996) argue that extensive games can be easier to dominance-solve than normal-form games, because backward induction provides guidance about the correct order to delete strategies. The standard revelation principle does not apply to their solution concept either.

[15] For readability, we generally suppress the latter notation, but the claims that follow hold for both deterministic and stochastic choice rules. Additionally, the set X could *itself* be a set of lotteries. The interpretation of this is that the planner can carry out one-time public lotteries at the end of the mechanism, where the randomization is observable and verifiable.

$(ii)\ \forall \theta_N \in \Theta_N$: $f(\theta_N) = g(z^G(h_\emptyset, (\mathbf{S}_i(\theta_i))_{i\in N}, \delta_c))$.

Similarly, we will say that G $\mathcal{C}$**-implements** f if there exists $\mathbf{S}_N$ that satisfies the requirements above. Function f is $\mathcal{C}$**-implementable** if there exists $(G, \mathbf{S}_N)$ that satisfies the requirements above.

Note that our concern is with weak implementation: we require that $\mathbf{S}_N \in \mathcal{C}(G)$, not $\{\mathbf{S}_N\} = \mathcal{C}(G)$. This is to preserve the analogy with canonical results for strategy-proofness, many of which assume weak implementation (Myerson 1981; Saks and Yu 2005).

DEFINITION 5 (Strategy-Proof): $\mathbf{S}_N \in \mathbf{SP}(G)$ *if for all* i *and for all* θ_i, $\mathbf{S}_i(\theta_i)$ *is weakly dominant.*

DEFINITION 6 (Obviously Strategy-Proof): $\mathbf{S}_N \in \mathbf{OSP}(G)$ *if for all* i *and for all* θ_i, $\mathbf{S}_i(\theta_i)$ *is obviously dominant.*

A mechanism is weakly group-strategy-proof if there does not exist a coalition that could deviate and all be *strictly* better off ex post.

DEFINITION 7 (Weakly Group-Strategy-Proof): $\mathbf{S}_N \in \mathbf{WGSP}(G)$ *if there does not exist a coalition* $\hat{N} \subseteq N$, *type profile* θ_N, *deviating strategies* $\hat{S}_{\hat{N}}$, *noncoalition strategies* $S_{N\setminus\hat{N}}$, *and chance moves* d_c *such that for all* $i \in \hat{N}$:

$$(4)\qquad u_i^G(h_\emptyset, \hat{S}_{\hat{N}}, S_{N\setminus\hat{N}}, d_c, \theta_i) > u_i^G(h_\emptyset, \mathbf{S}_{\hat{N}}(\theta_{\hat{N}}), S_{N\setminus\hat{N}}, d_c, \theta_i).$$

Obvious strategy-proofness implies weak group-strategy-proofness.[16]

PROPOSITION 1: *If* $\mathbf{S}_N \in \mathbf{OSP}(G)$, *then* $\mathbf{S}_N \in \mathbf{WGSP}(G)$.

PROOF:

Suppose $\mathbf{S}_N \notin \mathbf{WGSP}(G)$. Then there is a coalition $\hat{N}$ with types $\hat{\theta}_{\hat{N}}$ that could jointly deviate to strategies $\hat{S}_{\hat{N}}$ and all be strictly better off. Fix $S_{N\setminus\hat{N}}$ and d_c such that all agents in the coalition are strictly better off. Along the resulting terminal history, there is a first agent i in the coalition to deviate from $\mathbf{S}_i(\theta_i)$ to $\hat{S}_i$. That first deviation happens at some information set $I_i \in \alpha(\mathbf{S}_i(\theta_i), \hat{S}_i)$. Agent i strictly gains from that deviation, so $\mathbf{S}_N \notin \mathbf{OSP}(G)$.[17] ∎

COROLLARY 1: *If* $\mathbf{S}_N \in \mathbf{OSP}(G)$, *then* $\mathbf{S}_N \in \mathbf{SP}(G)$.

Proposition 1 suggests a question: is a choice rule OSP-implementable if and only if it is WGSP-implementable? Proposition 5 shows that this is not so.

[16] Barberà, Berga, and Moreno (2016) note that "many well-known individually strategy-proof mechanisms are also group strategy-proof, even if the latter is in principle a much stronger condition than the former."

[17] I thank Ilya Segal for suggesting this concise proof.

A. *Cognitive Limitations*

In what sense is obvious dominance obvious? Intuitively, to see that S_i weakly dominates S_i', the agent must compare *for each contingency* the outcome under S_i and the outcome under S_i'. By contrast, if S_i obviously dominates S_i', then when the two strategies first diverge, every outcome under S_i is at least as good as every outcome under S_i'. Thus, the agent can see that S_i is better than S_i', even if he does not understand how each strategy's outcome depends on unobserved contingencies. Obvious dominance can be recognized even if the agent has a simplified mental model of the world. We now make this point rigorously.

DEFINITION 8: *$\psi_i(h)$ denotes i's **experience** along history h; it is the sequence of information sets where i was called to play, and the actions that i took, in order. We construct this by starting at $h_\emptyset$, and moving step-by-step through predecessors of h. At each $h' \preceq h$, if $P(h') = i$, then we add the current information set to the sequence. If i has just played an action at h', then we add that action to the sequence.*

We use Ψ_i to denote the set $\{\psi_i(h) \,|\, h \in H\} \cup \psi_\emptyset$, where $\psi_\emptyset$ is the empty sequence.[18] *We define $\psi_i(I_i) := \psi_i(h) \,|\, h \in I_i$.*

We define an equivalence relation between mechanisms. In words, G and G' are i-indistinguishable if there exists a bijection from i's information sets and actions in G onto i's information sets and actions in G', such that

(i) ψ_i is an experience in G if and only if ψ_i's image is an experience in G';

(ii) Outcome x could follow experience ψ_i in G if and only if x could follow ψ_i's image in G'.

DEFINITION 9: *Take any $G, G' \in \mathcal{G}$, with information partitions $\mathcal{I}_i, \mathcal{I}_i'$ and experience sets Ψ_i, Ψ_i'. G and G' are **i-indistinguishable** if there exists a bijection $\lambda_{G,G'}$ from $\mathcal{I}_i \cup A(\mathcal{I}_i)$ to $\mathcal{I}_i' \cup A'(\mathcal{I}_i')$ such that*

(i) $\psi_i \in \Psi_i$ if and only if $\lambda_{G,G'}(\psi_i) \in \Psi_i'$;

(ii) $\exists z \in Z : g(z) = x$, $\psi_i(z) = \psi_i$ if and only if $\exists z' \in Z' : g'(z') = x$, $\psi_i'(z') = \lambda_{G,G'}(\psi_i)$,

where we use $\lambda_{G,G'}(\psi_i)$ to denote the sequence produced by passing every element of ψ_i through $\lambda_{G,G'}$.

For G and G' that are i-indistinguishable, we define $\lambda_{G,G'}(S_i)$ to be the strategy that, given information set I_i' in G', plays $\lambda_{G,G'}(S_i(\lambda_{G,G'}^{-1}(I_i')))$.

[18] Mandating the inclusion of the empty sequence has the following consequence: by looking at the set Ψ_i, it is not possible to infer whether $P(h_\emptyset) = i$.

The next theorem states that obviously dominant strategies are the strategies that can be recognized as weakly dominant, by an agent who has this simplified mental model of the world.

THEOREM 1: *For any i, θ_i it holds that S_i is obviously dominant in G if and only if for every G' that is i-indistinguishable from G, $\lambda_{G,G'}(S_i)$ is weakly dominant in G'.*

The "if" direction permits a constructive proof. Suppose S_i is not obviously dominant in G. Then there is some deviating strategy S_i' and some earliest point of departure I_i where the obvious dominance inequality does not hold. We construct a game form G' that is i-indistinguishable from G, such that, for some opponent strategies, if i plays $\lambda_{G,G'}(S_i)$ then play proceeds according to the worst-case scenario (consistent with reaching I_i), and if i deviates at $\lambda_{G,G'}(I_i)$, then play proceeds according to some best-case scenario (consistent with reaching I_i). The key is to find a general construction that always remains in the same equivalence class. The "only if" direction proceeds as follows. Suppose there exists some G' in the equivalence class of G, where $\lambda_{G,G'}(S_i)$ is not weakly dominant. There exists some earliest information set in G' where i could gain by deviating. We then use $\lambda_{G,G'}^{-1}$ to locate an information set in G, and a deviation S_i', that do not satisfy the obvious dominance inequality. The online Appendix provides the details.

One interpretation of Theorem 1 is that obviously dominant strategies are those that can be recognized as dominant given only a *partial description* of the game form. In particular, this partial description specifies, at every information set for agent i, the possible outcomes of each continuation strategy, but omits how those outcomes depend on the (unobserved) state—the opponent strategies and the chance moves. Thus, agent i is unable to break the state space into separate contingencies and compare strategies one contingency at a time. In this sense, obviously dominant strategies can be recognized as dominant without contingent reasoning.

Another interpretation of Theorem 1 is that obviously dominant strategies are robust to *local misunderstandings*. Suppose the agent could mistake any G for any i-indistinguishable G'. He has some belief about how his opponents are playing in G', and best responds to that belief. When is his (true) dominant strategy still a best response, for any such local mistake? By Theorem 1, this holds if and only if it is obviously dominant. For example, if the agent misunderstands the pricing rule in a second-price auction, then truthful bidding may not be a best response.[19] By contrast, in an ascending auction, the agent can substantially misunderstand how his clock price relates to other agents' clock prices, but truthful bidding is still a best response.

Many solution concepts in behavioral game theory specify that agents understand the game form, but best-respond to mistaken beliefs about opponent strategies.[20] Since a dominant strategy is always a best response, such theories predict no

[19] A third-price auction is i-indistinguishable from a second-price auction, and in a third-price auction bidding above one's value is a best response to the (symmetric) Nash equilibrium opponent strategies (Kagel and Levin 1993).

[20] Such concepts include level-k equilibrium (Stahl and Wilson 1994, 1995; Nagel 1995), cognitive hierarchy equilibrium (Camerer, Ho, and Chong 2004), cursed equilibrium (Eyster and Rabin 2005; Esponda 2008), and analogy-based expectation equilibrium (Jehiel 2005).

mistakes in strategy-proof mechanisms.[21] Obvious dominance captures misunderstandings about the game form, and is thus orthogonal to these theories.

B. *Supported by Bilateral Commitments*

Suppose the following *augmented* game form $\tilde{G}$ with consequences in X: as before we have a set of agents N, outcomes X, and preference profiles $\prod_{i\in N}\Theta_i$. However, there is one player in addition to N: Player 0, the Planner.

The Planner has an arbitrarily rich message space M. At the start of the game, each agent $i \in N$ privately observes θ_i. Play proceeds as follows:

(i) The Planner chooses one agent $i \in N$ and sends a query $m \in M$, along with a set of acceptable replies $R \subset M$.[22]

(ii) i observes (m, R), and chooses a reply $r \in R$.

(iii) The Planner observes r.

(iv) The Planner either selects an outcome $x \in X$, or chooses to send another query:
   (a) If the Planner selects an outcome, the game ends.
   (b) If the Planner chooses to send another query, go to Step (i).

For $i \in N$, i's type-strategy specifies what reply to give, as a function of his preferences, the past sequence of queries and replies between him and the Planner, and the current (m, R). That is,

$$\tilde{\mathbf{S}}_i(\theta_i, (m_k, R_k, r_k)_{k=1}^{t-1}, m_t, R_t) \in R_t. \tag{5}$$

Analogously, a strategy $\tilde{S}_i$ depends on $((m_k, R_k, r_k)_{k=1}^{t-1}, m_t, R_t)$, but not on θ_i. We use $\tilde{\mathbf{S}}_i^{\theta_i}$ to denote the strategy played by type θ_i of agent i. We abbreviate $(\tilde{\mathbf{S}}_i^{\theta_i})_{i\in N} \equiv \tilde{\mathbf{S}}_N^{\theta_N}$.

Here, $\tilde{S}_0$ denotes a pure strategy for the Planner; where $\mathbb{S}_0$ denotes the set of all pure strategies. We restrict the Planner to send only finitely many queries.

The standard full commitment paradigm is equivalent to allowing the Planner to commit to a unique $\tilde{S}_0 \in \mathbb{S}_0$ (or some probability measure over $\mathbb{S}_0$). Instead, we assume that for each agent, the Planner can commit to a subset $\hat{\mathbb{S}}_0^i \subseteq \mathbb{S}_0$ that is measurable with respect to that agent's observations in the game.

This is formalized as follows. Each $(\tilde{S}_0, \tilde{S}_N)$ results in some observation $o_i \equiv (o_i^C, o_i^X)$, consisting of a communication sequence between the Planner and agent i, $o_i^C = (m_k, R_k, r_k)_{k=1}^T$ for $T \in \mathbb{N}$, as well as some outcome $o_i^X \in X$.[23] The term $\mathcal{O}_i$ denotes the set of all possible observations for agent i. We define

[21] With the exception of level-0 agents in level-k and cognitive hierarchy models.

[22] This game could be made simpler without altering Theorem 2; the Planner could send only a set of acceptable replies $R \subset M$. Any information contained in the query could simply be "written into" every acceptable reply. However, distinguishing queries and replies makes the exposition more intuitive.

[23] The communication sequence might be empty, which we represent using $T = 0$.

$\phi_i : \mathbb{S}_0 \times \mathbb{S}_N \to \mathcal{O}_i$, where $\phi_i(\tilde{S}_0, \tilde{S}_N)$ is the unique observation resulting from $(\tilde{S}_0, \tilde{S}_N)$. Next we define, for any $\hat{\mathbb{S}}_0 \subseteq \mathbb{S}_0$:

$$\Phi_i\left(\hat{\mathbb{S}}_0\right) \equiv \left\{o_i \,|\, \exists \tilde{S}_0 \in \hat{\mathbb{S}}_0 \colon \exists \tilde{S}_N : o_i = \phi_i\left(\tilde{S}_0, \tilde{S}_N\right)\right\}. \tag{6}$$

For any $\hat{\mathcal{O}}_i \subseteq \mathcal{O}_i$:

$$\Phi_i^{-1}\left(\hat{\mathcal{O}}_i\right) \equiv \left\{\tilde{S}_0 \,|\, \forall \tilde{S}_N \colon \phi_i(\tilde{S}_0, \tilde{S}_N) \in \hat{\mathcal{O}}_i\right\}. \tag{7}$$

DEFINITION 10: $\hat{\mathbb{S}}_0$ *is* ***i-measurable*** *if there exists* $\hat{\mathcal{O}}_i$ *such that*

$$\hat{\mathbb{S}}_0 = \Phi_i^{-1}\left(\hat{\mathcal{O}}_i\right). \tag{8}$$

Intuitively, the *i*-measurable subsets of $\mathbb{S}_0$ are those such that, if the Planner deviates, then there exists an agent strategy profile such that agent *i* detects the deviation; that is, *i* has an observation that is not compatible with *any* strategy in the promised set. Formally, the *i*-measurable subsets of $\mathbb{S}_0$ are the σ-algebra generated by Φ_i (where we impose the discrete σ-algebra on $\mathcal{O}_i$).[24]

DEFINITION 11: *A* ***mixed strategy of finite length*** *over* $\hat{\mathbb{S}}_0$ *specifies a probability measure over a subset* $\overline{\mathbb{S}}_0 \subseteq \hat{\mathbb{S}}_0$ *such that: there exists* $k \in \mathbb{N}$ *such that: for all* $\tilde{S}_0 \in \overline{\mathbb{S}}_0$ *and all* $\tilde{S}_N$*:* $(\tilde{S}_0, \tilde{S}_N)$ *results in the Planner sending k or fewer total queries.*

We use $\Delta\hat{\mathbb{S}}_0$ to denote the mixed strategies of finite length over $\hat{\mathbb{S}}_0$. Denote by $\tilde{S}_0^{\Delta}$ an element of such a set.

DEFINITION 12: *A choice rule f is* ***supported by bilateral commitments*** $(\hat{\mathbb{S}}_0^i)_{i \in N}$ *if*

(*i*) *For all* $i \in N$: $\hat{\mathbb{S}}_0^i$ *is i-measurable*;

(*ii*) *There exist* $\tilde{S}_0^{\Delta}$, *and* $\tilde{\mathbf{S}}_N$ *such that*

- (*a*) *For all* θ_N: $(\tilde{S}_0^{\Delta}, \tilde{\mathbf{S}}_N^{\theta_N})$ *results in* $f(\theta_N)$.
- (*b*) $\tilde{S}_0^{\Delta} \in \Delta \cap_{i \in N} \hat{\mathbb{S}}_0^i$.
- (*c*) *For all* $i \in N$, θ_i, $\tilde{S}'_{N\setminus i}$, $\tilde{S}_0^{\Delta\prime} \in \Delta\hat{\mathbb{S}}_0^i$: $\tilde{\mathbf{S}}_i^{\theta_i}$ *is a best response to* $(\tilde{S}_0^{\Delta\prime}, \tilde{S}'_{N\setminus i})$ *given preferences* θ_i.

Requirement (*ii.a*) is that the Planner's mixed strategy and the agent's pure strategies result in the (distribution over) outcomes required by the choice rule. Requirement (*ii.b*) is that the Planner's strategy is a (possibly degenerate) mixture

[24] Notice that an observation for player *i* is simply a sequence of messages and responses, followed by an outcome. It does not include additional information about calendar time. A consequence of this formulation is that the Planner is unable to commit to the order in which she approaches the players.

over pure strategies compatible with *every* bilateral commitment.[25] $(ii.c)$ is that each agent i's assigned strategy is weakly dominant, when we consider the Planner as a player restricted to playing mixtures over strategies in $\hat{\mathbb{S}}_0^i$.

"Supported by bilateral commitments" is just one of many partial commitment regimes. This one requires that the commitment offered to each agent is measurable with respect to events that he can observe. In reality, contracts are seldom enforceable unless each party can observe breaches. Thus, "supported by bilateral commitments" is a natural case to study.

THEOREM 2: *f is OSP-implementable if and only if there exist bilateral commitments $(\hat{\mathbb{S}}_0^i)_{i\in N}$ that support f.*

The intuition behind the proof is as follows. A bilateral commitment $\hat{\mathbb{S}}_0^i$ is essentially equivalent to the Planner committing to "run" only games in some i-indistinguishable equivalence class of $\mathcal{G}$. Consequently, we can find a set of bilateral commitments that support f if and only if we can find some $(G, \mathbf{S}_N)$ such that, for every i, for every θ_i, for every G' that is i-indistinguishable from G, $\lambda_{G,G'}(\mathbf{S}_i(\theta_i))$ is weakly dominant in G'. By Theorem 1, this holds if and only if f is OSP-implementable. The online Appendix provides the details.

C. *The Pruning Principle*

When seeking SP-implementation, we can without loss of generality restrict attention to the class of direct revelation mechanisms, by the revelation principle. The standard revelation principle does not hold for OSP mechanisms. As the second-price auction illustrates, there exist OSP-implementable choice rules that cannot be implemented via a direct revelation mechanism.

However, there is a weaker principle that substantially simplifies the analysis. Here we define the *pruning* of a mechanism with respect to a type-strategy profile. If, for all type profiles, a history is never reached, then we delete that history (and redefine the other parts of the tuple so that what remains is well-formed).

DEFINITION 13 (Pruning): *Take any $G = \langle H, \prec, A, \mathcal{A}, P, \delta_c, (\mathcal{I}_i)_{i\in N}, g\rangle$, and $\mathbf{S}_N$. $\mathcal{P}(G, \mathbf{S}_N) \equiv \langle \tilde{H}, \tilde{\prec}, \tilde{A}, \tilde{\mathcal{A}}, \tilde{P}, \tilde{\delta}_c, (\tilde{\mathcal{I}}_i)_{i\in N}, \tilde{g}\rangle$ is the **pruning** of G with respect to $\mathbf{S}_N$, constructed as follows:*

(i) $\tilde{H} = \{h \in H \mid \exists \theta_N, d_c$ such that $h \preceq z^G(h_\emptyset, (\mathbf{S}_i(\theta_i))_{i\in N}, d_c)\}$.

(ii) *For all i, if $I_i \in \mathcal{I}_i$ then $(I_i \cap \tilde{H}) \in \tilde{\mathcal{I}}_i$.*[26]

(iii) $(\tilde{\prec}, \tilde{A}, \tilde{\mathcal{A}}, \tilde{P}, \tilde{\delta}_c, \tilde{g})$ *are* $(\prec, A, \mathcal{A}, P, \delta_c, g)$ *restricted to $\tilde{H}$.*

[25] This requirement prevents the Planner from extracting arbitrary information by making promises and breaking them. Otherwise, the Planner could promise to implement some constant outcome, ask every agent to report their type, and then do whatever she pleased with that information.

[26] Note that the initial history $h_\emptyset$ is distinct from the empty set. That is to say, $(I_i \cap \tilde{H}) = \emptyset$ does *not* entail that $\{h_\emptyset\} \in \tilde{I}_i$.

It turns out that, if some mechanism OSP-implements a choice rule, then the pruning of that mechanism with respect to the equilibrium strategies OSP-implements that same choice rule. Thus, while we cannot restrict attention to direct revelation mechanisms, we can restrict attention to "minimal" mechanisms, such that no histories are off the path of play. This is used both in this paper and in the subsequent literature[27] to state clean results.

PROPOSITION 2 (The Pruning Principle): *Let* $\tilde{G} \equiv \mathcal{P}(G, \mathbf{S}_N)$, *and* $\tilde{\mathbf{S}}_N$ *be* $\mathbf{S}_N$ *restricted to* $\tilde{G}$. *If* $(G, \mathbf{S}_N)$ *OSP-implements* f, *then* $(\tilde{G}, \tilde{\mathbf{S}}_N)$ *OSP-implements* f.

III. Applications

A. *Binary Allocation Problems*

We now consider a canonical environment, $(N, X, \Theta_N, (u_i)_{i\in N})$. Let $Y \subseteq 2^N$ be the set of feasible *allocations*, with representative element $y \in Y$. An outcome consists of an allocation $y \in Y$ and a transfer for each agent, $X = Y \times \mathbb{R}^n$. Define $t \equiv (t_i)_{i\in N}$ to denote a profile of transfers.

Preferences are quasilinear. Let $\Theta_N = \prod_{i\in N} \Theta_i$, where $\Theta_i = [\underline{\theta}_i, \overline{\theta}_i]$, for $0 \leq \underline{\theta}_i < \overline{\theta}_i < \infty$. For $\theta_i \in \Theta_i$,

$$u_i(\theta_i, y, t) = 1_{i\in y}\theta_i + t_i. \tag{9}$$

For instance, in a private value auction with unit demand, $i \in y$ if and only if agent i receives at least one unit of the good under allocation y. In a procurement auction, $i \in y$ if and only if i does not incur costs of provision under allocation y. The term θ_i is agent i's cost of provision (equivalently, benefit of nonprovision). In a binary public goods game, $Y = \{\emptyset, N\}$.

An *allocation rule* is a function $f_y : \Theta \to Y$. A choice rule is thus a combination of an allocation rule and a payment rule, $f = (f_y, f_t)$, where $f_t : \Theta \to \mathbb{R}^n$. Similarly, for each game form G, we disaggregate the outcome function, $g = (g_y, g_t)$. In this part, we concern ourselves only with deterministic allocation rules and payment rules, and thus suppress notation involving δ_c and d_c.

DEFINITION 14: *An allocation rule* f_y *is* $\mathcal{C}$*-implementable if there exists* f_t *such that* (f_y, f_t) *is* $\mathcal{C}$*-implementable.* G $\mathcal{C}$*-implements* f_y *if there exists* f_t *such that* G $\mathcal{C}$*-implements* (f_y, f_t).

In a binary allocation problem, f_y is SP-implementable if and only if f_y is monotone.[28] This result is implicit in Spence (1974) and Mirrlees (1971), and is proved explicitly in Myerson (1981).[29] Moreover, if an allocation rule f_y is SP-implementable, then the accompanying transfer rule f_t is essentially unique:

[27] Ashlagi and Gonczarowski (2015); Bade and Gonczarowski (2016); Pycia and Troyan (2016).

[28] f_y is monotone if for all i, for all θ_{-i}, $1_{i\in f_y(\theta_i, \theta_{-i})}$ is weakly increasing in θ_i.

[29] These monotonicity results for are for weak SP-implementation rather than full SP-implementation implementation. Weak SP-implementation requires $\mathbf{S}_N \in \mathrm{SP}(G)$. Full SP-implementation requires $\mathbf{S}_N = \mathrm{SP}(G)$. There are monotone allocation rules for which the latter requirement cannot be satisfied. For example, suppose two agents

$$f_{t,i}(\theta_i, \theta_{-i}) = -1_{i \in f_y(\theta)} \inf\left\{\theta_i' \mid i \in f_y(\theta_i', \theta_{-i})\right\} + r_i(\theta_{-i}), \tag{10}$$

where r_i is some arbitrary function of the other agents' preferences. This follows easily by arguments similar to those in Green and Laffont (1977) and Holmström (1979).

In this standard environment, what happens when we require OSP-implementation? In particular, what restrictions does OSP-implementation place on the transfer rules and the extensive game form?

In binary allocation problems, every OSP mechanism is essentially a *personal-clock auction*. It is personal in two respects: firstly, the clock price and closing rule could vary across agents. Secondly, the clock price could be associated with different consequences for different agents. The full statement of the definition is novel, but we will build it out of familiar parts.

Consider how the ascending auction appears to a single bidder: at each point, there is a "going transfer" associated with being in the allocation (winning the object). The agent either plays *quit* or *continue*; if he plays *quit*, then he is out of the allocation and makes no payments. If he plays *continue*, then either he is in the allocation and has the going transfer, or the going transfer falls (the going price rises) and he faces the same decision again.

We can generalize this procedure a bit, while still ensuring that the agent has an obviously dominant strategy. The going transfer could fall in any increments, and how much it falls could depend on other agents' actions. The transfer associated with being out of the allocation could be nonzero, as long as it is some (known) fixed amount. Even if the agent chooses *continue*, we could sometimes mandate that he quits, in which case he is out of the allocation (and receives the fixed transfer). The agent need not choose between *continue* and *quit* at every information set; he need only be offered a choice when the going transfer strictly falls. There could be multiple actions that *quit*, all having the same consequence for him.[30] There could be multiple actions that *continue*, provided that the going transfer will not fall in future and at least one such action guarantees that he is in the allocation. The agent could receive arbitrary information about the history of play. We call this generalized procedure *In-Transfer Falls*.

Consider how a descending-price procurement auction appears to a single supplier (holding one unit of an indivisible good): at each point, there is a "going transfer" associated with being *out* of the allocation. The agent either plays *quit* or *continue*; if he plays *quit*, then he is in the allocation (keeps the good) and receives nothing. If he plays *continue*, then either he is out of the allocation (sells the good) and receives the going transfer, or the going transfer falls and he faces the same decision again. Notice that now the transfer associated with being in the allocation is fixed and the transfer associated with being out of the allocation falls monotonically. We could generalize this in the same ways as before, and call the resulting procedure *Out-Transfer Falls*.

with unit demand. Agent 1 receives one unit if and only if $v_1 > 0.5$. Agent 2 receives one unit if and only if $v_2 > v_1$.

[30] Although they could affect what happens to other agents—it is incentive compatible for the agent to reveal any information about his type at the point when he quits, and the allocation rule could in principle depend on this information.

In a *personal-clock auction*, starting from any point an agent first has a nonsingleton information set, that agent either faces *In-Transfer Falls* or *Out-Transfer Falls*.

DEFINITION 15: *G is a* ***personal-clock auction*** *if, for every* $i \in N$, *at every earliest information set* I_i^* *such that* $|A(I_i^*)| > 1$*:*

(i) *Either* (***In-Transfer Falls***) *there exists a fixed transfer* $\bar{t}_i \in \mathbb{R}$, *a going transfer* $\tilde{t}_i : \{I_i | I_i^* \preceq I_i\} \to \mathbb{R}$, *and a set of "quitting" actions* A^q *such that*

(a) *For all z such that* $I_i^* \prec z$,
(i) *either* $i \notin g_y(z)$ *and* $g_{t,i}(z) = \bar{t}_i$,
(ii) *or*, $i \in g_y(z)$ *and*

$$g_{t,i}(z) = \inf_{I_i | I_i^* \preceq I_i \prec z} \tilde{t}_i(I_i). \tag{11}$$

(b) *For all* $a \in A^q$, *for all z such that* $a \in \psi_i(z)$: $i \notin g_y(z)$.

(c) $A^q \cap A(I_i^*) \neq \emptyset$.

(d) *For all* $I_i', I_i'' \in \{I_i | I_i^* \preceq I_i\}$:
(i) *If* $I_i' \prec I_i''$, *then* $\tilde{t}_i(I_i') \geq \tilde{t}_i(I_i'')$.
(ii) *If* $I_i' \prec I_i''$, $\tilde{t}_i(I_i') > \tilde{t}_i(I_i'')$, *and there does not exist* I_i''' *such that* $I_i' \prec I_i''' \prec I_i''$, *then* $A^q \cap A(I_i'') \neq \emptyset$.
(iii) *If* $I_i' \prec I_i''$ *and* $\tilde{t}_i(I_i') > \tilde{t}_i(I_i'')$, *then* $|A(I_i') \backslash A^q| = 1$.
(iv) *If* $|A(I_i') \backslash A^q| > 1$, *then there exists* $a \in A(I_i')$ *such that: for all z such that* $a \in \psi_i(z)$: $i \in g_y(z)$.

(ii) *Or* (***Out-Transfer Falls***), *as above, but we substitute every instance of "*$i \in g_y(z)$*" with "*$i \notin g_y(z)$*" and vice versa.*

Notice what this definition does not require. The going transfer need not be equal across agents. Whether and how much one agent's going transfer changes could depend on other agents' actions. Some agents could face In-Transfer Falls, and other agents could face Out-Transfer Falls (a two-sided clock auction).[31] Which procedure an agent faces could even depend on other agents' past actions.

THEOREM 3: *If* $(G, \mathbf{S}_N)$ *OSP-implements* f_y, *then* $\mathcal{P}(G, \mathbf{S}_N)$ *is a personal-clock auction. If G is a personal-clock auction, then there exist* $\mathbf{S}_N$ *and* f_y *such that* $(G, \mathbf{S}_N)$ *OSP-implements* f_y.

For any normal-form mechanism, there are typically many equivalent extensive forms. The theory of mechanism design seldom provides general criteria to choose between them. In binary allocation problems, obvious dominance pins down many

[31] Loertscher and Marx (2017) investigate clock auctions that are prior-free, asymptotically optimal, and obviously strategy-proof.

extensive-form details of the ascending auction, and provides an answer to the question, "Why are ascending auctions so common?"

B. *Top Trading Cycles*

We now produce an impossibility result in a classic matching environment (Shapley and Scarf 1974). There are n agents in the market, each endowed with an indivisible good. An agent's type is a vector $\theta_i \in \mathbb{R}^n$. Here, Θ_N is the set of all n by n matrices of real numbers. An outcome assigns one object to each agent. If agent i is assigned object k, he has utility θ_i^k. There are no money transfers.

Given preferences θ and agents $R \subseteq N$, a *top trading cycle* is a set $\emptyset \subset R' \subseteq R$ whose members can be indexed in a cyclic order:

$$R' = \{i_1, i_2, \ldots, i_r = i_0\}, \tag{12}$$

such that each agent i_k likes i_{k+1}'s good at least as much as any other good in R. Following Roth (1982), we assume that the algorithm in question has an arbitrary, fixed way of resolving ties.

DEFINITION 16: *f is a top trading cycle rule if, for all θ, $f(\theta)$ is equal to the output of the following algorithm:*

(i) *Set* $R^1 = N$,

(ii) *For* $l = 1, 2, \ldots$:

(a) *choose some top trading cycle* $R' \subseteq R^l$,
(b) *carry out the indicated trades*,
(c) *set* $R^{l+1} := R^l \backslash R'$,
(d) *terminate if* $R^{l+1} = \emptyset$.

The algorithm is of economic interest, because it finds a core allocation in an economy with indivisible goods (Shapley and Scarf 1974).

PROPOSITION 3: *If f is a top trading cycle rule, then there exists G that SP-implements f* (Roth 1982).

PROPOSITION 4: *If f is a top trading cycle rule, then there exists G that WGSP-implements f* (Bird 1984).

PROPOSITION 5: *If f is a top trading cycle rule and $n \geq 3$, then there does not exist G that OSP-implements f.*

PROOF:

OSP-implementability is a hereditary property of functions. That is, if f is OSP-implementable given domain Θ_N, then the subfunction $f' = f$ with domain

$\Theta'_N \subseteq \Theta_N$ is OSP-implementable. Thus, to prove Proposition 5, it suffices to produce a subfunction that is not OSP-implementable.

Consider the following subset $\Theta'_N \subset \Theta_N$. Take agents a, b, c, with endowed goods A, B, C. Agent a has only two possible types, θ_a and θ'_a, such that

$$\text{(13)} \qquad \text{either} \quad B \succ_a C \succ_a A \succ_a \cdots$$

$$\text{or} \quad C \succ_a B \succ_a A \succ_a \cdots.$$

We make the symmetric assumption for b and c.

We now argue by contradiction. Take any G pruned with respect to the truthful strategy profiles, such that (by Proposition 2) G OSP-implements $f' = f$ for domain Θ'. Consider some history h at which $P(h) = a$ with a nonsingleton action set. This cannot come before all such histories for b and c.

Suppose not, and suppose $B \succ_a C$. If a chooses the action corresponding to $B \succ_a C$, and faces opponent strategies corresponding to $C \succ_b A$ and $B \succ_c A$, then a receives good A. If a chooses the action corresponding to $C \succ_a B$, and faces opponent strategies corresponding $A \succ_c B$, then a receives good C. Thus, it is not an obviously dominant strategy to choose the action corresponding to $B \succ_a C$. So a cannot be the first to have a nonsingleton action set.

By symmetry, this argument applies to b and c as well. So all of the action sets for a, b, and c are singletons, and G does not OSP-implement f', a contradiction. ∎

Proposition 5 implies that the OSP-implementable choice rules are not identical to the WGSP-implementable choice rules.

IV. Laboratory Experiment

Are obviously strategy-proof mechanisms easier for real people to understand? The following laboratory experiment provides a straightforward test: we compare pairs of mechanisms that implement the same choice rule. One mechanism in each pair is SP, but not OSP. The other mechanism is OSP. Standard game theory predicts that both mechanisms will produce the same outcome. We are interested in whether subjects play the dominant strategy at higher rates under OSP mechanisms.

A. *Experiment Design*

The experiment is an across-subjects design, comparing three pairs of games. There are four players in each game.

For the first pair, we compare the second-price auction (2P) and the ascending clock auction (AC). In both these games, subjects bid for a money prize. Subjects have induced affiliated private values; if a subject wins the prize, he earns an amount equal to the value of the prize, minus his payments from the auction. For each subject, his value for the prize is equal to a group draw plus a private adjustment. The group draw is uniformly distributed between \$10 and \$110. The private adjustment

is uniformly distributed between \$0 and \$20. All money amounts in these games are in \$0.25 increments. Each subject knows his own value, but not the group draw or the private adjustment.[32]

2P is SP, but not OSP. In 2P, subjects submit their bids simultaneously. The highest bidder wins the prize, and makes a payment equal to the second-highest bid. Bids are constrained to be between \$0 and \$150.[33]

AC is OSP. In AC, the price starts at a low value (the highest \$25 increment that is below the group draw), and counts upward, up to a maximum of \$150. Each bidder can quit at any point.[34] When only one bidder is left, that bidder wins the object at the current price.

Previous studies comparing second-price auctions to ascending clock auctions have small sample sizes, given that when the same subjects play a sequence of auctions, these are plainly not independent observations. Kagel, Harstad, and Levin (1987) compare two groups playing second-price auctions to two groups playing ascending clock auctions. Harstad (2000) compares five groups playing second-price auctions to three groups playing ascending clock auctions. (The comparison is not the main goal of either experiment.) Other studies find similar results for second-price auctions (Kagel and Levin 1993) and for ascending clock auctions (McCabe, Rassenti, and Smith 1990), but these do not directly compare the two formats with the same value distribution and the same subject pool. When we compare 2P and AC, we can see this as a high-powered replication of Kagel, Harstad, and Levin (1987), since we now observe 18 groups playing 2P and 18 groups playing AC.[35]

For the second pair, we compare the second-price plus-X auction (2P+X) and the ascending clock plus-X auction (AC+X). Subjects' values are drawn as before. However, there is an additional random variable X, which is uniformly distributed between \$0 and \$3. Subjects are not told the value of X until after the auction.

In 2P+X, subjects submit their bids simultaneously. The highest bidder wins the prize if and only if his bid exceeds the second-highest bid plus X. If the highest bidder wins the prize, then he makes a payment equal to the second-highest bid plus X. Otherwise, no agent wins the prize, and no payments are made. In 2P+X, submitting a bid equal to your value is a dominant strategy, but it is not obviously dominant.

In AC+X, the price starts at a low value (the highest \$25 increment that is below the group draw), and counts upward. Each bidder can quit at any point.[36] When only one bidder is left, the price *continues to rise* for another X dollars, and then freezes. If the highest bidder keeps bidding until the price freezes, then she wins the prize at the final price. Otherwise, no agent wins the prize and no payments are made. In

[32] I use affiliated private values for two reasons. First, in strategy-proof auctions with independent private values, incentives for truthful bidding are weak for bidders with values near the extremes. Affiliation strengthens incentives for these bidders. Second, Kagel, Harstad, and Levin (1987) use affiliated private values, and the first part of the experiment is designed to replicate their results.

[33] In both 2P and AC, if there is a tie for the highest bid, then no bidder wins the object.

[34] During the auction, each bidder observes the number of active bidders.

[35] I am not aware of any previous laboratory experiment that directly compares second-price and ascending clock auctions, holding constant the value distribution and subject pool, with more than five groups playing each format.

[36] As in AC, each bidder observes the number of active bidders. However, if the number of active bidders is 1 or 2, then the computer display informs bidders that the number of active bidders is "1 or 2."

AC+X, it is an obviously dominant strategy to keep bidding if the price is strictly below your value, and quit otherwise.

Some subjects might find 2P or AC familiar, since such mechanisms occur in some natural economic environments. Differences in subject behavior might be caused by different degrees of familiarity with the mechanism. 2P+X and AC+X are novel mechanisms that subjects are unlikely to find familiar. 2P+X and AC+X can be seen as perturbations of 2P and AC; the underlying choice rule is made more complex while preserving the SP-OSP distinction. Thus, comparing 2P+X and AC+X indicates whether the distinction between SP and OSP mechanisms holds for novel and more complicated auction formats.

In the third pair of games, subjects may receive one of four common-value money prizes. The four prize values are drawn, uniformly at random and without replacement, from the set {$0.00, $0.25, $0.50, $0.75, $1.00, $1.25}. Subjects observe the values of all four prizes at the start of each game.

In a strategy-proof random serial dictatorship (SP-RSD), subjects are informed of their priority score, which is drawn uniformly at random from the integers 1 to 10. They then simultaneously submit ranked lists of the four prizes. Players are processed sequentially, from the highest priority score to the lowest. Ties in priority score are broken randomly. Each player is assigned the highest-ranked prize on his list, among the prizes that have not yet been assigned. It is a dominant strategy to rank the prizes in order of their money value. SP-RSD is SP, but not OSP.[37]

In an obviously strategy-proof random serial dictatorship (OSP-RSD), subjects are informed of their priority score. Players take turns, from the highest priority score to the lowest. When a player takes his turn, he is shown the prizes that have not yet been taken, and picks one of them. It is an obviously dominant strategy to pick the available prize with the highest money value.

SP-RSD and OSP-RSD differ from the auctions in several ways. The auctions are private-value games of incomplete information, whereas SP-RSD and OSP-RSD are common-value games of complete information. In the auctions, subjects face two sources of strategic uncertainty: they are uncertain about their opponents' valuations, and they are uncertain about their opponents' strategies (a function of valuations). By contrast, in SP-RSD and OSP-RSD, subjects face no uncertainty about their opponents' valuations.

Unlike the auctions, SP-RSD and OSP-RSD are constant-sum games, such that one player's action cannot affect total player surplus. Any effect that persists in both the auctions and the serial dictatorships is difficult to explain using social preferences, since such theories typically make different predictions for constant-sum and nonconstant-sum games. Thus, in comparing SP-RSD and OSP-RSD, we test whether the SP-OSP distinction has empirical support in mechanisms that are very different from auctions.

At the start of the experiment, subjects are randomly assigned into groups of four. These groups persist throughout the experiment. Consequently, each group's play can be regarded as a single independent observation in the statistical analysis.

[37] SP-RSD is SP, but not OSP since, if a player swaps the order of the highest and second-highest prizes, he might win the second-highest prize. If he reports his true rank-order list, he might win the third-highest prize.

TABLE 2—MECHANISMS IN EACH TREATMENT (*Ten Rounds*)

Treatment 1	AC	AC+X	OSP-RSD
Treatment 2	2P	2P+X	SP-RSD
Treatment 3	AC	AC+X	SP-RSD
Treatment 4	2P	2P+X	OSP-RSD

Each group either plays ten rounds of AC, followed by ten rounds of AC+X, or plays ten rounds of 2P, followed by ten rounds of 2P+X.[38] At the end of each round, subjects are shown the auction result, their own profit from this round, the winning bidder's profit from this round, and the bids (in order from highest to lowest). Notice that subjects have ten rounds of experience with a standard auction, before being presented with its unusual +X variant. Thus, the data from +X auctions record moderately experienced bidders grappling with a new auction format.

Next, groups are rerandomized into either ten rounds of OSP-RSD or ten rounds of SP-RSD. At the end of each round, subjects see which prize they have obtained, and whether their priority score was the highest, or second-highest, and so on.

Table 2 summarizes the design. Subjects had printed copies of the instructions, and the experimenter read aloud the part pertaining to each 10-round segment just before that segment began. The instructions (correctly) informed subjects that their play in earlier segments would not affect the games in later segments. The instructions did not mention dominant strategies or provide recommendations for how to play, so as to prevent confounds from the experimenter demand effect. Instructions for both SP and OSP mechanisms are of similar length and similar reading levels.[39]

In every SP mechanism, each subject had 90 seconds to make his choice. Each subject could revise his choice as many times as he desired during the 90 seconds, and only his final choice would count. For OSP mechanisms, mean time to completion was 113.0 seconds in AC, 121.4 seconds in AC+X, and 40.5 seconds in OSP-RSD. However, the rules of the OSP mechanisms imply that not every subject was actively choosing throughout that time.

B. *Administrative Details*

Subjects were paid $20 for participating, in addition to their profits or losses from every round of the experiment. On average, subjects made $37.54, including the participation payment. Subjects who made negative net profits received just the $20 participation payment.

I conducted the experiment at the Ohio State University Experimental Economics Laboratory in August 2015, using z-Tree (Fischbacher 2007). I recruited subjects

[38] If a stage game with dominant strategies is repeated finitely many times, then the resulting repeated game typically does not have a dominant strategy. The same holds for obviously dominant strategies. Consequently, in interpreting these results as informing us about dominant strategy play, we invoke an implicit narrow framing assumption. The same assumption is made for other experiments in this literature, such as Kagel, Harstad, and Levin (1987) and Kagel and Levin (1993).

[39] Both sets of instructions are approximately at a fifth-grade reading level according to the Flesch-Kincaid readability test, which is a standard measure for how difficult a piece of text is to read (Kincaid et al. 1975). The instructions are available in the online Appendix.

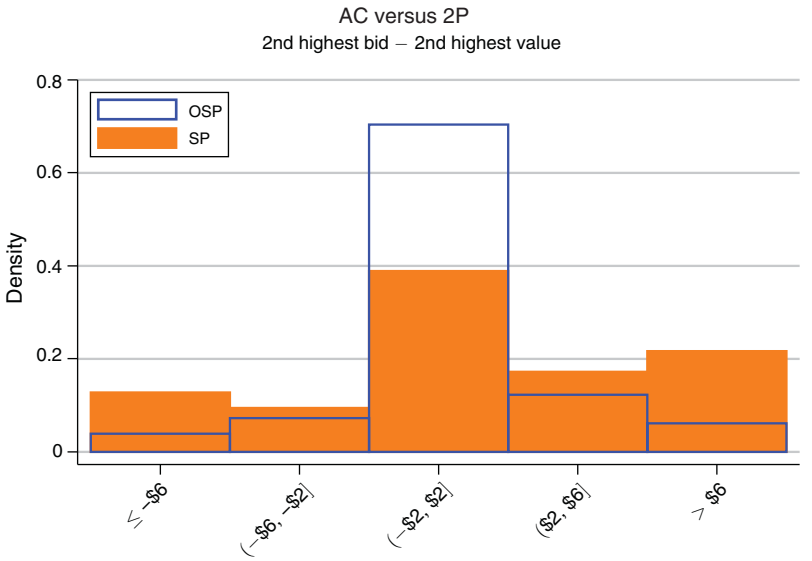

FIGURE 2. STANDARD AUCTIONS

from the student population using an online system (Greiner 2015). I administered 16 sessions, where each session involved 1 to 3 groups. Each session lasted about 90 minutes. In total, the data include 144 subjects in 36 groups of 4 (with 9 groups in each treatment).[40]

C. *Statistical Analysis*

The data include 4 different auction formats, with 180 auctions per format, for a total of 720 auctions.[41]

One natural summary statistic for each auction is the difference between the second-highest bid and the second-highest value. This is, equivalently, the difference between that auction's closing price, and the closing price that would have occurred if all bidders played the dominant strategy. Figure 2 displays histograms of the second-highest bid minus the second-highest value, for AC and 2P. Figure 3 does the same for AC+X and 2P+X. If all agents are playing the dominant strategy in an auction, then the histogram for that auction will be a point mass at zero.

There is a substantial difference between the empirical distributions for OSP and SP mechanisms. If we choose a random auction from the data, how likely is it to have a closing price within $2.00 of the equilibrium price? An auction is 31 percentage points more likely to have a closing price within $2.00 of the equilibrium price under AC (OSP) compared to 2P (SP). An auction is 28 percentage points more likely to have a closing price within $2.00 of the equilibrium price under AC+X

[40] In two cases, network errors caused crashes which prevented a group from continuing in the experiment. I recruited new subjects to replace these groups.

[41] In 2 out of 720 auctions, computer errors prevented bidders from correctly entering their bids. We omit these two observations, but including them has little effect on any of the results that follow.

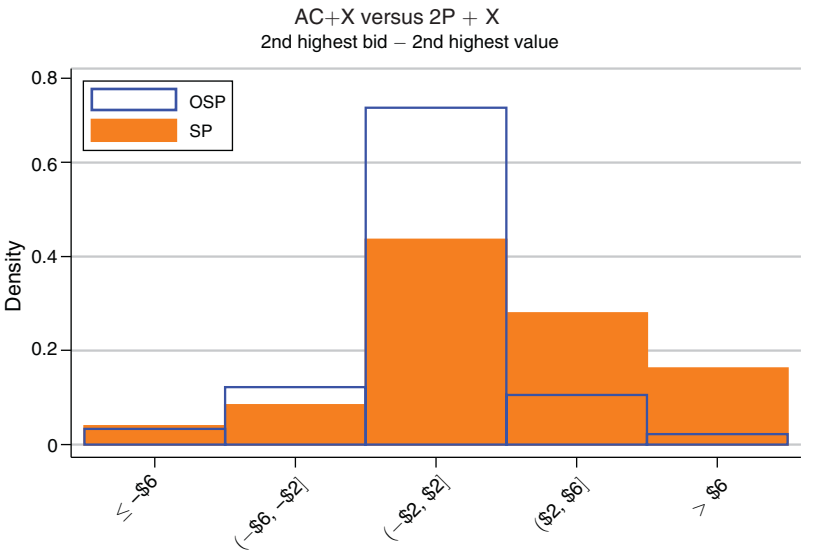

FIGURE 3. +X AUCTIONS

TABLE 3—MEAN(ABS(*2nd bid − 2nd value*))

Format	Rounds	SP	OSP	*p*-value
Auction	1–5	8.04	3.19	0.006
		(1.25)	(1.05)	
	6–10	4.99	1.77	0.016
		(1.18)	(0.33)	
+X Auction	1–5	3.99	1.83	0.006
		(0.60)	(0.41)	
	6–10	3.69	1.29	0.017
		(0.87)	(0.33)	

Notes: For each group, we take the mean absolute difference over each 5-round block. We then compute standard errors counting each group's 5-round mean as a single observation. (Eighteen observations per cell, standard errors in parentheses.) *p*-values are computed using a two-sample *t*-test, allowing for unequal variances. Other empirical strategies yield similar results; see the online Appendix for details.

(OSP) compared to 2P+X (SP). Closing prices under 2P+X are systematically biased upward ($p = 0.0031$).[42]

Table 3 displays the mean absolute difference between the second-highest bid and the second-highest value, for the first five rounds and the last five rounds of each auction. This measures the magnitude of errors under each mechanism. (Alternative measures of errors are in the online Appendix.) Errors are systematically larger under SP than under OSP, and this difference is significant in both the standard

[42] For each group, we take the mean difference between the second-highest bid and the second-highest value. This produces one observation per group playing 2P+X, for a total of 18 observations, and we use a *t*-test for the null that these have zero mean.

Table 4—Proportion of Serial Dictatorships Not Ending in Dominant Strategy Outcome

	SP	OSP	*p*-value
Rounds 1–5	43.3% (7.3%)	7.8% (3.3%)	0.0002
Rounds 6–10	28.9% (5.2%)	6.7% (3.2%)	0.0011
p-value	0.1026	0.7492	

Notes: For each group, for each 5-round block, we record the error rate. We then compute standard errors counting each group's observed error rate as a single observation. (Eighteen observations per cell, standard errors in parentheses.) When comparing SP to OSP, we compute *p*-values using a two-sample *t*-test, allowing for unequal variances. (Alternative empirical strategies yield similar results. See the online Appendix for details.) When comparing early to late rounds of the same game, we compute *p*-values using a paired *t*-test.

auctions and the novel +X auctions, and in both early and late rounds. To build intuition for effect sizes, consider that the expected profit of the winning bidder in 2P and AC is about \$4.00 (given dominant strategy play). Thus, the average errors under 2P are larger than the theoretical prediction for total bidder surplus.

There is some evidence of learning in 2P; errors are smaller in the last five rounds compared to the first five rounds ($p = 0.045$, paired *t*-test). For the other three auction formats, there is no significant evidence of learning.[43]

To compare subject behavior under SP-RSD and OSP-RSD, we compute the proportion of games that *do not* end in the dominant strategy outcome. Under SP-RSD, 36 percent of games do not end in the dominant strategy outcome. Under OSP-RSD, 7 percent of games do not end in the dominant strategy outcome. Table 4 displays the empirical frequency of nondominant strategy outcomes, by format and by 5-round blocks. Deviations from the dominant strategy outcome happen more frequently under SP-RSD than under OSP-RSD, and these differences are highly significant in both early and late rounds.

In SP-RSD, 29 percent of submitted rank-order lists contain errors. The most common error under SP-RSD is to swap the ranks of the highest and second-highest prizes, and report the list in order 2nd–1st–3rd–4th. This error accounts for 18 percent of incorrect rank-order lists.

D. *One-Shot Treatments*

At the suggestion of a referee, I ran additional treatments that investigate SP-RSD and OSP-RSD without repeated play and with higher stakes, at the same laboratory as the original treatments. These treatments consist of a one-shot serial dictatorship, followed by a demographic questionnaire. The stakes are 12 times as high; 4 money prizes are drawn uniformly at random and without replacement from the set {\$0, \$3, \$6, \$9, \$12, \$15}.

[43] $p = 0.173$ for AC, $p = 0.694$ for 2P+X, and $p = 0.290$ for AC+X.

I administered 24 sessions, 12 in each treatment, in February 2017. In total, the data include 404 subjects in 101 groups, 48 in one-shot SP-RSD and 53 in one-shot OSP-RSD.[44]

Under one-shot SP-RSD, 40 percent of games do not end in the dominant strategy outcome, compared to 8 percent under one-shot OSP-RSD. This difference is statistically significant ($p = 0.0001$, Fisher's exact test). For comparison, the corresponding error rates in the first round of the ten-round treatments are 44 percent for SP-RSD and 17 percent for OSP-RSD. Error rates in the one-shot treatments are not significantly different from first-round error rates in the ten-round treatments.[45] In one-shot SP-RSD, subjects who submit incorrect lists lose \$2.30 on average, compared to the amount they would have earned by playing the dominant strategy.[46]

E. *Summary of Experiment*

In summary, subjects play the dominant strategy at higher rates in OSP mechanisms, as compared to SP mechanisms that should (according to standard theory) implement the same choice rule. This difference is significant and substantial across all three pairs of mechanisms, and persists after five rounds of repeated play with feedback. Of course, the set of all strategy-proof mechanisms cannot be tested exhaustively in a single experiment. This is only preliminary evidence that obvious dominance correctly classifies strategy-proof mechanisms according to their cognitive complexity.

There are not (yet) well-formed alternative theories that explain the data presented here. The following theories predict dominant-strategy play in any strategy-proof mechanism: level-k equilibrium,[47] cognitive hierarchy equilibrium,[48] cursed equilibrium,[49] and analogy-based expectation equilibrium.[50] Standard specifications of quantal response equilibrium[51] predict substantial under-bidding in ascending auctions, and simulations suggest that it predicts *larger* errors in ascending auctions than in second-price auctions, which is the opposite of the effect in the data (the online Appendix provides details).

V. Discussion

Sometimes, it is important to design simple mechanisms. If simple auctions attract more participants, then they can raise more revenue than complex optimal auctions (Bulow and Klemperer 1996). Moreover, human beings tend to play their equilibrium strategies in simple mechanisms, but make large mistakes in complex

[44] The imbalance between treatments is due to subjects who registered to participate, but were absent at the time their session started. Each session lasted about ten minutes, and subjects were paid \$5 for participating, in addition to their incentive payments.

[45] $p = 0.802$ for SP-RSD and $p = 0.176$ for OSP-RSD, Fisher's exact test.

[46] In one-shot SP-RSD, 33 percent of subjects submit incorrect rank-order lists. The most common error is still to report the order 2nd–1st–3rd–4th, which accounts for 30 percent of incorrect lists. Demographic analysis is in the online Appendix.

[47] Stahl and Wilson (1994, 1995); Nagel (1995).

[48] Camerer, Ho, and Chong (2004).

[49] Eyster and Rabin (2005); Esponda (2008).

[50] Jehiel (2005).

[51] McKelvey and Palfrey (1995, 1998).

mechanisms.[52] This affects the performance of complex optimal mechanisms, often in ways that are difficult to foresee. Simplicity can be a real constraint—every bit as real as the constraints imposed by technology or incentives.[53]

Consequently, we need a precise definition of simplicity. Of course, mechanisms can be simple or complex in many ways, and the definition I have provided captures only one of them. Any definition must trade off multiple worthy concerns. It must be empirically accurate, but analytically tractable. It must be weak enough to allow positive results, but strong enough to discipline our choice of mechanisms.

Here are some dimensions of simplicity that OSP does not capture, which might be fruitful for investigation. Firstly, since OSP invokes only standard game-theoretic primitives, it does not capture framing effects. Any multiple-choice mathematics quiz has an obviously dominant strategy, since there is a deterministic correct answer for every question. We can embed puzzles of arbitrary complexity into the labels of moves, which OSP entirely neglects. Secondly, OSP abstracts from the complexity of searching a deep game tree for the optimal strategy. The agent compares his obviously dominant strategy to deviations at each information set, but how does he find that strategy in the first place? In a well-engineered mechanism, the answer may be that the designer draws the agent's attention to it.

A formal definition lets us quantify the cost of simplicity. For instance, it is complex to sell multiple objects in a combinatorial auction, because the welfare-optimal mechanism must elicit preferences over the power set of objects. In such environments, it can be that no OSP mechanism yields first-best welfare.[54] However, when objects are substitutes and types are independently distributed, there always exists an OSP mechanism that delivers at least half of first-best expected welfare (Feldman, Gravin, and Lucier 2014; Dütting et al. 2016).[55] In these environments, the "cost of simplicity" is bounded by one-half.

When designing mechanisms, the appropriate constraints depend on the intended application. (The requirement that a ship withstand 30 atmospheres of pressure is superfluous for many vessels, but vital for submarines.) If the agents are expert strategists, and the mechanism will be thoroughly audited by a trusted third-party, then OSP is too demanding. Strategically sophisticated agents and full commitment power are the ideal conditions for mechanism design. These conditions do not always obtain, so it is valuable to design mechanisms that are robust to weaker assumptions.

Appendix

An *extensive game form with consequences in* X is a tuple $\langle H, \prec, A, \mathcal{A}, P, \delta_c, (\mathcal{I}_i)_{i \in N}, g \rangle$, such that

(i) H is a set of histories, along with a binary relation $\prec$ on H that represents precedence.

[52] Such behavior has been found in the laboratory (Kagel, Harstad, and Levin 1987; Kagel and Levin 1993; Chen and Sönmez 2006), and in high-stakes decisions in the field (Hassidim et al. 2016; Rees-Jones forthcoming).

[53] This paraphrases an observation that Roth (2007) makes about repugnance.

[54] This holds even when objects have additively separable values (Bade and Gonczarowski 2016).

[55] This statement is entailed by Lemma 3.4 of Feldman, Gravin, and Lucier (2014). It even holds for the more general class of fractionally subadditive preferences.

(a) $\prec$ is a partial order, and $(H, \prec)$ forms an arborescence.[56]
(b) $h_\emptyset$ denotes $h \in H | \neg\exists h' : h' \prec h$.
(c) $Z \equiv \{h \in H | \neg\exists h' : h \prec h'\}$.
(d) $\sigma(h)$ denotes the set of immediate successors of h.

(ii) A is a set of actions.

(iii) $\mathcal{A} : H\backslash h_\emptyset \to A$ labels each non-initial history with the last action taken to reach it.

(a) For all h, $\mathcal{A}$ is one-to-one on $\sigma(h)$.
(b) $A(h)$ denotes the actions available at h.

$$A(h) \equiv \bigcup_{h' \in \sigma(h)} \mathcal{A}(h'). \tag{A1}$$

(iv) P is a player function. $P : H\backslash Z \to N \cup c$, where c is chance.

(v) δ_c is the chance function. It specifies a probability measure over chance moves. d_c denotes some realization of chance moves: for any h where $P(h) = c, d_c(h) \in A(h)$.[57]

(vi) $\mathcal{I}_i$ is a partition of $\{h | P(h) = i\}$ such that

(a) $A(h) = A(h')$ whenever h and h' are in the same cell of the partition.
(b) For any $I_i \in \mathcal{I}_i$, we denote $P(I_i) \equiv P(h)$ for any $h \in I_i$. $A(I_i) \equiv A(h)$ for any $h \in I_i$.
(c) Each action is available at only one information set: if $a \in A(I_i)$, $a' \in A(I'_j), I_i \neq I'_j$ then $a \neq a'$.

(vii) g is an outcome function. It associates each terminal history with an outcome. $g : Z \to X$.

REFERENCES

Ashlagi, Itai, and Yannai A. Gonczarowski. 2015. "No Stable Matching Mechanism Is Obviously Strategy-Proof." *arXiv* 1511.00452.

Ausubel, Lawrence M. 2004. "An Efficient Ascending-Bid Auction for Multiple Objects." *American Economic Review* 94 (5): 1452–75.

Bade, Sophie, and Yannai A. Gonczarowski. 2016. "Gibbard-Satterthwaite Success Stories and Obvious Strategyproofness." https://arxiv.org/abs/1610.04873 (accessed September 13, 2017).

Barberà, Salvador, Dolors Berga, and Bernardo Moreno. 2016. "Group Strategy-Proofness in Private Good Economies." *American Economic Review* 106 (4): 1073–99.

Bartal, Yair, Rica Gonen, and Noam Nisan. 2003. "Incentive Compatible Multi Unit Combinatorial Auctions." *TARK '03*, Proceedings of the 9th Conference on Theoretical Aspects of Rationality and Knowledge, 72–87.

[56] That is, a directed rooted tree such that every edge points away from the root.

[57] We could make the addition assumption that δ_c has full support on the available moves $A(h)$ when it is called to play. This ensures a pleasing invariance property: it rules out games with zero-probability chance moves that do not affect play, but do affect whether a strategy is obviously dominant. However, a full support assumption is not necessary for any of the results that follow, so we do not make it here.

Bird, Charles G. 1984. "Group Incentive Compatibility in a Market with Indivisible Goods." *Economics Letters* 14 (4): 309–13.

Bulow, Jeremy, and Paul Klemperer. 1996. "Auctions versus Negotiations." *American Economic Review* 86 (1): 180–94.

Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong. 2004. "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics* 119 (3): 861–98.

Charness, Gary, and Dan Levin. 2009. "The Origin of the Winner's Curse: A Laboratory Study." *American Economic Journal: Microeconomics* 1 (1): 207–36.

Chen, Yan, and Tayfun Sönmez. 2006. "School Choice: An Experimental Study." *Journal of Economic Theory* 127 (1): 202–31.

Cramton, Peter. 1998. "Ascending Auctions." *European Economic Review* 42 (3–5): 745–56.

Dütting, Paul, Michal Feldman, Thomas Kesselheim, and Brendan Lucier. 2016. "Posted Prices, Smoothness, and Combinatorial Prophet Inequalities." *arXiv* 1612.03161.

Edelman, Benjamin, Michael Ostrovsky, and Michael Schwarz. 2007. "Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords." *American Economic Review* 97 (1): 242–59.

Elmes, Susan, and Philip J. Reny. 1994. "On the Strategic Equivalence of Extensive Form Games." *Journal of Economic Theory* 62 (1): 1–23.

Esponda, Ignacio. 2008. "Behavioral Equilibrium in Economies with Adverse Selection." *American Economic Review* 98 (4): 1269–91.

Esponda, Ignacio, and Emanuel Vespa. 2014. "Hypothetical Thinking and Information Extraction in the Laboratory." *American Economic Journal: Microeconomics* 6 (4): 180–202.

Esponda, Ignacio, and Emanuel Vespa. 2016. "Contingent Preferences and the Sure-Thing Principle: Revisiting Classic Anomalies in the Laboratory." https://pdfs.semanticscholar.org/b533/1bd9de240fabfd5278cbdfb7d0b4bcb77398.pdf (accessed September 13, 2017).

Eyster, Erik, and Matthew Rabin. 2005. "Cursed Equilibrium." *Econometrica* 73 (5): 1623–72.

Feldman, Michal, Nick Gravin, and Brendan Lucier. 2014. "Combinatorial Auctions via Posted Prices." *arXiv* 1411.4916 [cs.GT].

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10 (2): 171–78.

Friedman, Eric J. 2002. "Strategic Properties of Heterogeneous Serial Cost Sharing." *Mathematical Social Sciences* 44 (2): 145–54.

Friedman, Eric, and Scott Shenker. 1996. "Synchronous and Asynchronous Learning by Responsive Learning Automata." https://pdfs.semanticscholar.org/9b36/9bdc873b5e3b40a2c22017d4efa90ba3c8df.pdf (accessed September 13, 2016).

Gkatzelis, Vasilis, Evangelos Markakis, and Tim Roughgarden. 2017. "Deferred-Acceptance Auctions for Multiple Levels of Service." http://theory.stanford.edu/~tim/papers/reverse2.pdf (accessed September 13, 2017).

Glazer, Jacob, and Ariel Rubinstein. 1996. "An Extensive Game as a Guide for Solving a Normal Game." *Journal of Economic Theory* 70 (1): 32–42.

Green, Jerry, and Jean-Jacques Laffont. 1977. "Characterization of Satisfactory Mechanisms for the Revelation of Preferences for Public Goods." *Econometrica* 45 (2): 427–38.

Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." *Journal of the Economic Science Association* 1 (1): 114–25.

Harstad, Ronald M. 2000. "Dominant Strategy Adoption and Bidders' Experience with Pricing Rules." *Experimental Economics* 3 (3): 261–80.

Hassidim, Avinat, Deborah Marciano-Romm, Assaf Romm, and Ran I. Shorrer. 2016. "'Strategic' Behavior in a Strategy-Proof Environment." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2784659 (accessed September 13, 2017).

Holmström, Bengt. 1979. "Groves' Scheme on Restricted Domains." *Econometrica* 47 (5): 1137–44.

Jehiel, Philippe. 2005. "Analogy-Based Expectation Equilibrium." *Journal of Economic Theory* 123 (2): 81–104.

Kagel, John H., Ronald M. Harstad, and Dan Levin. 1987. "Information Impact and Allocation Rules in Auctions with Affiliated Private Values: A Laboratory Study." *Econometrica* 55 (6): 1275–304.

Kagel, John H., and Dan Levin. 1993. "Independent Private Value Auctions: Bidder Behaviour in First-, Second- and Third-Price Auctions with Varying Numbers of Bidders." *Economic Journal* 103 (419): 868–79.

Kincaid, J. Peter, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. "Derivation of New Readibility Formulas (Automated Readibility Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel." DTIC Document. http://www.dtic.mil/docs/citations/ADA006655 (accessed September 13, 2017).

Li, Shengwu. 2017. "Obvious Ex Post Equilibrium." *American Economic Review* 107 (5): 230–34.
Li, Shengwu. 2017. "Obviously Strategy-Proof Mechanisms: Dataset." *American Economic Review*. https://doi.org/10.1257/aer.20160425.
Loertscher, Simon, and Leslie M. Marx. 2017. "Optimal Clock Auctions." https://faculty.fuqua.duke.edu/~marx/bio/papers/PriorFree.pdf (accessed September 13, 2017).
McCabe, Kevin A., Stephen J. Rassenti, and Vernon L. Smith. 1990. "Auction Institutional Design: Theory and Behavior of Simultaneous Multiple-Unit Generalizations of the Dutch and English Auctions." *American Economic Review* 80 (5): 1276–83.
McKelvey, Richard D., and Thomas R. Palfrey. 1995. "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior* 10 (1): 6–38.
McKelvey, Richard D., and Thomas R. Palfrey. 1998. "Quantal Response Equilibria for Extensive Form Games." *Experimental Economics* 1 (1): 9–41.
Milgrom, Paul, and Ilya Segal. 2015. "Deferred-Acceptance Auctions and Radio Spectrum Allocation." https://faculty.fuqua.duke.edu/~marx/bio/papers/PriorFree.pdf (September 13, 2017).
Mirrlees, James A. 1971. "An Exploration in the Theory of Optimum Income Taxation." *Review of Economic Studies* 38 (114): 175–208.
Myerson, Roger B. 1981. "Optimal Auction Design." *Mathematics of Operations Research* 6 (1): 58–73.
Nagel, Rosemarie. 1995. "Unraveling in Guessing Games: An Experimental Study." *American Economic Review* 85 (5): 1313–26.
Ngangoue, Kathleen, and Georg Weizsäcker. 2015. "Learning from Unrealized versus Realized Prices." German Institute for Economic Research Berlin Discussion Paper 1487.
Pathak, Parag A., and Tayfun Sönmez. 2008. "Leveling the Playing Field. Sincere and Sophisticated Players in the Boston Mechanism." *American Economic Review* 98 (4): 1636–52.
Pycia, Marek, and Peter Troyan. 2016. "Obvious Dominance and Random Priority." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2853563 (accessed September 13, 2017).
Rees-Jones, Alex. Forthcoming. "Suboptimal Behavior in Strategy-Proof Mechanisms: Evidence from the Residency Match." *Games and Economic Behavior.*
Roth, Alvin E. 1982. "Incentive Compatibility in a Market with Indivisible Goods." *Economics Letters* 9 (2): 127–32.
Roth, Alvin E. 2007. "Repugnance as a Constraint on Markets." *Journal of Economic Perspectives* 21 (3): 37–58.
Rothkopf, Michael H., and Ronald M. Harstad. 1995. "Two Models of Bid-Taker Cheating in Vickrey Auctions." *Journal of Business* 68 (2): 257–67.
Rothkopf, Michael H., Thomas J. Teisberg, and Edward P. Kahn. 1990. "Why Are Vickrey Auctions Rare?" *Journal of Political Economy* 98 (1): 94–109.
Saks, Michael, and Lan Yu. 2005. "Weak Monotonicity Suffices for Truthfulness on Convex Domains." Paper presented at Proceedings of the 6th Association for Computing Machinery Conference on Electronic Commerce, Vancouver, BC.
Savage, Leonard J. 1954. *Foundations of Statistics.* New York: John Wiley and Sons.
Shapley, Lloyd, and Herbert Scarf. 1974. "On Cores and Indivisibility." *Journal of Mathematical Economics* 1 (1): 23–37.
Shimoji, Makoto, and Joel Watson. 1998. "Conditional Dominance, Rationalizability, and Game Forms." *Journal of Economic Theory* 83 (2): 161–95.
Spence, A. Michael. 1974. "Competitive and Optimal Responses to Signals: An Analysis of Efficiency and Distribution." *Journal of Economic Theory* 7 (3): 296–332.
Stahl, Dale O., II, and Paul W. Wilson. 1994. "Experimental Evidence on Players' Models of Other Players." *Journal of Economic Behavior and Organization* 25 (3): 309–27.
Stahl, Dale O., and Paul W. Wilson. 1995. "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Economic Behavior* 10 (1): 218–54.
Thompson, F. B. 1952. "Equivalence of Games in Extensive Form." RAND Corporation Working Paper RM-759.
Troyan, Peter. 2016. "Obviously Strategyproof Implementation of Allocation Mechanisms." http://people.virginia.edu/~pgt8y/Troyan-TTC-OSP.pdf (accessed September 13, 2017).
Vickrey, W. 1961. "Counterspeculation, Auctions, and Competitive Sealed Tenders." *Journal of Finance* 16: 8–37.
Zhang, Luyao, and Dan Levin. 2017. "Bounded Rationality and Robust Mechanism Design: An Axiomatic Approach." *American Economic Review* 107 (5): 235–39.

This article has been cited by:

1. Alvin E. Roth. 2018. Marketplaces, Markets, and Market Design. *American Economic Review* **108**:7, 1609-1658. [Abstract] [View PDF article] [PDF with links]