Empirical strategy-proofness[*]

Rodrigo A. Velez[†] and Alexander L. Brown[‡]

Department of Economics, Texas A&M University, College Station, TX 77843

June 14, 2022

Abstract

Empirical tests of strategy-proof mechanisms often demonstrate that agents persistently report non-truthful messages. One possibility is that this behavior is consistent with equilibrium play, albeit an equilibrium not intended by the mechanism's designer. While these undesirable equilibria may exist under many mechanisms, we determine that the actual, empirical observation of such equilibria is only likely when the underlying social choice function violates a non-bossiness condition and information is not interior. Our analysis introduces and relies upon an empirically-based approach to the refinement of Nash equilibrium. A survey of experimental and empirical results on games of this type supports our findings.

JEL classification: C72, D47, D91.

Keywords: behavioral mechanism design; empirical equilibrium; robust mechanism design; strategy-proofness.

1 Introduction

Strategy-proofness (Gibbard, 1973; Satterthwaite, 1975) requires that truthful reports be dominant strategies in the simultaneous direct-revelation game associated with a social choice function (scf). Despite its theoretical appeal (see Barbera, 2010, for a survey),

[*] Thanks to James Andreoni, Antonio Cabrales, Marco Castillo, Yeon-Koo Che, Cary Deck, Cary Fryman, Huiyi Guo, Yuval Heller, Mark Isaac, Salvatore Nunnari, Marek Pycia, Utku Unver and seminar participants in Boston College, NC State U., Ohio State U., UCSD, UT Dallas, U. Zurich, Georgetown/Maryland, UN San Luis, SAET19, 7th TETC, and North American Meetings ESA 2019, COMID20, GAMES2020 for useful comments. Special thanks to the authors of Attiyeh et al. (2000); Cason et al. (2006); Chen and Sönmez (2006); Healy (2006); Andreoni et al. (2007); and Li (2017) whose data is either publicly available or has been made available for our analysis. All errors are our own.

[†] rvelezca@tamu.edu; https://sites.google.com/site/rodrigoavelezswebpage/home

[‡] alexbrown@tamu.edu; http://people.tamu.edu/∼alexbrown

experimental and empirical evidence suggests that agents may persistently exhibit weakly-dominated behavior when a strategy-proof scf is operated. This includes experiments with a wide variety of mechanisms (Coppinger et al., 1980; Kagel et al., 1987; Kagel and Levin, 1993; Harstad, 2000; Attiyeh et al., 2000; Chen and Sönmez, 2006; Cason et al., 2006; Healy, 2006; Andreoni et al., 2007; Li, 2017; Masuda et al., 2022), survey evidence from matching platforms (Rees-Jones, 2017; Hassidim et al., 2020), and empirical evidence from school-choice mechanisms (Artemov et al., 2021; Chen and Pereyra, 2019; Shorrer and Sóvágó, 2022). More strikingly, studies have documented persistent weakly-dominated behavior that supports the Nash equilibrium hypothesis and produces outcomes that differ from those selected by the scf (Sec. 5.2). This paper identifies the conditions under which this type of behavior may plausibly occur.

To understand why some, but not all, suboptimal equilibria of strategy-proof games are empirically relevant, it is natural to analyze them based on an equilibrium refinement. The popular tremble-based refinements are not suitable for this purpose. From Selten (1975), Myerson (1978) and Kohlberg and Mertens (1986) to their most recent forms in Milgrom and Mollner (2021, 2018) and Fudenberg and He (2021), these refinements discard all weakly-dominated behavior as implausible (see van Damme (1991) for an early survey).[1]

We propose an alternative path to refine Nash equilibrium. We partition strategy space into strategy profiles that are either plausible or implausible. Our partition is validated by over thirty years of empirical research. Equilibria that can be approximated by profiles in plausible strategy space are plausible equilibria; the other are not. Thus, our selection is *empirical*. Data supporting a plausible equilibrium is compatible with plausible behavior being played. Data supporting an implausible equilibrium rejects the hypothesis that plausible behavior is being played.

We use a non-parametric theory, *weak payoff monotonicity*, to determine plausibility of behavior. This property of the full profile of empirical distributions of play in a game requires that for each agent, differences in frequency of play reveal differences in expected utility. That is, between two alternative actions for an agent, say a and b, if the agent plays a with higher frequency than b, it is because given what the other agents are doing, a has higher expected utility than b (see Sec. 2 for an intuitive example). Weak payoff monotonicity is satisfied by monotone noisy best-response models.[2] In game experiments repeated multiple times, where behavior has a chance to converge, monotone noisy best-

[1] Economic theorists have seldom addressed the plausibility of wekly-dominated behavior. There are three notable exceptions. Nachbar (1990) and Dekel and Scotchmer (1992) observed that weakly-dominated behavior can result from the evolution of strategies that are updated by means of simple intuitive rules. Perhaps the study that is most skeptical of discarding all weakly-dominated behavior is Samuelson (1992), who shows that it has no solid epistemic foundation in all games.

[2] These include the exchangeable randomly perturbed payoff models (Harsanyi, 1973; van Damme, 1991), the control cost model (van Damme, 1991), the monotone structural Quantal Response Equilibrium (QRE)

response models typically do a good job at predicting final period averages as well as comparative statics across treatments (Goeree et al., 2018).[3]

Empirical equilibrium is the refinement that selects each Nash equilibrium for which there is a sequence of weakly payoff monotone behavior that converges to it. Empirical equilibria exist for each finite game and may admit weakly-dominated behavior.[4] Data that supports an equilibrium that is not empirical, necessarily refutes *all* monotone noisy best-response models. Thus, if the researcher endorses the hypothesis that behavior will be rationalized by a monotone noisy best-response model, they can confidently discard all equilibria that are not empirical as implausible.

We can considerably advance our understanding of the direct-revelation game of a strategy-proof scf by calculating its empirical equilibria. If, on the one hand, we find that for a certain game each empirical equilibrium is truthful equivalent, then, we may suspect that should behavior approach equilibrium, we should not be concerned with non-truthful strategies. On the other hand, if we find that some empirical equilibria are not truthful-equivalent, this alerts us to the possibility that we may observe persistent behavior that generates undesirable outcomes and approximates mutual best responses.

We present two main results. First, non-bossiness—i.e., the requirement on an scf that no agent be able to change the outcome without changing her own welfare—is *necessary and sufficient* to guarantee that for each common prior type space, each empirical equilibrium of the direct-revelation game of a strategy-proof scf in a private values environment, produces, with certainty, the truthful outcome (Theorem 1). Second, the requirement that a strategy-proof scf have no bossy dominant strategy characterizes this form of robust implementation for type spaces with full support (Theorem 2).[5]

Our results provide sharp predictions on which scfs should and should not reliably produce their theoretically-intended outcomes when behavior supports the Nash equilibrium hypothesis. Our theory is silent about the propensity of a mechanism to induce Nash behavior or dominant strategy play. It is often accepted that agents gain greater strategic sophistication with experience and repetition of game (e.g., Davis and Holt, 1993; Plott and

model (McKelvey and Palfrey, 1995), and the regular QRE models (McKelvey and Palfrey, 1996; Goeree et al., 2005).

[3] Goeree et al. (2005) argue in favor of a stronger form of weak payoff monotonicity that requires frequencies of play be ordinally equivalent to expected payoffs, to discipline unrestricted noisy best-response models that are not falsifiable (Haile et al., 2008). Our construction can be equivalently founded on this stronger non-parametric theory (Velez and Brown, 2020b).

[4] Indeed, the limits of logistic QRE (as the noisy best responses converge to best responses) are empirical equilibria (McKelvey and Palfrey, 1995). It is known that for each finite game these limits exist and that they may admit weakly-dominated behavior (McKelvey and Palfrey, 1995).

[5] A strategy is bossy if for some report of the other agents it changes the outcome compared with the truthful report and keeps the payoff of the agent constant.

Zeiler, 2005; Goeree et al., 2016), so our theory is more applicable for experienced rather than initial play, although we do not endorse or require use of any particular model of equilibrium dynamics.

The Top Trading Cycles mechanism (Shapley and Scarf, 1974), the median voting rules (Moulin, 1980), the Uniform rule (Benassy, 1982; Sprumont, 1983), and all securely implementable mechanisms (Saijo et al., 2007) are all non-bossy. Thus, if we operate these scfs we should not expect frequencies of play be close to an equilibrium that produces unintended outcomes. Another category of scfs are all bossy but have no bossy dominant strategy. For these scfs, as long as there is enough uncertainty we can still expect the same. They include the second-price auction, the Pivotal mechanism (see Green and Laffont, 1977), and Student Proposing Deferred Acceptance (Gale and Shapley, 1962; Abdulkadiroğlu and Sönmez, 2003). Outside of a few specific theoretical thought experiments (see e.g., Repullo, 1985), there are few scfs that fall outside of these first two categories. This is good news to mechanism designers: such scfs would always have open the possibility of undesirable equilibria where agents play weakly-dominated strategies.

The sharp predictions of our theorems are consistent with experimental and empirical evidence on strategy-proof mechanisms (Sec. 5). Indeed, they are in line with some of the most puzzling evidence on the second-price auction, a strategy-proof mechanism that violates non-bosiness but has no bossy dominant strategy. Consistent with our predictions, consequential deviations from truthful behavior producing undesirable outcomes are persistently observed when this mechanism is operated but only when agents' types are common information (Andreoni et al., 2007).

Our results have important consequences for robust mechanism design. Robust full implementation requires mechanisms produce the right outcomes in each and every predicted behavior for a rich family of information structures (Bergemann and Morris, 2011). We characterize robust full implementation based on the empirical equilibrium prediction (Sec. 4.3). Full implementation theory has been limited to two extremes. In one extreme, the researcher uses Nash equilibrium as prediction. This leads mainly to impossibility results that depend crucially on *all* Nash equilibria being plausible (Saijo et al., 2007; Adachi, 2014; Bochet and Sakai, 2010; Fujinaka and Wakayama, 2011). Thus it is not clear that these are hard constraints of design. On the other extreme, the researcher uses undominated equilibrium as prediction. This imposes no restrictions (Jackson, 1992). Some weakly dominated equilibria are empirically relevant, however. Thus, the current state of the art in full implementation theory is either unnecessarily pessimistic or unrealistically optimistic. Our proposal, designing mechanism based on empirical equilibrium, bridges these two approaches with a tractable prediction that is informed by the accumulated empirical evidence.

All in all, our study finds meaningful differences among strategy-proof scfs that explain to a great extent why unwanted Nash equilibria are empirically relevant only for some of these scfs in only some information structures. It is not the first to note differences between strategy-proof scfs. Besides the work on robust implementation, a growing literature searches for mechanisms with *simple* dominant strategies (Li, 2017; Pycia and Troyan, 2022; Mackenzie, 2020).[6] With only few exceptions (e.g., Arribilaga et al., 2020), these simplicity requirements are only satisfied by priority-like scfs in problems of interest that admit more symmetric strategy-proof scfs that are non-bossy (c.f. Ashlagi and Gonczarowski, 2018; Bade and Gonczarowski, 2017; Troyan, 2019). Even though we share the common objective of identifying subclasses of strategy-proof mechanisms that perform better in practice, our emphasis is on behavior among experienced players (Nash behavior), and not on identifying conditions that foster dominant strategy play.

Our work belongs to the growing literature on behavioral mechanism design, which aims to inform the design of mechanisms with regularities observed in laboratory experiments and empirical data. These papers can be classified in two different approaches. First, Cabrales and Ponti (2000), Healy (2006), and Tumennasan (2013) study the performance of mechanisms for solutions concepts defined by a convergence process.[7] They identify properties of mechanisms that guarantee their convergence to desired allocations under certain dynamics. These conditions turn out to be strong and essentially require implementation in strict equilibria. The second approach in this literature is to analyze the design of mechanisms accounting for behavior that is not utility maximizing for specific alternative behavior models (c.f., Eliaz, 2002; de Clippel, 2014; de Clippel et al., 2018; Kneeland, 2022). Our work bridges these two approaches. It informs us about the performance of mechanisms when behavior satisfies a weak form of rationality and also is approximately in equilibrium. Note that even though we assume that relevant behavior is in the proximity of a Nash equilibrium, we are not assuming that agents are approximate utility maximizers. Agents may seem as utility maximizers because indeed they are, or because the behavior of the other agents makes them look as if they were utility maximizers. Thus, when we design mechanisms based on empirical equilibrium, we are implicitly evaluating the system based on a theory of boundedly rational behavior.[8]

[6] Bo and Hakimov (2019, 2020) have also indentified mechanisms that may perform better than direct-revelation mechanisms of strategy proof scfs in experimental environments.

[7] Tumennasan (2013) defines implementation as the conjunction of two phenomena. First, all limits of logistic Quantal Response Equilibrium behavior are optimal, a requirement similar in nature to implementation in empirical equilibrium. Second, at least one of these sequences exhibits a strong form of convergence. This second requirement implies the existence of a strict equilibrium. Thus, the game forms associated with most strategy-proof social choice functions do not satisfy this requirement.

[8] In this sense we share part of the philosophy of analysis of strategic behavior with misspecified models (Esponda and Pouzo, 2016).

		Agent c	
		L	R
Agent r	T	1,1	0,0
	B	0,0	0,0

		Agent B	
		other	own
Agent A	other	$u_A(h_B), u_B(h_A)$	0,0
	own	0,0	0,0

Table 1: Identical normal form of games. Basic example game for demonstrating empirical equilibrium (left). The equivalent game in a house-trading game under TTC (right); $u_i(h_j) > 0$ is agent i's utility from trading with j. Though both agents have one weakly dominant strategy, there are two Nash equilibrium in pure strategies.

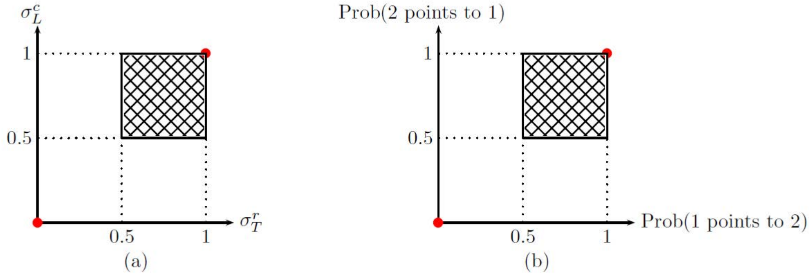

Figure 1: Weakly payoff monotone profiles (shaded area) and empirical equilibria in basic coordination game (a) and equivalent TTC game with complete information (b). Agents coordinate with probability one in the unique empirical equilibrium: the only Nash equilibrium that is in the closure of weakly payoff monotone behavior.

2 The intuition: empirical equilibrium, TTC, and second-price auction

A basic game illustrates empirical equilibrium (Table 1(left)). Consider a pair of agents $\{r, c\}$ who have action spaces $A_r = \{T, B\}$ and $A_c = \{L, R\}$. If they coordinate on (T, L) they get one dollar, otherwise they get zero. Let σ_T^r be the probability with which r chooses T, and so on. Behavior in this game is represented by the pair $(\sigma_T^r, \sigma_L^c) \in [0,1] \times [0,1]$ (Fig. 1(a)). There are two Nash equilibria in this game. They either coordinate, $(\sigma_T^r, \sigma_L^c) = (1,1)$; or they miscoordinate, $(\sigma_T^r, \sigma_L^c) = (0,0)$. Nash equilibrium $(0,0)$ is not empirical. The closure of the weakly payoff monotone distributions in this game is the north-east quadrant of the strategy space. Indeed, in a weakly payoff monotone distribution no agent plays the miscoordinating action with probability greater than one half. This action is weakly dominated by the coordinating action, so it will never have a higher payoff. Thus, if one were to have data supporting $(0,0)$, that data would reject each noisy best response model satisfying weak payoff monotonicity. Once can easily see that $(1,1)$ is an empirical equilibrium of this game.

Two mechanisms illustrate how empirical equilibrium informs us about the performance of dominant strategy mechanisms: The top trading cycles (TTC) mechanism for the reallocation of indivisible goods from individual endowments (Shapley and Scarf, 1974); and the second-price auction. For simplicity, we consider two-agent, stylized versions of these market design environments.

Consider two agents, A and B, with strict preferences over the houses they each own. TTC is the mechanism that operates as follows. Each agent points to a house. They swap houses if each agent points to the other agent's house; they remain in their houses otherwise. It is a dominant strategy for each agent to point to her preferred house. Thus, if one predicts that truthful dominant strategies will result when this mechanism is operated, one would obtain efficient trade. Other Nash equilibria exist. Consider the strategy profile where each agent unconditionally points to her own house. Regardless of information structure, these strategies are mutual best responses for expected-utility-maximizing agents. They do not produce the same outcomes as the truthful profile when trade is efficient. Simple analysis reveals an identical structure to our example game (see Table 1(right) and Fig. 1(b)).

The second-price auction is a mechanism for the allocation of a good by a seller among some buyers. We suppose that there are two buyers, A and B, who may have a type $\theta_i \in \{L, M, H\}$. The value that an agent assigns to the object depends on her type: $v_L = 0$, $v_M = 1/2$, and $v_H = 1$. Each agent has quasi-linear preferences, i.e., assigns zero utility to receiving no object, and $v_{\theta_i} - x_i$ to receiving the object and paying x_i for it. In the second-price auction each agent reports his or her value for the object.[9] Then an agent with the higher valuation receives the object and pays the seller the valuation of the other agent. Ties are decided uniformly at random. Truthfully revealing types is weakly dominant for each agent under this mechanism. In its truthful, dominant-strategy equilibrium, it obtains an efficient assignment of the object, i.e., an agent with higher value receives the object. Moreover, the revenue of the seller is the second highest valuation. Other Nash equilibria exist. Suppose that agent A has type M, agent B has type H, and both agents have complete information of their types. Table 2 presents the normal form of the complete information game that ensues. There are infinitely many Nash equilibria of this game. For instance, agent B reports her true type and agent A randomizes in some arbitrary way between L and M. In these equilibria, the seller generically obtains lower revenue than in the truthful equilibrium.

We intend to determine which, if any, of the sub-optimal equilibria of TTC, the second-price auction, or any other strategy-proof mechanism, should concern a social planner who operates one of these mechanisms. We proceed by calculating the empirical equilibria of the games induced by the operation of these mechanisms. The properties of the Nash equilibria of the TTC and the second-price auction differ significantly. No sub-optimal

[9] We use the term "second-price auction" to denote the scf defined by the dominant-strategy outcomes of the mechanism in which agents bid for the good and the winner pays the second highest bid.

		Agent B		
		H	M	L
Agent A	H	-1/4,0	0,0	1/2,0
	M	0,1/2	0,1/4	1/2,0
	L	0,1	0,1	1/4,1/2

Table 2: Normal form of second-price auction with complete information when $\theta_A = M$ and $\theta_B = H$.

Nash equilibrium of the TTC game is an empirical equilibrium. By contrast, for some information structures, the second-price auction has empirical equilibria whose outcomes differ from those of the truthful ones. Note that the sub-optimal equilibria of the TTC that we exhibit are prior free, i.e., they are strategy profiles that constitute equilibria independently of the information structure. However, as our analysis unveils, this property turns out to be unrelated to the empirical plausibility of equilibria.

First, consider the TTC game in a complete information environment in which both agents prefer to trade. This game is equivalent to the basic game we used to illustrate empirical equilibrium where the coordinating action is to point to the other agent (Fig.1(b)). If behavior can be fit satisfactorily by a model satisfying weak payoff monotonicity, the only equilibrium that has the chance to be approximated by empirical distributions is the efficient equilibrium.

This argument does not depend on the assumption of complete information. Consider the TTC game when information is summarized by a common prior.[10] Since revealing true preferences is a dominant strategy, agents, regardless of type, must reveal their truthful preferences with probability at least 1/2 in each weakly payoff monotone profile. Thus, in any limit of a sequence of weakly-payoff-monotone strategies, each agent reveals her true preference with probability at least 1/2. Consequently, in each empirical equilibrium there is a lower bound on the probability with which each agent is truthful. Given the realization of agents' types, each agent always believes the true payoff type of the other agent is possible. Then, in each empirical equilibrium of the TTC, whenever trade is efficient (for the true types of the agents), each agent will place positive probability on the other agent pointing to her. Consequently, in each empirical equilibrium of the TTC, given that an agent prefers to trade, this agent will point to the other agent with probability *one* whenever efficient trade is possible. Thus, each empirical equilibrium of the TTC obtains the truthful outcome with certainty.

For the second-price auction, consider the complete information structure whose associated normal form game is presented in Table 2. Let $\varepsilon > 0$ and $\sigma := (\sigma_A, \sigma_B)$ be the pair of probability distributions on each agent's action space defined as follows. Agent A places ε

[10] This can be relaxed to some extent. See Sec. 3.

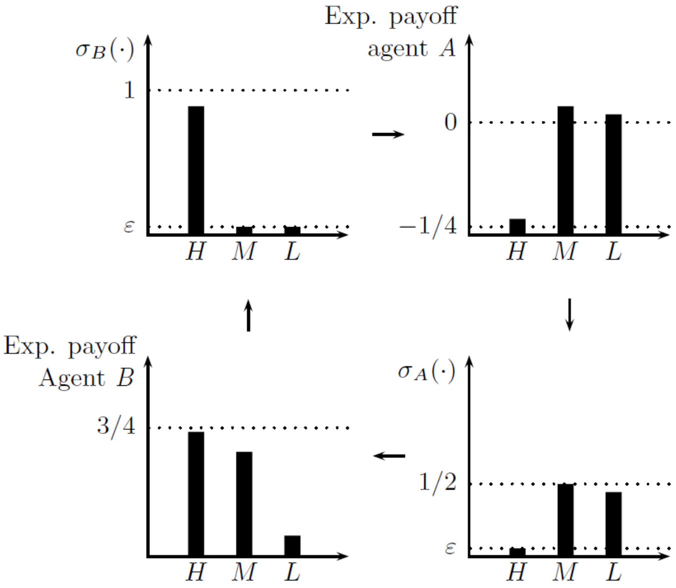

Figure 2: Weakly monotone distribution in the second price auction with complete information for agents M and H. As ε vanishes the distributions converge to a Nash equilibrium in which the lower value agent randomizes between L and M.

probability on H, $1/2$ on M and $1/2-\varepsilon$ on L. Agent B places $1-2\varepsilon$ probability on H, and ε on each of the other two actions. For small ε this profile is weakly payoff monotone. This can be easily seen in Fig. 2. Starting at the top left of the figure is σ_B. This distribution induces the expected payoffs for agent A shown at the top right of the figure. Since M is the unique weakly dominant strategy for agent A and σ_B is interior, the highest payoff for agent A is achieved by M. Between L and H, agent A obtains a higher payoff with L, because with H she ends up buying the object for a price above her value with positive probability. Thus, expected payoffs for agent A are ordered exactly as σ_A, which is shown at the bottom right of Fig. 2. If agent A plays σ_A, agent B's expected utility is that shown at the bottom left of Fig. 2. Agent B's utility is maximized at her unique weakly dominant strategy. Thus, differences in σ_B reveal differences in expected utility. More precisely, agent B playing H with higher probability than both M and L is consistent with H having higher expected payoff than both M and L. As ε vanishes, this profile converges to a Nash equilibrium in which agent A randomizes between M and L with equal probability.[11]

The concept of empirical equilibrium differentiates TTC and the second-price auction.

[11] Note that we can easily modify our construction to have interior profiles that are ordinally equivalent to expected payoffs. This is always possible. That is, each empirical equilibrium of a finite game is the limit of interior distributions that are, agent-wise, ordinally equivalent to expected payoffs (Velez and Brown, 2020b). Note also that we could easily modify our construction so agent A places probability $1/2+\alpha$ on M and $1/2-\alpha$ on L for any $0 \leq \alpha \leq 1/2$.

Suppose that agents' behavior is weakly payoff monotone. Then, if these mechanisms are operated, one will never observe that empirical distributions of play in TTC approximate an equilibrium producing a sub-optimal outcome. By contrast, this possibility is not ruled out for the second-price auction.

It turns out that these differences among these two mechanisms can be pinned down to a property that TTC satisfies and the second-price auction violates: non-bossiness, i.e., in the direct-revelation game of the mechanism, an agent cannot change the outcome by lying without changing her welfare (Theorem 1).

For strategy-proof mechanisms that do violate non-bossiness, it is useful to examine which information structures produce undesirable empirical equilibria. For strategy-proof mechanisms in which no agent can be bossy with a dominant strategy (e.g., the second-price auction) undesirable equilibria cannot be an empirical equilibrium under interior information structures (Theorem 2). Thus, our previous example was dependent on the type of information structure we used (i.e., complete information).

Together, Theorems 1 and 2 produce sharp predictions about the type of behavior that is plausible when a strategy-proof scf is operated in different information structures. Our review of the relevant experimental and empirical literature is largely consistent with these predictions (see Sec. 5).

3 Model

There is a group of agents $N := \{1, \ldots, n\}$ and an arbitrary set of alternatives X. Agents have private values, i.e., each $i \in N$ has a payoff type θ_i, determining an expected utility index $u_i(\cdot|\theta_i) : X \to \mathbb{R}$. The set of possible payoff types for agent i is Θ_i and the set of possible payoff type profiles is $\Theta := \prod_{i \in N} \Theta_i$. We assume that Θ is finite. For each $S \subseteq N$, Θ_S is the cartesian product of the type spaces of the agents in S. The generic element of Θ_S is θ_S. When $S = N \setminus \{i\}$ we simply write Θ_{-i} and θ_{-i}. We concatenate partial profiles, as in (θ_{-i}, μ_i). We use this notation consistently when operating with vectors (as in strategy profiles). We assume that information is summarized by a common prior $p \in \Delta(\Theta)$.[12] For each θ in the support of p and each $i \in N$, let $p(\cdot|\theta_i)$ be the distribution p conditional on agent i drawing type θ_i.[13]

[12] For a finite set F, $\Delta(F)$ denotes the simplex of probability measures on F.

[13] Our results can be extended for general type spaces à la Bergemann and Morris (2005) when one requires the type of robust implementation in our theorems only for the common support of the priors. We prefer to present our payoff-type model for two reasons. First, it is simpler. Second, since our theorems are robust implementation characterizations, they are not stronger results when stated for larger sets of priors. By stating our theorems in our domain, the reader is sure that we do not make use of the additional freedom that games with non-common priors allow.

A social choice function (scf) selects an alternative for each possible state. The generic scf is $g : \Theta \to X$. Three properties of scfs play an important role in our results. An scf g,

1. is *strategy-proof* if for each $\theta \in \Theta$, each $i \in N$, and each $\tau_i \in \Theta_i$, $u_i(g(\theta)|\theta_i) \geq u_i(g(\theta_{-i}, \tau_i)|\theta_i)$.

2. is *non-bossy* if for each $\theta \in \Theta$, each $i \in N$, and each $\tau_i \in \Theta_i$, $u_i(g(\theta)|\theta_i) = u_i(g(\theta_{-i}, \tau_i)|\theta_i)$ implies that $g(\theta) = g(\theta_{-i}, \tau_i)$.

3. has *no bossy dominant strategy* if for each $\theta \in \Theta$, each $i \in N$, and each $\tau_i \in \Theta_i$, if $u_i(g(\theta)|\theta_i) = u_i(g(\theta_{-i}, \tau_i)|\theta_i)$ and $g(\theta) \neq g(\theta_{-i}, \tau_i)$, then there is $\tau_{-i} \in \Theta_{-i}$ such that $u_i(g(\tau_{-i}, \theta_i)|\theta_i) > u_i(g(\tau)|\theta_i)$.

The first property is well-known. The second property, first introduced by Saijo et al. (2007), requires that no agent, when telling the truth (in the direct-revelation mechanism associated with the scf), be able to change the outcome by changing her report without changing her welfare. It is satisfied by TTC, the Median Voting rule, and the Uniform rule. It is violated by the Pivotal mechanism, the second-price auction, and SPDA.

Starting from Satterthwaite and Sonnenschein (1981), a variety of related axioms, usually referred to as non-bossiness, have played a role in mechanism design, implementation, and social choice theory. Thomson (2016) surveys the definition and the normative content of these different notions of non-bossiness.

The third property requires that any consequential deviation from a truthful report by an agent can have adverse consequences for her. Restricted to strategy-proof scfs, this property says that, in the direct-revelation game associated with the scf, for each agent, all dominant strategies are equivalent. This property is satisfied whenever each agent type has a unique weakly dominant action, as in the second-price auction. It is not necessary that weakly dominant actions be unique for this property to be satisfied. A student in a school choice environment with strict preferences and in which Student Proposing Differed Acceptance is operated, may have multiple dominant strategies (e.g., a student that each school ranks first). However, any misreport that is also a dominant strategy for this student, cannot change the outcome (see Online Appendix for a formal argument).

A finite mechanism is a pair (M, φ) where $M := (M_i)_{i \in N}$ are finite message spaces and $\varphi : M \to \Delta(X)$ is an outcome function. Given the common prior p, (M, φ) determines a standard Bayesian game $\Gamma := (M, \varphi, p)$. When the prior is degenerate, i.e., places probability one on a payoff type $\theta \in \Theta$, we refer to this as a game of complete information and denote it simply by (M, φ, θ). A strategy for agent i in Γ is a function that assigns to each $\theta_i \in \Theta_i$ that happens with positive probability under p, a function $\theta_i \mapsto \sigma_i(\cdot|\theta_i) \in \Delta(M_i)$.

We denote a profile of strategies by $\sigma := (\sigma_i)_{i\in N}$. For each $S \subseteq N$, and each $\theta_S \in \Theta_S$, $\sigma_S(\cdot|\theta_S)$ is the corresponding product measure $\prod_{i\in S} \sigma_i(\cdot|\theta_i)$. When $S = N$ we simply write $\sigma(\cdot|\theta)$. We denote the measure that places probability one on $m_i \in M_i$ by δ_{m_i}. With a complete information structure, we simplify notation and do not condition strategies on an agent's type, which is uniquely determined by the prior. Thus, in game (M, φ, θ) we write σ_i instead of $\sigma_i(\cdot|\theta_i)$.

Let $\theta_i \in \Theta_i$ be realized with positive probability under p. The expected utility of agent i with type θ_i in Γ from playing strategy μ_i when the other agents select actions as prescribed by σ_{-i} is

$$U_\varphi(\sigma_{-i}, \mu_i|p, \theta_i) := \sum u(\varphi(m)|\theta_i)p(\theta_{-i}|\theta_i)\sigma_{-i}(m_{-i}|\theta_{-i})\mu_i(m_i|\theta_i),$$

where the summation is over all $\theta_{-i} \in \theta_{-i}$ and $m \in M$. A profile of strategies σ is an (Bayesian Nash) *equilibrium* of Γ if for each $\theta \in \Theta$ in the support of p, each $i \in N$, and each $\mu_i \in \Delta(M_i)$, $U_\varphi(\sigma_{-i}, \mu_i|p, \theta_i) \leq U_\varphi(\sigma_{-i}, \sigma_i|p, \theta_i)$. The set of equilibria of Γ is $N(\Gamma)$. Given an scf g, we say that a strategy profile σ *implements* g if for each θ in the support of p and each profile of reports m in the support of $\sigma(\cdot|\theta)$, $\varphi(m) = g(\theta)$.

We say that $m_i \in M_i$ is a *weakly dominant action* for agent i with type $\theta_i \in \Theta_i$ in (M, φ) if for each $r_i \in M_i$, and each $m_{-i} \in M_{-i}$, $u_i(m|\theta_i) \geq u_i(m_{-i}, r_i|\theta_i)$. A *dominant-strategy mechanism* is that in which each agent type has at least one weakly dominant action.

Our basis for plausibility of behavior is the following weak form of utility maximization.

Definition 1. A profile of strategies for $\Gamma := (M, \varphi, p)$, $\sigma := (\sigma_i)_{i\in N}$, is *weakly payoff monotone for* Γ if for each $\theta \in \Theta$ in the support of p, each $i \in N$, and each pair $\{m_i, n_i\} \subseteq M_i$ such that $\sigma_i(m_i|\theta_i) > \sigma_i(n_i|\theta_i)$, $U_\varphi(\sigma_{-i}, \delta_{m_i}|p, \theta_i) > U_\varphi(\sigma_{-i}, \delta_{n_i}|p, \theta_i)$.

We then identify the Nash equilibria that can be approximated by plausible behavior.

Definition 2. An *empirical equilibrium* of $\Gamma := (M, \varphi, p)$ is an equilibrium of Γ that is the limit of a sequence of weakly payoff monotone strategies for Γ.

In any finite game, firm equilibria, approachable equilibria, and the limits of logistic Quantal Response Equilibrium (as the sophistication parameter converges to infinite sophistication) are empirical equilibria. Thus, at least one empirical equilibrium exists for every finite game (see van Damme, 1991; McKelvey and Palfrey, 1995).

4 Results

4.1 When are all empirical equilibria optimal?

Our first result is a characterization of the social choice functions whose direct-revelation game obtains, for all common-prior information structures and for each possible state, the scf's optimal outcome.

Theorem 1. Let g be an scf. The following statements are equivalent.

1. For each common prior p each empirical equilibrium of (Θ, g, p) implements g.
2. g is strategy-proof and non-bossy.

We divide the proof of our characterization results, Theorems 1 and 2, in five lemmas of independent interest that we discuss in Sec. 4.2.

Theorem 1 states that the robust performance of a direct-revelation mechanism of an scf, evaluated with the empirical equilibrium prediction, depends exclusively on two of its properties: strategy-proofness and non-bossiness. Our examples in Sec. 2 illustrate it. TTC is both strategy-proof and non-bossy. In the TTC game, regardless of the information structure, none of the sub-optimal equilibria survive the empirical equilibrium refinement. The second-price auction is strategy-proof, but violates non-bossiness. In this auction, there are information structures for which some sub-optimal Nash equilibria are empirical equilibria.

It is somehow unexpected that the differences in performance of the TTC and second-price auction can be traced down to simple properties of the scfs that can be articulated in any private-values model. Each of these mechanisms operates in structured environments. For instance, some of the strategic properties of TTC crucially depend on particular assumptions on preferences. This scf selects the unique core allocation (existence and uniqueness of core allocation is rare on allocation environments). These further features of the housing-market model are not essential for the empirical equilibrium robustness of TTC. Similarly, a violation of strategy-proofness or non-bossiness is sufficient to find an information structure for which the direct-revelation game of an scf g has an empirical equilibrium that does not implement g. The particular additional features of our example (e.g., a specific number of agents, quasi-linear preferences) are not essential.

Our second result is a characterization of the social choice functions whose direct-revelation game obtains, for all *interior* common-prior information structures and for each possible state, the scf's recommended outcome.

Theorem 2. Let g be an scf. The following statements are equivalent.

1. For each interior common prior p, each empirical equilibrium of (Θ, g, p) implements g.

2. g is strategy-proof and has no bossy dominant strategy.

Theorem 2 reveals a connection between the information structure in which an scf g is operated and its performance. If, conditional on her type no agent can rule out a profile of types of the other agents, a direct-revelation game of g can have empirical equilibria that do not implement g only if g has bossy dominant strategies. The second-price auction, for instance, has no bossy dominant strategy. Thus, it is not by chance that our example in Sec. 2 of an empirical equilibrium with low revenue for the seller is in an environment with complete information.

Theorems 1 and 2 together allow us to better understand the Nash equilibria of the direct-revelation games of a strategy-proof scf g. If at least one agent has a bossy weakly dominant action, there are Nash equilibria in weakly dominant actions that do not implement g. One of these equilibria is empirical. If the scf has no bossy dominant strategy but violates non-bossiness, there could be empirical equilibria that do not implement g depending on the information structure in which the scf is operated. On the one hand, if information is interior, all empirical equilibria implement g. On the other hand, one can always find a non-interior information structure for which there is an empirical equilibrium that do not implement g. Finally, if g is non-bossy, then all empirical equilibria of a direct-revelation game of g implement g.

Our analysis articulates a folk wisdom about the performance of strategy-proof scfs: (some) dominant-strategy mechanisms should perform better when agents face enough uncertainty.[14] By contrast, refinements that discard Nash equilibria that involve weakly-dominated strategies produce a single prediction as long as a strategy-proof scf has no bossy-dominant strategies: only truthful equilibrium is plausible independently of the information structure.

4.2 Five basic lemmas

We divide the proof of the theorems in the previous section into five lemmas. First, let us see why under the conditions in the theorems each empirical equilibrium of the direct-revelation game of a strategy-proof scf g implements g.

We start with a key lemma. It states that, in each empirical equilibrium of a direct-revelation game of a strategy-proof scf, truthful reports are played with at least the frequency of uniform random play. We present the proof of all lemmas in the Appendix.

[14] For instance, it is commonly accepted within the experimental economics literature that private rather than common information of values may be beneficial for market outcomes (e.g., Smith, 1994). The general justification is that when more information is available about others' valuations, individuals may strive to deviate from the single-shot Nash equilibrium in order to capture more economic rents.

Definition 3. Let g be an scf and p a common prior. A strategy profile in (Θ, g, p), σ, is *baseline-truthful* if it prescribes each agent be truthful with a frequency that is at least that of uniform random play. That is, for each $i \in N$ and each $\theta_i \in \Theta_i$ that happens with positive probability under p, $\sigma_i(\theta_i|\theta_i) \geq 1/|\Theta_i|$.

Lemma 1. Let g be strategy-proof, p a common prior, and σ an empirical equilibrium of (Θ, g, p). Then, σ is a baseline-truthful strategy profile in (Θ, g, p).

The intuition behind Lemma 1 is simple. Truthful reports are weakly dominant strategies for each agent in a strategy-proof game. Thus, in a weakly-payoff-monotone distribution, truthful reports should have the highest frequency, which is always at least as great as under uniform-random play. This property is inherited by any limit of these distributions. In particular, it is inherited by all empirical equilibria.

With our next two results we show that a baseline-truthful equilibrium of the direct-revelation game of a strategy-proof scf always implements g when the conditions in our theorems are satisfied. In the next lemma, we identify the defining characteristic of a baseline-truthful equilibrium that does not implement g.

Lemma 2. Let g be strategy-proof, p a common prior, and σ a baseline-truthful equilibrium of (Θ, g, p). Suppose that σ does not implement g. Then, there are $\theta \in \Theta$ in the support of p, τ in the support of $\sigma(\cdot|\theta)$, and $i \in N$ such that $u_i(g(\tau)|\theta_i) = u_i(g(\tau_{-i}, \theta_i)|\theta_i)$ and $g(\tau) \neq g(\tau_{-i}, \theta_i)$.

Let g be strategy-proof and consider an equilibrium of its direct-revelation game. Lemma 2 states that if the equilibrium does not implement g, at least one agent ends up being bossy with respect to the reports of the other agents. Thus, the conditions that guarantee there is no bossy best-response to a baseline-truthful strategy in g's direct-revelation game also guarantee these equilibria all implement g. The following lemma identifies two of these conditions.

Lemma 3. Let g be strategy-proof, p a common prior, and σ a baseline-truthful strategy profile in (Θ, g, p). Assume that at least one of the following conditions holds:

1. g is non-bossy.
2. g has no bossy dominant strategy and p has full support.

Let $\theta \in \Theta$ in the support of p, τ_{-i} in the support of $\sigma_{-i}(\cdot|\theta_{-i})$, $i \in N$, and $\tau_i \in \Theta_i$ such that $g(\tau) \neq g(\tau_{-i}, \theta_i)$. Then, τ_i is not a best response to σ_{-i} for i.

Lemmas 1–3 allow us to complete the proof that Statement 2 implies Statement 1 in our theorems (see Appendix).

We now prove that the conditions in Statement 2 in our theorems are necessary for the robustness property in the corresponding Statement 1. We start with the second theorem, whose robustness requirement is weaker and thus admits a larger class of scfs.

Lemma 4. Let g be an scf. Suppose that for each interior prior p each empirical equilibrium of (Θ, g, p) implements g. Then, g is strategy-proof and has no bossy dominant strategy.

The necessity of strategy-proofness in Lemma 4 is closely related to Bergemann and Morris (2005, Proposition 3). These authors show that for each interior common-prior information structure the direct-revelation game of an scf g has a pure strategy equilibrium that implements g only if g is strategy-proof. Since there are games in which no empirical equilibrium is in pure strategies, our result does not immediately follow as a consequence of the earlier result. The argument advanced by these authors can be easily modified to prove our result, however. We present it in the Appendix for completeness.

Finally, it is necessary that an scf be non-bossy for it to satisfy the robustness property in Statement 1 in Theorem 1. (Strategy-proofness is already implied by Lemma 4.)

Lemma 5. Let g be strategy-proof. Suppose that for each common prior p each empirical equilibrium of (Θ, g, p) implements g. Then, g is non-bossy.

Our proof of Lemma 5 is by contradiction. We show that it is impossible for an scf to satisfy the robustness property in the lemma and also violate non-bossiness. If the scf violates non-bossiness, then there is some state θ at which at least one agent is bossy. That is, if all other agents report truthfully, the agent can change the outcome by lying without changing her payoff. We prove that one can always construct an empirical equilibrium in which the agent is bossy with positive probability for the complete information game for state θ. This proves the lemma because the equilibrium we construct does not implement g. It is worth noting that we prove this equilibrium is empirical by constructing a sequence of weakly payoff monotone distributions based on noisy best-response distributions that are popular for the analysis of experimental data (Goeree et al., 2016). The equilibrium play of the bossy agent in our construction is approximated by Logistic Quantal responses. The equilibrium play of all other agents are approximated by uniform perturbations of equilibrium play. The detail in our proof consists on carefully selecting the equilibrium and the speed at which the noisy best-responses move towards their limit. This construction guarantees these sequences converge to equilibrium and maintain weak payoff monotonicity along the way.

The proof of Lemma 5 relies on uniform perturbations, perturbations where agents play actions that have different expected utility with the same frequency. This is compatible with weak-payoff monotonicity. Indeed, this property only requires that between two actions that are played with different probability, the one with higher frequency has higher expected utility. It is possible to modify the construction and approximate the equilibrium by means of interior distributions in which frequencies of play are ordinally equivalent to expected utility. This is a consequence of a more general result. Each empirical equilibrium is the limit of behavior generated by a sequence of monotone noisy best-response operators (regular quantal response functions as defined by Goeree et al., 2005) that are utility maximizers in the limit (see Velez and Brown, 2020b, Theorem 2).

Our lemmas have independent substantive content. Lemmas 1–3 inform the mechanism designer that it is enough for behavior to be baseline-truthful for the robust performance of an scf satisfying Statement 2 in either Theorems 1 or 2. They also inform the empirical economist that when the conditions in Theorems 1 and 2 are satisfied, one can confirm behavior is not approaching a sub-optimal Nash equilibrium by simply checking that behavior is baseline-truthful. Importantly, baseline-truthfulness is an easily testable property that is overwhelmingly supported by experimental data (Sec. 5).

Additionally, Lemma 3 gives us a powerful shortcut for the analysis of experimental data from strategy-proof games. Suppose that one finds subjects in experimental games are truthful at least as frequently as uniform-random play. Additionally, subjects also play best responses with high probability. Then, if any of the conditions in the lemma are satisfied, the mechanism is necessarily performing well even when behavior stays persistently away from the truthful equilibrium. The reason is that under these conditions, any "consequential" lie is not a best response. (A consequential lie is one that, conditional on what the other agents are reporting, ends up changing the outcome of the game.) Thus, if one observes most agents are playing best responses, but a considerable share of these are lying, their lies cannot be consequential.

If our theorems' requirements on an scf are violated, then one cannot rule out weakly payoff monotone behavior can be close to mutual best responses that produce undesirable outcomes. Our proofs of Lemmas 4 and 5 reveal that this risk does not disappear even if one requires frequencies of play be ordinally equivalent to expected utility, or be generated by some of the most popular models for the analysis of experimental data.

4.3 Secure and robust implementation

It is informative to compare our results with secure implementation and robust implementation of scfs (Saijo et al., 2007; Bergemann and Morris, 2011). A mechanism fully-implements an scf g in a given solution concept for a set of information structures, if each predicted behavior in each game induced by the mechanism for each admissible information structure implements g.

Secure implementation requires the existence of a dominant-strategy mechanism that fully-implements the scf in Nash equilibrium for all complete information priors. This requirement is equivalent to the existence of a mechanism that fully-implements the scf in Bayesian Nash equilibrium for all common-prior information structures (Adachi, 2014). This later form of implementation is also referred to as robust implementation. The following theorem states necessary and sufficient conditions for secure/robust implementation in our environment.

Theorem 3 (Secure/robust implementation)**.** Let g be an scf. The following are equivalent.

1. There is a finite mechanism (M, φ) such that for each common prior p, each equilibrium of (M, φ, p) implements g.

2. There is a dominant-strategy finite mechanism (M, φ) such that for each complete information prior θ, each equilibrium of (M, φ, θ) implements g.

3. For each common prior p, each equilibrium of (Θ, g, p) implements g.

4. (i) g is strategy-proof and non-bossy, and

   (ii) g satisfies the outcome rectangular property, i.e., for each pair of payoff types $\{\theta, \tau\} \subseteq \Theta$, if for each $i \in N$, $g(\theta_i, \tau_{-i}) = g(\tau)$, then $g(\theta) = g(\tau)$.

A parallel result to Theorem 3 is due to Saijo et al. (2007) ($2 \Leftrightarrow 4$ and 2 & $4 \Rightarrow 3 \Rightarrow 1$) and Adachi (2014) ($1 \Rightarrow 4$) in an environment in which they restrict to pure-strategy equilibria and they consider implementation for type spaces larger than our payoff-type space. Our statement includes mixed-strategy equilibria and does not make any requirement for type spaces in which payoff types can be "cloned." Thus, Saijo et al. (2007) and Adachi (2014)'s results do not trivially imply Theorem 3 by means of Bergemann and Morris (Sec. 6.3, 2011)'s purification argument. The proof of Theorem 3 can be completed by adapting the arguments in these papers, however. We include it in an online Appendix.

In the language of full implementation, Theorem 1 states that an scf is fully implementable in empirical equilibrium for all common-prior information structures by its direct-revelation mechanism if and only if it is strategy-proof and non-bossy. Theorem 2 states that an scf is fully implementable in empirical equilibrium for all interior common-prior information structures by its direct-revelation mechanism if and only if it is strategy-proof and has no bossy dominant strategy. Theorem 3 implies that an scf is fully implementable in Bayesian Nash equilibrium for all common-prior information structures by its direct-revelation mechanism if and only if it is strategy-proof, non-bossy, and satisfies the outcome-rectangular property.

Secure/robust implementation and our empirical equilibrium analysis pursue a similar objective. We all identify mechanisms that perform well independently of informational assumptions based on the Nash equilibrium prediction. Secure/robust implementation guarantees optimal performance for *each* equilibrium of the system. We only make this requirement for the empirical equilibria of the games.

Since each empirical equilibrium is an equilibrium, the properties that characterize the forms of implementation in our theorems are necessary for secure/robust implementation. They are not sufficient, however. Besides strategy-proofness and non-bossiness, secure/robust implementation also requires that the outcome rectangular property be satisfied. Remarkably, several prominent environments that admit non-dictatorial, strategy-proof and non-bossy scfs, only admit (serially) dictatorial securely implementable scfs (Saijo et al., 2007; Bochet and Sakai, 2010; Fujinaka and Wakayama, 2011). Table 3 presents some of the most prominent scfs that satisfy the conditions in our theorems, but are not securely/robustly implementable.[15] Thus, all these scfs are not securely/robustly implementable because of equilibria that are not empirical.

Experimental studies show us that behavior in strategy-proof games may approximate equilibria in weakly-dominated strategies that obtain outcomes not intended by the mechanism designer (see Sec. 5). Thus, similar to the secure/robust implementation literature, we find that it is indeed justified to evaluate these systems with a prediction beyond dominant strategy equilibrium. We do not think that each possible Nash equilibrium of a system poses a real risk to it, however. Instead of making a blanket requirement for all Nash equilibria, as secure/robust implementation does, we refine these equilibria by requiring proximity to specific behavior. Recall, this specific behavior is in the empirical content of some of the most prominent theories that have been successful in fitting data and replicating comparative statics in experiments. Since the experimental data is the one alerting us about the shortcomings of the dominant strategy equilibrium prediction, we let this data guide us in the selection of the equilibria that are relevant in these games.

Bochet and Tumennassan (2020) provide an alternative characterization of secure/robust implementation. Their basic definitions are *resilience* and *group resilience*, two properties of an scf. The former property requires that for any profile of reports that leads to an outcome not intended by the scf, at least one agent benefits by deviating to the truthful report. The later property extends the requirement to groups of agents. On the one hand, an scf satisfies resilience if and only if it is securely implementable. Thus, interestingly, reversion to truthful reports without cooperation does not generate the same refinement

[15] Note that the secure/robust implementation theorem guarantees a revelation principle holds for this type of implementation. Thus, if an scf violates the outcome-rectangular property, it cannot be securely/robustly implemented by a general mechanism.

scf (preference domain)	Strategy-proofness	Essentially unique dominant strategies	Non-bossiness	outcome rectangular property
TTC (strict)	+	+	+	−
Uniform rule (single peaked)	+	+	+	−
Median voting (single peaked)	+	+	+	−
Second price auction (quasi-linear)	+	+	−	−
Pivotal (quasi-linear)	+	+	−	−
SPDA (strict)	+	+	−	−

Table 3: Strategy-proof scfs and the outcome rectangular property; + indicates that the property labeling the column is satisfied by the scf, and − the opposite. These statements refer to the usual preference spaces in which these scfs are defined.

as approximation by weakly-payoff-monotone behavior. On the other hand, an scf satisfies group resilience if and only if it is strategy-proof and non-bossy. Thus, in environments in which communication and cooperation is feasible, the undesirable equilibria of a strategy-proof and non-bossy scf are refined away by coalitional incentives. At a conceptual level, our approach differs in that our basis to refine equilibria, weak-payoff monotonicity, is an equilibrium property that requires no cooperation. At a technical level our results differ in that these authors restrict their model to pure strategy equilibria and complete information. Thus, our work is the first to establish a meaningful connection between the information structure in which a strategy-proof scf is operated and its performance.

5 Empirical evidence on dominant strategy mechanisms and policy-relevant conclusions

The purpose of our paper is to make a theoretical contribution to mechanism design and implementation theory that is both empirically motivated and empirically supported. This section substantiates this empirical motivation and support.

Ideal experiments for our purpose are those which (1) evaluate whether subject play is baseline-truthful in strategy-proof games; (2) compare two, largely equivalent direct-revelation games with complete information, one under a strategy-proof scf that is bossy and one that is non-bossy; (3) compare bossy mechanisms with no bossy dominant strategies under interior and non-interior information; (4) evaluate whether subject play obeys our assumption of weak payoff monotonicity and thus when data approximates mutual best responses it is consistent with empirical equilibrium.

Experiments fulfilling these objectives are readily available in the literature (Coppinger et al., 1980; Kagel et al., 1987; Kagel and Levin, 1993; Harstad, 2000; Attiyeh et al., 2000;

Chen and Sönmez, 2006; Cason et al., 2006; Healy, 2006; Andreoni et al., 2007; Li, 2017). We review them and categorize general findings as empirical results.[16]

5.1 Data is consistent with the properties identified by our theorems precluding undesirable equilibria

There is no evidence of a strategy-proof scf producing subject behavior approximating undesirable equilibria when the conditions in at least one of our theorems are satisfied. Indeed, we can find no evidence in any study of subject play resembling an undesirable equilibrium under a non-bossy mechanism or a mechanism with no bossy dominant strategies under interior information (i.e., Andreoni et al., 2007; Attiyeh et al., 2000; Cason et al., 2006; Chen and Sönmez, 2006; Healy, 2006; Kawagoe and Mori, 2001; Li, 2017). In other words, we observe no evidence of undesirable equilibria that are not empirical.

We show an additional level of robustness: in all the studies that we have surveyed, dominant strategy play is higher than the corresponding probability under uniform-random play. Using Lemmas 1–3, if this condition is satisfied, it means subjects will encounter significant disincentives to deviate whenever their actions lead to outcomes different from the social planner's objective. In other words, if subjects were to play strategies that produce the outcomes of an undesirable equilibrium, they would not be playing best-responses to empirical frequencies of play to the other agents.

In appendix table A.1, we survey the literature for experimental results with dominant strategy mechanisms. We find ten studies across a variety of mechanisms. In every experiment, rates of dominant strategy play exceed the threshold of uniform play. A simple binomial test—treating each of these ten papers as a single observation—rejects any null hypothesis that these rates of dominant strategy play are drawn from a random distribution with median probability at or below these levels ($p < 0.001$). Thus one would reasonably conclude that rates of dominant strategy play should exceed that under uniform support.

It is evident then that the accumulated experimental data supports the first policy relevant conclusion of our theory. It is worth noting that the simplicity and cleanness of our analysis is possible because our the strength of our results. Our lemmas transform the problem of testing whether behavior does not approximate a bad equilibrium into a single basic statistic test.

Policy-relevant conclusion: It is "safe"—in terms of the existence of strict incentives against undesirable behavior and no documented evidence of the persistence of that behavior—to operate a strategy-proof scf when the conditions in at least one of our theorems is satisfied.

[16] Given the extensive experimental evidence we have available, we feel that replicating some of these experiments or performing equivalent ones would not produce additional insights.

5.2 The risks identified by our theorems are substantiated by data

Evidence exists of behavior approximating undesirable equilibria in a variety of experimental studies. In all cases, the experimental environment does not satisfy the conditions of either of our theorems.

If the conditions on our theorems are not satisfied, we cannot rule out agents' play approximating an undesirable equilibrium for a strategy-proof scf in its direct form. Our survey of existing experimental literature reveals this possibility is substantive. Among the experiments we surveyed, Cason et al. (2006), Healy (2006), and Andreoni et al. (2007) involve the operation of a strategy-proof scf that violates non-bossiness in an information environment in which information is not interior. These are the only conditions for which our theory predicts the possibility of observing Nash equilibria that are not intended by the mechanism designer.

Andreoni et al. (2007) provide an excellent example for convergence to an undesirable equilibrium when a strategy-proof scf is operated. In their experiment, groups of four subjects play second-price auction games, simultaneously, for 30 rounds. Groups are rematched each round and play with the same private values across all different games. One game involves no information about the other players' valuations beyond the distribution from which they are drawn. A second game involves complete information. As Appendix figures Fig. A.1 and A.3 demonstrate, frequencies of play in both games converge toward mutual best response. However, in the complete information game these Nash equilibria have outcomes that are inconsistent with the intentions of the mechanisms designers, namely, bids deviate far more from their truthful valuation, inefficiency is possible and the tendency of low-valuation bidders to report zero persists under complete information.

Cason et al. (2006) provide a nearly ideal experiment comparison between a bossy and non-bossy mechanism in comparable table-game formats. Their conclusion is that in the bossy mechanism they cannot rule out unwanted Nash equilibria. Their non-bossy mechanism also satisfies rectangularity, thus unwanted Nash equilibrium does not exist for this game. Healy (2006) also provides similar conclusions.[17]

Policy-relevant conclusion: There are substantive risks—in terms of agents' play resembling an undesirable equilibrium—to operating a strategy-proof scf when the conditions on both our theorems are violated.

[17] In both Cason et al. (2006) and Healy (2006) agents do not know the structure of payoffs of the other agents, but know their payoffs are fixed throughout the experiment. Thus, when play stabilizes agents are responding to a conjecture of behavior of the same type of agent, a corner information structure. At the cost of notation and formalism our results can be extended to this environment.

5.3 Information matters

For strategy-proof scfs without a bossy dominant strategy, undesirable equilibria are generally not observed under an interior information structure. By contrast, undesirable equilibria are observed under complete information. That is, the observance of undesirable equilibria is dependent on whether the information structure allows undesirable empirical equilibria.

Andreoni et al. (2007)'s experiment gives us an ideal window to observe this phenomenon. As seen in Sec. 5.2, behavior in the complete information environment is characteristic of undesirable Nash equilibria. The second-price auction has no bossy dominant strategies: when information is interior, behavior cannot approximate an undesirable Nash equilibrium provided dominant strategy play is frequent enough. As Sec. 5.1 demonstrates, dominant strategy play is sufficiently frequent. Thus, if behavior approximates a Nash equilibrium in this informational setting, the outcomes obtained are necessarily close to the ideal, intended, equilibria of the second-price auction.

Behavior in Andreoni et al.'s second-price auction experiment approximates a Nash equilibrium under incomplete information. As in the complete information treatment, by the second half of the experiment, virtually all agents are playing best responses. Unlike in the complete information treatment, agents' deviations from their dominant strategies do not produce outcomes that differ greatly from the truthful equilibrium outcomes. As Fig. A.2 shows after the initial five rounds, median bids are the agents' own values, outcomes generally truthful and efficient, i.e., such that a highest valuation agent wins the auction at prices very close to the second valuation. The mechanism is achieving the social planner's objectives.

Our theory explains the differences in behavior between treatments in Andreoni et al.'s experiment. Under incomplete information, there is a penalty for a player to deviate too much from their dominant strategy. There is no corresponding penalty under complete information. In the complete information case, the highest valuation agent persistently overbids and the other agents persistently bid on a wide range under the highest valuation agent's value. As long as these behaviors are essentially separated, they are mutual best responses. Thus, the penalty from playing a weakly-dominated action is negligible given that all agents stick to these patterns of play. On the other hand, in the incomplete information treatment, for each bid, there is a positive probability that at least one agent bids her valuation. Since agents bid their values with high probability (68.2% on average), there is a non-trivial chance that a significant deviation from truthful behavior is suboptimal. Together, these experiments reveal that agents do react to pecuniary incentives and use information and observed frequencies of play of the other agents in a meaningful way. They do not preemptively react to a hypothetical tremble of the other agents, however.[18]

Policy-relevant conclusion: Performance of a strategy-proof scf that has no bossy dominant strategy is affected by the information structure in which it is operated. A designer should focus on operating such mechanisms only when it is plausible that information is interior.

5.4 Weak payoff monotonicity

Though a direct test of weak payoff monotonicity is impractical, the success of empirical models satisfying this property to fit data provides suggestive evidence of the applicability of this commonly-held assumption. A weaker test shows, across a wide variety of experiments, that frequency of subject play is highly correlated with expected payoffs. Violations of weak payoff monotonicity generally involve preferences unrelated to the numerical payoffs "induced" on subjects within an experimental game.

In experiments with games repeated for a number of periods with the same population and in which behavior becomes stable, weakly payoff monotone models typically do a good job at predicting final-period averages and comparative statics across different treatments. Experimental economists usually arrive at this conclusion by calculating maximum likelihood estimates of these models like the logistic Quantal Response Equilibrium (QRE).[19] Goeree et al. (2016) catalogue several hundred experiments, all which have successfully calibrated QRE models on their data.

To our knowledge there has been no formal econometric analysis of the specification of weakly payoff monotone models.[20] There are two main issues that one encounters to rigorously evaluate weak payoff monotonicity in a game. First, the empirical content defined by the closure of this theory is not a convex set. So it cannot be expressed as the intersection of linear inequalities, an essential feature for the application of most econometric approaches.

[18] Since the first experiments on the second-price auctions with private values of Coppinger et al. (1980) and Kagel and Levin (1993), experimental economists have observed that even though agents do not play their dominant strategy in these games, the probability with which they would have ended up disciplined by the market given other subject strategies is very low. Our analysis goes beyond this observation by showing that as predicted by empirical equilibrium analysis, the degree to which these deviations cause actual loss of subject earnings is linked to the non-bossiness properties of the scf and the information structure.

[19] Haile et al. (Footnote 2, 2008) note that it is indeed considered "unusual" when experimental data cannot be reconciled with monotone structural QRE, a more general monotone noisy best-response model. Recent experiments reported by Goeree and Louis (2021) provide further support for weak payoff monotonicity.

[20] Melo et al. (2019) develop a test for the structural Quantal Response hypothesis. This parametric hypothesis does not imply weak payoff monotonicity and constrains behavior between games with different payoff matrices for the same action spaces. These authors leverage these restrictions to construct a test of this model. Weak payoff monotonicity imposes no restriction on behavior across different payoff matrices. Even for a single game matrix it may impose less restrictions than the monotone structural Quantal Response hypothesis (Velez and Brown, 2020b).

Second, evaluating this property requires that one has a good estimate of agents' frequencies of play and expected utility for all actions. In games with relatively large action spaces, like those we surveyed (eg., in Andreoni et al. (2007) agents bid integers in the interval $\{0, ..., 200\}$; in Attiyeh et al., 2000 each agent has 2001 actions available), inference requires an impractically large number of observations.

Even though fully testing weak payoff monotonicity is not obvious, one can test for certain markers of this property that are less demanding on data. First, in weakly payoff monotone data sets there should be a positive correlation between the frequencies with which actions are played and their empirical expected utility. For the four studies where we have sufficient data (Andreoni et al., 2007; Attiyeh et al., 2000; Cason et al., 2006; Li, 2017), we can compare the actual payoffs earned with each action choice with the counterfactual payoffs had a subject chosen a different action. If subjects choose actions independent of payoffs—a gross violation of weak payoff monotonicity—we should suspect the differences between the average payoffs of played strategies and counterfactual payoffs of non-played strategies to be evenly distributed around zero. Instead we find in all cases the average payoffs of played strategies *exceed* those of non-played strategies.[21] Treating the 30 total sessions across these four studies as independent observations, we can easily reject the null hypothesis that strategies are played independent of expected payoffs ($p < 0.001$).[22]

Weak payoff monotonicity has implications that can be tested independently. For instance, it implies weakly dominant actions are played at least as frequently as uniform random play (see Sec. 5.1). Another useful implication is between actions related by weak domination. Suppose that action a_i weakly dominates action b_i for agent i with type θ_i. Independent of the behavior of the other agents, the expected payoff of a_i is greater than or equal to the expected payoff of b_i for this agent type. In a weakly payoff monotone distribution in the game, agent i with type θ_i will never play b_i more frequently than a_i. Thus, if σ is the profile of distributions of play and one can reject the hypothesis that $\sigma_i(a_i|\theta_i) \geq \sigma_i(b_i|\theta_i)$, one can also reject weak payoff monotonicity.

Using this strategy one can find evidence against weak payoff monotonicity. We are aware of three instances. First, in the Pivotal mechanism experiment of Cason et al. (2006), there are two dominant strategies for each agent. While the Column agent chooses them with similar frequencies (36.1% and 38.3%), the Row agent chooses them with frequencies 51.1% and 19.4%. Parametric paired t-tests and non-parametric signed rank and sign tests

[21] Using a conditional-logistic regression also produces positive coefficients in all cases. It also assumes a specific formalized structure on subject choice, making it a less general test.

[22] Specifically, in 30 out of 30 sessions the average strategy subjects played in a round had higher expected payoffs than those they didn't play. If we exclude all instances where subjects played a dominant strategy, this result holds in 28 out of 30 sessions.

suggest the latter difference is statistically significant at the subject level ($p < 0.01$), but not the former.

Second, there is a well documented propensity of overbidding in second-price auctions. This does not have to be necessarily at odds with weak payoff monotonicity. Agents who draw larger values will find overbidding with respect to value to be a less costly mistake than underbidding. Low-value agents will have fewer bids below their value than above their value. Thus, such a distribution of play can still be weakly payoff monotone and in aggregate overbid more than underbid. However, Figure 1 in Andreoni et al. (2007), which depicts the frequency of the difference between the bid of the low value agents and the maximal value, shows that these agents place significantly higher weight in the bids that are close to the maximal value agent. This is a clear violation of weak payoff monotonicity, because in the second-price auction a bid b above an agent's value v_i is weakly-dominated by all bids between v_i and b. Andreoni et al. (2007) argue that this behavior may be due to spiteful preferences of the low-value agents. Finally, a simple behavioral regularity as rounding to multiples of five, can easily induce violations of weak payoff monotonicity. Such patterns are present, among bids that are related by weak dominance, in the auction data of Andreoni et al. (2007), Brown and Velez (2020), and Li (2017).

With the exception of rounding, the aforementioned violations all come from data where there is evidence of subject play resembling an undesirable equilibrium. Interestingly, our model predicts these possibilities despite small violations of its underlying assumptions. We do not believe this is coincidental. When agents are indifferent over several different strategies, they may favor some strategies over others for reasons independent of personal payoffs. Such a practice would violate weak-payoff monotonicity. Because agents are not choosing these strategies based on the mechanism designer's objectives, there is a substantial danger that their play may approximate an undesirable equilibrium.

Consider an abstract example in labor management. There are two incentive structures A and B. Under incentive structure A, it is a (strictly) dominant strategy for employees to do their job correctly. Under B, it is a (weakly) dominant strategy for employees to do their job, but there are also 999 shirking activities that produce similar payoffs. Our model would correctly identify the dangers of incentive structure B. Imagine that data show employees do their job correctly with high probability under A, but do their job correctly and shirking activity no. 333 roughly half of the time under B. This latter behavior violates our model's assumption of weak-payoff monotonicity, since all shirking options should be played with the same probability. Nonetheless, our model's prediction is still very useful for mechanism designers of labor incentives.

What about activity no. 333? There must be some appeal of this activity that goes beyond simple payoff structures. We have been testing weak payoff monotonicity under the assumption that numerical payoffs induced in an experimental game are representative of actual payoffs perceived by subjects. Admittedly, this in an imperfect assumption. Agents may have different objectives when they engage in strategic interactions. For instance, they can have other regarding preferences. They could be attracted to some actions because of their labels. They could use mental shortcuts as rounding. If we allowed all these other options to be included, we would lose the discipline of falsifiability on our model's predictions. We also would lose the applicability of empirical equilibrium to the universe of simultaneous-move finite games.

This example has an experimental analogue with the complete information treatment of Andreoni et al. (2007). Our model predicts undesirable equilibria are possible within this environment, something that is indeed found within the data. However, the authors note that specific equilibria that are most played are consistent with spiteful bidding. Much like the appeal of activity no. 333 in our previous example, we do not view these explanations as competing, rather they are complementary. Our model answers the question on why the desired equilibrium is not being played; the incentives are not strict enough to rule out other options. The explanation of spiteful bidding explains why a particular undesirable equilibrium is so appealing. As we do not consistently observe spite dominating payoff-maximizing behavior in general, we conclude that an examination of existing incentive structures is necessary before other explanations are used to explain subject behavior.

Policy-relevant conclusion: The fundamental assumption of our model, weak payoff monotonicity, generally describes behavioral data well, but there are some anomalies. Even in data where these anomalies are present, it is still a good idea to follow the general guidelines presented here on operating a strategy-proof scf.

5.5 Empirical findings in the field

Strategy-proof mechanisms have been operated for some time in the field. Empirical studies of such mechanisms have generally corroborated the observations from laboratory experiments (e.g. Hassidim et al., 2020; Rees-Jones, 2017; Artemov et al., 2021; Chen and Pereyra, 2019; Shorrer and Sóvágó, 2022), in such high stakes environments as career choice (Roth, 1984) and school choice (Abdulkadiroğlu and Sönmez, 2003). Among these papers, Artemov et al. (2021) and Chen and Pereyra (2019) are the closest to ours. Besides presenting empirical evidence of persistent violations of the dominant-strategy hypothesis, they propose theoretical explanations for it. They restrict to school choice environments in which a particular mechanism is used. Artemov et al. (2021) study a continuum model in which the SPDA mechanism is operated in an interior incomplete information environment. They conclude that it is reasonable that one can observe equilibria in which agents deviate from the ideal but only make mistakes that do not affect equilibrium outcomes. Their construction is based on the approximation of the continuum economy by means of finite realizations of it in which agents are allowed to make mistakes that vanish as the population grows. Chen and Pereyra (2019) study a finite school choice environment in which there is a unique ranking of students across all schools. Based on the analysis of an ordinal form of equilibrium, the authors argue that only when information is not interior can an agent be expected to deviate from her truthful report. Our study substantially differs in its scope with these two papers, because our results apply to all private values environments that admit a strategy-proof scf. When applied to a school choice problem, our results are qualitatively in line with those in these two studies and thus provide a rationale for their empirical findings. However, our results additionally explain the causes of behavior in these environments (informational assumptions and specific properties of the mechanisms) and provide exact guidelines of when these phenomena will be present in any other environment that accepts a strategy-proof mechanism.

6 Discussion and concluding remarks

Under the direct-revelation mechanism of a strategy-proof scf, agents' behavior may approximate Nash equilibria with outcomes that differ from the intent of the mechanism's designer. We address this possibility theoretically and empirically. We tailor our analysis to a mechanism designer concerned about the plausibility of this possibility.

Three novel theoretical insights come from our analysis. First, as long as agents are truthful frequently enough, the behavior in a direct-revelation game of a strategy-proof scf, g, will only support Nash equilibria that implement g whenever one of two conditions are satisfied: either g is non-bossy, or g has no bossy dominant strategy and information is interior. Second, whenever a strategy-proof scf that violates non-bossiness is operated in an environment in which information is not interior, behavior may approximate a Nash equilibrium that produces outcomes that do not implement g. Third, the performance of a strategy-proof scf that has no bossy dominant strategy is affected by the information environment in which it is operated. In other words, we articulate a folk theorem stating that some dominant strategy mechanisms should perform better whenever there is enough uncertainty.

Empirical evidence supports our findings. There is strong evidence that agents are truthful frequently enough in direct-revelation games of strategy-proof scfs. Thus, under the conditions on an scf in our theorems, Nash behavior in the direct-revelation game of the scf can only essentially produce the outcomes selected by the scf for the true types. Laboratory experiments also confirm our second theoretical insight. There is evidence of behavior in direct-revelation games of strategy-proof scfs violating non-bosiness that converges to sub-optimal equilibria when information is complete. As predicted by our theorems, the same mechanisms operated in an interior information structure are immune to these problems.

Our theoretical analysis is based upon a refinement we introduce, empirical equilibrium. This refinement selects the equilibria that can be approached by weakly payoff monotone behavior. Weak payoff monotonicity is satisfied by popular noisy best-response models that have been successful in replicating comparative statics in laboratory experiments. Since weak payoff monotonicity is compatible with noisy best-responses, the refinement does not rule out all equilibria involving weakly-dominated behavior.

The idea to refine Nash equilibrium by means of the proximity to plausible behavior has precedent in the literature. Harsanyi (1973) addressed the plausibility of Nash equilibrium itself by approximating in each game at least one Nash equilibrium by means of behavior that is unambiguosly determined by utility maximization in additive randomly perturbed payoff models with vanishing perturbations. The main difference with our construction is that the theory in which we base approximation is non-parametric and disciplined by an a priori restriction that allows us to narrow the set of equilibria that can be approximated.[23] Our refinement is also closely related to Rosenthal (1989)'s approximation of equilibria by a particular linear random choice model that evolves towards best responses and is defined only in games with two actions, van Damme (1991)'s firm equilibria and vanishing control costs approachable equilibria, and McKelvey and Palfrey (1996)'s logistic QRE approachable equilibria. These authors propose parametric theories to account for deviations from utility maximization in games and require equilibria to be approachable by the empirical content of these theories. Behavior generated by each of these theories satisfies weak payoff monotonicity. Thus they generate further refinements of empirical equilibrium.

These previous attempts to refine Nash equilibrium by means of approachability were never studied as stand alone refinements. van Damme (1991) developed firm equilibria and vanishing control costs approachable equilibria as basic frameworks to add restrictions and provide foundations for other equilibrium refinements that do eliminate weakly-dominated behavior, e.g., Selten (1975)'s perfect equilibria. McKelvey and Palfrey (1996) observed that Nash equilibrium can be refined based on approachability by logistic QRE behavior, but did not pursue the study of this refinement further. Thus, a significant contribution of our work is to show that a robust non-parametric generalization of these refinements can inform us about the incentives for truthful revelation in dominant strategy games.

[23] Each Nash equilibrium can be approached by a sequence of behavior in Harsanyi (1973)'s randomly perturbed payoff models with vanishing perturbations (Velez and Brown, 2020b).

Besides the application of empirical equilibrium to the analysis of strategy-proof mechanisms in the present paper, we have advanced, in three companion papers, the foundations of empirical equilibrium (Velez and Brown, 2020b) and the study of this equilibrium concept in a partnership-dissolution environment (Velez and Brown, 2020a; Brown and Velez, 2020). In Velez and Brown (2020b) we show that empirical equilibrium can be equivalently defined by means of approximation of behavior in van Damme (1991)'s control costs games and Goeree et al. (2005)'s regular QRE. We also show that there is a meaningful difference between this refinement and every possible refinement based on monotone additive randomly perturbed payoff models (e.g., firm equilibria, logistic QRE approachable equilibria). In Velez and Brown (2020a) we advance empirical equilibrium analysis of partnership-dissolution auctions, a family of non-strategy-proof mechanisms in which no agent has available a weakly dominant strategy. In Brown and Velez (2020) we experimentally test the comparative statics predicted by empirical equilibrium in partnership-dissolution auctions. We encourage others to identify other environments in which empirical equilibrium analysis produces results that are relevant for mechanism design and game theory. In particular it is interesting to generalize empirical equilibrium to extensive form games and explore its applications.

Finally, some technical remarks. We have deliberately concentrated our analysis on direct-revelation mechanisms. Our results obviously extend to mechanisms that are strategically equivalent to these direct-revelation games. For this reason, we have made no distinction between some direct-revelation mechanisms and their alternative abstract forms in our presentation. For instance, the second-price auction is a mechanism in which agents simply bid for a good, one of the highest bidders get the good and pays the second-highest bid. This mechanism is equivalent to the direct-revelation game of the social choice function that assigns the object to an agent with the highest valuation and charges this agent the second highest valuation among all agents.

It is an open question whether there is an environment that admits a strategy-proof scf violating non-bossiness and for which there is a mechanism that implements the scf in empirical equilibrium for all common-prior information structures. At the cost of slightly heavier notation one can show that the statements in Theorem 1 are equivalent to the existence of a finite *dominant-strategy* mechanism (M, φ) that implements the scf in empirical equilibria for all common-prior information structures. Thus, our restriction to direct-revelation games is to some extent without loss of generality. If one were to implement a strategy-proof scf that violates non-bossiness robustly in empirical equilibria, one would have to renounce partial implementation in dominant strategies.

There is a sense in which our restriction to direct-revelation games is not without loss of generality, however. A mechanism designer may be interested not in precluding the existence of suboptimal equilibria, but in minimizing the probability that undesirable outcomes occur. For this purpose, general mechanisms may offer improvements over direct-revelation mechanisms. Think for instance of a social planner who oversees the lunch choices of an agent. There are two options, healthy and unhealthy. The social planner would like the agent to choose the healthy choice whenever it is the best for them. Suppose that the social planner uses a direct-revelation mechanism. That is, the social planner asks the agent for their preference and then assigns the agent's preferred option breaking ties in favor of the healthy choice. Then in each empirical equilibrium of the game (choice problem in this case) the social planner would obtain their objective for sure when the agent is not indifferent between options and with two thirds probability whenever the agent is indifferent between both choices. Suppose that instead the social planner offers the agent a menu of $k+1$ choices, where the first $k > 2$ choices lead to the healthy option and the last choice to the unhealthy. Then in each empirical equilibrium of the game the social planner would obtain their objective for sure when the agent is not indifferent between options and with $k/(k+1)$ probability whenever the agent is indifferent between both choices. Thus, the social planner can virtually implement their objective in empirical equilibria with a general mechanism.

Finally, it is known that the restriction to social choice *functions* is not without loss of generality in robust implementation. Bergemann and Morris (2005, Example 2) show that "partial" robust implementation can be achieved for a "social choice correspondence" that does not posses any strategy-proof single-valued selection. Their argument can be adapted to account for mixed strategies, which are essential in our analysis, and to show that the same phenomenon happens in our environment (see Example 1 in our Online Appendix).

Appendix

Proof of Lemma 1. Let g be strategy-proof, p a common prior, and σ an empirical equilibrium of $\Gamma := (\Theta, g, p)$. We prove that for each $i \in N$ and each $\theta_i \in \Theta_i$, θ_i is in the support of $\sigma_i(\cdot|\theta_i)$. Consider a sequence of weakly payoff monotone distributions for Γ, $\{\sigma^\lambda\}_{\lambda\in\mathbb{N}}$, such that for each $i \in N$ and each $\theta_i \in \Theta_i$, as $\lambda \to \infty$, $\sigma^\lambda(\cdot|\theta_i) \to \sigma(\cdot|\theta_i)$. Let $\lambda \in \mathbb{N}$ and $\theta_{-i} \in \Theta_{-i}$. Since θ_i is a weakly dominant action for agent i with type θ_i in (Θ, g), for each $\tau_i \in \Theta_i$, $u_i(\varphi(\theta_{-i}, \theta_i)|\theta_i) \geq u_i(\varphi(\theta_{-i}, \tau_i)|\theta_i)$. Thus, $U_\varphi(\sigma^\lambda_{-i}, \delta_{\theta_i}|p, \theta_i) \geq U_\varphi(\sigma^\lambda_{-i}, \delta_{\tau_i}|p, \theta_i)$. Since σ is weakly payoff monotone for Γ, we have that for each $\tau_i \in \Theta_i$, $\sigma^\lambda_i(\theta_i|\theta_i) \geq \sigma^\lambda_i(\tau_i|\theta_i)$. Convergence implies that $\sigma_i(\theta_i|\theta_i) \geq \sigma_i(\tau_i|\theta_i)$. Thus, θ_i is in the support of $\sigma_i(\cdot|\theta_i)$. □

Proof of Lemma 2. Let us make three assumptions.

Assumption (i): Let g be strategy-proof, p a common prior, and $\sigma \in N(\Theta, g, p)$.

Assumption (i'): σ is a baseline-truthful strategy profile.

Assumption (ii): Suppose that for each $\theta \in \Theta$ in the support of p, each τ in the support of $\sigma(\cdot|\theta)$, and each $i \in N$ such that $u_i(g(\tau)|\theta_i) = u_i(g(\tau_{-i}, \theta_i)|\theta_i)$ we have that $g(\tau) = g(\tau_{-i}, \theta_i)$.

We claim that if Assumptions (i,i') and (ii) are satisfied, σ is necessarily optimal. To see this, let $\theta \in \Theta$ be in the support of p and τ be in the support of $\sigma(\cdot|\theta)$. Consider first agent 1. Since τ is in the support of the equilibrium in state θ, the expected utility of τ_1 for agent 1 is greater than or equal to the expected utility of θ_1. Since θ_1 is a weakly dominant strategy for agent 1 in this state, these utilities are equal. Moreover, the integrand of the expected utility of θ_1 dominates point-wise the integrand of the expected utility of τ_1. This means that these integrands need to be equal in the support of the common integrating measure of both integrals. Since τ_{-1} is in the support of this integrating measure, we have that $u_1(g(\tau)|\theta_1) = u_1(g(\tau_{-1}, \theta_1)|\theta_1)$. Thus, by Assumption (ii), $g(\tau) = g(\tau_{-1}, \theta_1)$. Now, by Assumption (i'), we have that (τ_{-1}, θ_1) is also in the support of $\sigma(\cdot|\theta)$. We can iterate then and conclude that $g(\tau) = g(\theta)$.

Thus, we have proved that if Assumptions (i,i') are satisfied and σ is sub-optimal, Assumption (ii) needs to be violated. That is, there must be $\theta \in \Theta$ in the support of p, τ in the support of $\sigma(\cdot|\theta)$, and $i \in N$ such that $u_i(g(\tau)|\theta_i) = u_i(g(\tau_{-i}, \theta_i)|\theta_i)$, and $g(\tau) \neq g(\tau_{-i}, \theta_i)$. □

Proof of Lemma 3. Let g be strategy-proof, p a common prior, and σ a baseline-truthful strategy profile in (Θ, g, p). Let $\theta \in \Theta$ in the support of p, τ_{-i} in the support of $\sigma_{-i}(\cdot|\theta_{-i})$, $i \in N$, and $\tau_i \in \Theta_i$ such that $g(\tau) \neq g(\tau_{-i}, \theta_i)$. We prove that each of the additional assumptions in the lemma imply that τ_i is not a best response to σ_{-i} for agent i.

Suppose first that g is non-bossy. Since $g(\tau) \neq g(\tau_{-i}, \theta_i)$, it must be the case that $u_i(\tau|\theta_i) \neq u_i(\tau_{-i}, \theta_i|\theta_i)$. Since g is strategy-proof $u_i(\tau|\theta_i) < u_i(\tau_{-i}, \theta_i|\theta_i)$. Since τ_{-i} is played with positive probability by $N \setminus \{i\}$, and truthful reports are weakly dominant strategies, $U_g(\sigma_{-i}, \tau_i|p, \theta_i) < U_g(\sigma_{-i}, \theta_i|p, \theta_i)$. Thus, τ_i is not a best response to σ_{-i} for agent i.

Suppose now that g has no bossy dominant strategy and p has full support. Since $g(\tau) \neq g(\tau_{-i}, \theta_i)$, it must be the case that τ_i is not a weakly dominant strategy for agent i with type θ_i. Thus, there is τ'_{-i} such that $u_i(\tau'_{-i}, \tau_i|\theta_i) < u_i(\tau'_{-i}, \theta_i|\theta_i)$. Since p has full support, (τ'_{-i}, θ_i) is in the support of p. Since truthful reports are played with positive probability and truthful reports are weakly dominant strategies, $U_g(\sigma_{-i}, \tau_i|p, \theta_i) < U_g(\sigma_{-i}, \theta_i|p, \theta_i)$. Thus, τ_i is not a best response to σ_{-i} for agent i. □

Proof of Lemma 4. Let g be an scf. Suppose that for each interior prior p and each empirical equilibrium of (Θ, g, p), σ, we have that for each pair $\{\theta, \tau\} \subseteq \Theta$ where τ is in the support of $\sigma(\cdot|\theta)$, $g(\theta) = g(\tau)$.

We claim that g is strategy-proof. Our proof of this claim follows Bergemann and Morris (2005, Proposition 3). We spell out the details because our statement includes mixed strategy equilibria. Let $\theta \in \Theta$, $i \in N$, and $\tau_i \in \Theta_i$. Let $\varepsilon \in (0,1)$. Consider the common prior p that places probability $1/2-\varepsilon/2$ on each element of $\{\theta, (\theta_{-i}, \tau_i)\}$, and places uniform probability on all other payoff types. Thus, p is interior. Let σ be an equilibrium of (Θ, g, p) that, for each $\mu \in \Theta$ and each message in the support of $\sigma(\cdot|\mu)$, produces $g(\mu)$. Thus, the expected value of a report in the support of $\sigma_i(\cdot|\theta_i)$ has an expected value for type θ_i that is greater than or equal to the expected value of a report in the support of $\sigma_i(\cdot|\tau_i)$, i.e.,

$$\begin{aligned}&p(\theta_{-i}|\theta_i)u_i(g(\theta)|\theta_i) + \textstyle\sum_{\mu_{-i}\in\theta_{-i}} p(\mu_{-i}|\theta_i)u_i(g(\mu_{-i},\theta_i)|\theta_i) \geq \\ &p(\theta_{-i}|\theta_i)u_i(g(\theta_{-i},\tau_i)|\theta_i) + \textstyle\sum_{\mu_{-i}\in\theta_{-i}} p(\mu_{-i}|\theta_i)u_i(g(\mu_{-i},\tau_i)|\theta_i).\end{aligned}$$

Since as $\varepsilon \to 0$, $p(\theta_{-i}|\theta_i) \to 1$, we have that $u_i(g(\theta)|\theta_i) \geq u_i(g(\theta_{-i},\tau_i)|\theta_i)$. Thus, g is strategy-proof.

We now claim that g has no bossy dominant strategy. Suppose by contradiction that there are $i \in N$, $\theta \in \Theta$, $\tau_i \in \Theta_i$, such that $u_i(g(\theta)|\theta_i) = u_i(g(\theta_{-i},\tau_i)|\theta_i)$, $g(\theta) \neq g(\theta_{-i},\tau_i)$, and for each $\tau_{-i} \in \Theta_{-i}$, $u_i(g(\tau_{-i},\theta_i)|\theta_i) \leq u_i(g(\tau)|\theta_i)$. Let p have full support. Let σ be an empirical equilibrium of (Θ, g, p). Since g is strategy-proof, τ_i is a weakly dominant action for agent i with type θ_i in (Θ, g), and for each $j \in N \setminus \{i\}$, θ_j is a dominant strategy for agent j with type θ_j. By Lemma 1, $\sigma(\cdot|\theta)$ places positive probability on (θ_{-i},τ_i). This contradicts Statement 1 in the theorem. □

Proof: 2 ⇒ 1 *in Theorems 1 and 2.* Suppose that g is strategy-proof and let σ be an empirical equilibrium of the direct revelation game of g. From Lemma 1 we know that σ is a baseline-truthful equilibrium. We claim that σ implements g. Suppose by contradiction that it is not. By Lemma 2 we know that at least an agent ends up being bossy with positive probability. This contradicts, Lemma 3, which states that under the assumptions of the theorems, no empirical equilibrium of a direct revelation game of g admits an agent be bossy with positive probability. □

Proof of Lemma 5. Let g be strategy-proof. Suppose that for each common prior p and each empirical equilibrium of (Θ, g, p), σ, we have that for each pair $\{\theta, \tau\} \subseteq \Theta$ where θ is in the support of p and τ is in the support of $\sigma(\cdot|\theta)$, $g(\theta) = g(\tau)$. We claim that g is non-bossy. Suppose by contradiction that g violates this property. We can suppose without loss of generality that there is $\theta \in \Theta$ and $\tau_1 \in \Theta_1$ such that $u_1(g(\theta)|\theta) = u_1(g(\theta_{-1},\tau_1)|\theta)$ and $g(\theta) \neq g(\theta_{-1},\tau_1)$.

Consider the complete information prior p that places probability one on state θ. For each $i \in N$, let D_i be the set of weakly dominant actions for agent i with type θ_i in (Θ, g, θ). Let σ be the strategy profile in which each agent $i > 1$ uniformly randomizes on the set D_i, and agent 1 uniformly randomizes on the set of best responses to σ_{-1}. By construction σ is a Nash equilibrium of (Θ, g, θ). We claim that σ is an empirical equilibrium of (Θ, g, θ) in which agent 1 plays τ_1 with positive probability.

Note that g satisfies the hypotheses of Lemma 4. Thus, g has no bossy dominant strategy.

We prove first that agent 1 plays τ_1 with positive probability in σ. That is, τ_1 is a best response to σ_{-1} for agent 1 with type θ_1 in (Θ, g, θ). Let θ'_{-1} be in the support of σ_{-1}. Recall that by hypothesis, $u_1(g(\theta)|\theta_1) = u_1(g(\theta_{-1}, \tau_1)|\theta_1)$. Since θ'_2 is a dominant strategy for agent 2 with type θ_2 in (Θ, g, θ), and g has no bossy dominant strategy, we have that $g(\theta) = g(\theta_{-\{1,2\}}, \theta_1, \theta'_2)$ and $g(\theta_{-1}, \tau_1) = g(\theta_{-\{1,2\}}, \tau_1, \theta'_2)$. By repeating this step we get that $g(\theta) = g(\theta'_{-1}, \theta_1)$ and $g(\theta_{-1}, \tau_1) = g(\theta'_{-1}, \tau_1)$. Then, $u_1(g(\theta'_{-1}, \theta_1)|\theta_1) = u_1(g(\theta'_{-1}, \tau_1)|\theta_1)$. Thus, $U_g(\sigma_{-1}, \tau_1|p, \theta_1) = U_g(\sigma_{-1}, \theta_1|p, \theta_1)$. Since θ_1 is a weakly dominant action for agent 1 with type θ_1 in (Θ, g, θ), τ_1 is a best response to σ_{-1} for agent 1. Consequently, agent 1 plays τ_1 with positive probability in σ_1.

We finally show that σ is an empirical equilibrium of (Θ, g, θ). We construct a sequence of weakly payoff monotone distributions for this game that converges to σ. Let $k \in \mathbb{N}$ and $0 < \varepsilon < 1$. If for some $i > 1$ all actions are in D_i, let $\sigma_i^k := \sigma_i$ and $\sigma_i^{k,\varepsilon} := \sigma_i$. Note that if for each $i > 1$ all actions are in D_i, σ is itself a weakly payoff monotone distribution. So σ is an empirical equilibrium. Thus, we can suppose without loss of generality that for some $i > 1$, there are some possible reports that are not in D_i. For each such agent let $\sigma_i^{k,\varepsilon}$ be the strategy that distributes $1 - \varepsilon$ uniformly among D_i and distributes ε unifromly among the actions not in D_i. Let $\sigma_1^{k,\varepsilon}$ be the strategy for agent 1 defined as follows: for each $\theta'_1 \in \Theta_1$,

$$\sigma_1^{k,\varepsilon}(\theta'_1) := \frac{e^{kU_g(\sigma_{-1}^{k,\varepsilon}, \theta'_1|p,\theta_1)}}{\sum_{\theta''_1 \in \Theta_1} e^{kU_g(\sigma_{-1}^{k,\varepsilon}, \theta''_1|p,\theta_1)}}.$$

Since the exponential function is positive, $\sigma_1^{k,\varepsilon}$ is an interior probability distribution. Thus, $\sigma^{k,\varepsilon}$ is an interior strategy profile.

We claim that if $\varepsilon < \min_{i>1, |D_i|<|\Theta_i|} |D_i|/|\Theta_i|$, $\sigma^{k,\varepsilon}$ is weakly payoff monotone for (Θ, g, θ). Since the exponential function is strictly increasing, $\sigma_1^{k,\varepsilon}$ is ordinally equivalent to the expected utility vector $(U_g(\sigma_{-1}^{k,\varepsilon}, \theta''_1|p, \theta_1))_{\theta''_1 \in \Theta_1}$. Thus, for any two actions $\{\theta'_1, \theta''_1\} \subseteq \Theta_1$, $U_g(\sigma_{-1}^{k,\varepsilon}, \theta'_1|p, \theta_1) \geq U_g(\sigma_{-1}^{k,\varepsilon}, \theta''_1|p, \theta_1)$ if and only if $\sigma_1^{k,\varepsilon}(\theta'_1) \geq \sigma_1^{k,\varepsilon}(\theta''_1)$. If $i > 1$ is such that $|D_i| = |\Theta_i|$, $\sigma_i^{k,\varepsilon}$ places equal probability on all actions. Thus, it induces no violation of weak payoff monotonicity. Finally, let $i > 1$ be such that $|D_i| < |\Theta_i|$. Since $\varepsilon < \min_{i>1, |D_i|<|\Theta_i|} |D_i|/|\Theta_i|$, $\sigma_i^{k,\varepsilon}$ places higher probability on each element of D_i than on each element of $\Theta_i \setminus D_i$. Since no element of $\Theta_i \setminus D_i$ is a weakly dominant strategy for agent i with type θ_i in (Θ, g, θ) and $\sigma^{k,\varepsilon}$ is interior, for each $\theta'_i \in D_i$ and each $\theta''_i \in \Theta_i \setminus D_i$, $U_g(\sigma^{k,\varepsilon}_{-i}, \theta'_i|p, \theta_i) > U_g(\sigma^{k,\varepsilon}_{-i}, \theta''_i|p, \theta_i)$. Consequently, $\sigma^{k,\varepsilon}_i$ induces no violation of weak payoff monotonicity.

Note that as $\varepsilon \to 0$, for each $i > 1$, $\sigma^{k,\varepsilon}_i \to \sigma_i$. By continuity of expected utility and exponential operators, as $\varepsilon \to 0$,

$$\frac{e^{kU_g(\sigma^{k,\varepsilon}_{-1}, \theta'_1|p,\theta_1)}}{\sum_{\theta''_1 \in \Theta_1} e^{kU_g(\sigma^{k,\varepsilon}_{-1}, \theta''_1|p,\theta_1)}} \to \frac{e^{kU_g(\sigma_{-1}, \theta'_1|p,\theta_1)}}{\sum_{\theta''_1 \in \Theta_1} e^{kU_g(\sigma_{-1}, \theta''_1|p,\theta_1)}}.$$

Since for each θ'_1 in the support of σ_1

$$\frac{e^{kU_g(\sigma_{-1}, \theta'_1|p,\theta_1)}}{\sum_{\theta''_1 \in \Theta_1} e^{kU_g(\sigma_{-1}, \theta''_1|p,\theta_1)}} = \frac{e^{kU_g(\sigma_{-1}, \theta_1|p,\theta_1)}}{\sum_{\theta''_1 \in \Theta_1} e^{kU_g(\sigma_{-1}, \theta''_1|p,\theta_1)}},$$

we have that there is $\varepsilon(k) \in (0, 1/k)$ for which for each θ'_1 in the support of σ_1,

$$\left| \frac{e^{kU_g(\sigma^{k,\varepsilon(k)}_{-1}, \theta'_1|p,\theta_1)}}{\sum_{\theta''_1 \in \Theta_1} e^{kU_g(\sigma^{k,\varepsilon(k)}_{-1}, \theta''_1|p,\theta_1)}} - \frac{e^{kU_g(\sigma_{-1}, \theta_1|p,\theta_1)}}{\sum_{\theta''_1 \in \Theta_1} e^{kU_g(\sigma_{-1}, \theta''_1|p,\theta_1)}} \right| < \frac{1}{k}. \tag{1}$$

Let $\sigma^k := \sigma^{k,\varepsilon(k)}$. By construction, σ^k is weakly payoff monotone for (Θ, g, θ). Moreover, as $k \to \infty$, $\varepsilon(k) \to 0$. Thus, as $k \to \infty$, for each $i > 1$, $\sigma^k_i \to \sigma_i$. By (1), as $k \to \infty$, for each θ'_1 in the support of σ_1,

$$\frac{e^{kU_g(\sigma^{k}_{-1}, \theta'_1|p,\theta_1)}}{\sum_{\theta''_1 \in \Theta_1} e^{kU_g(\sigma^{k}_{-1}, \theta''_1|p,\theta_1)}} \to \frac{e^{kU_g(\sigma_{-1}, \theta_1|p,\theta_1)}}{\sum_{\theta''_1 \in \Theta_1} e^{kU_g(\sigma_{-1}, \theta''_1|p,\theta_1)}}.$$

Finally, let θ'_1 be outside the support of σ_1. Then, $U_g(\sigma_{-1}, \theta'_1|p, \theta_1) < U_g(\sigma_{-1}, \theta_1|p, \theta_1)$. It follows that,

$$\sigma^k_1(\theta'_1)/\sigma^k_1(\theta_1) = e^{k\left(U_g(\sigma_{-1}, \theta'_1|p,\theta_1) - U_g(\sigma_{-1}, \theta_1|p,\theta_1)\right)} \underset{k\to\infty}{\to} 0.$$

Thus, as $k \to \infty$, $\sigma^k_1 \to \sigma_1$ and consequently, $\sigma^k \to \sigma$.

In summary, σ is a Nash equilibrium of (Θ, g, θ) that is the limit of a sequence of weakly payoff monotone strategies for (Θ, g, θ); moreover, σ prescribes each agent be truthful with positive probability and agent 1 play τ_1 with positive probability. Thus, σ is an empirical equilibrium of (Θ, g, θ) in which in state θ, which is realized with probability one in this environment, the profile of reports (θ_{-1}, τ_1) is realized with positive probability. Since $g(\theta) \neq g(\theta_{-1}, \tau_1)$, this is a contradiction to the hypothesis of the lemma. □

References

Abdulkadiroğlu, A., Sönmez, T., June 2003. School choice: A mechanism design approach. Amer Econ Review 93 (3), 729–747.

Adachi, T., 2014. Robust and secure implementation: equivalence theorems. Games Econ Behavior 86 (0), 96 – 101.

Andreoni, J., Che, Y.-K., Kim, J., 2007. Asymmetric information about rivals' types in standard auctions: An experiment. Games Econ Behavior 59 (2), 240 – 259.

Arribilaga, P., Masso, J., Neme, A., 2020. All sequential allotment rules are obviously strategy-proof, Working paper U. A. Barcelona and U. N. San Luis.

Artemov, G., Che, Y.-K., He, Y., 2021. Strategic 'mistakes': Implications for market design research, Mimeo.

Ashlagi, I., Gonczarowski, Y. A., 2018. Stable matching mechanisms are not obviously strategy-proof. J Econ Theory 177, 405 – 425.

Attiyeh, G., Franciosi, R., Isaac, R. M., Jan 2000. Experiments with the pivot process for providing public goods. Public Choice 102 (1), 93–112.

Bade, S., Gonczarowski, Y. A., 2017. Gibbard-satterthwaite success stories and obvious strategyproofness.

Barbera, S., 2010. Strategy-proof social choice. In: Arrow, K., Sen, A., Suzumura, K. (Eds.), Handbook of Social Choice and Welfare. Vol. 2. North-Holland, Amsterdam, New York, Ch. 25, pp. 731–831.

Benassy, J. P., 1982. The economics of market disequilibrium. New York: Academic Press.

Bergemann, D., Morris, S., 2005. Robust mechanism design. Econometrica 73 (6), 1771–1813.

Bergemann, D., Morris, S., 2011. Robust implementation in general mechanisms. Games and Economic Behavior 71 (2), 261 – 281.

Bo, I., Hakimov, R., 07 2019. Iterative Versus Standard Deferred Acceptance: Experimental Evidence. The Economic Journal 130 (626), 356–392.

Bo, I., Hakimov, R., 2020. Pick an object mechanism.

Bochet, O., Sakai, T., 2010. Secure implementation in allotment economies. Games Econ Behavior 68 (1), 35 – 49.

Bochet, O., Tumennassan, N., 2020. Prevalence of truthtelling and implementation, mimeo.

Brown, A. L., Velez, R. A., 2020. Empirical bias and efficiency of alpha-auctions: experimental evidence.
URL https://arxiv.org/abs/1905.03876

Cabrales, A., Ponti, G., 2000. Implementation, elimination of weakly dominated strategies and evolutionary dynamics. Review of Economic Dynamics 3 (2), 247 – 282.

Cason, T. N., Saijo, T., Sjöström, T., Yamato, T., 2006. Secure implementation experiments: Do strategy-proof mechanisms really work? Games Econ Behavior 57 (2), 206 – 235.

Chen, L., Pereyra, J., 2019. Self-selection in school choice. Games Econ Behavior 117, 59 – 81.

Chen, Y., Sönmez, T., 2006. School choice: an experimental study. Journal of Economic Theory 127 (1), 202 – 231.

Cooper, D. J., Fang, H., 2008. Understanding overbidding in second price auctions: An experimental study*. The Economic Journal 118 (532), 1572–1595.

Coppinger, V. M., Smith, V. L., Titus, J. A., 1980. Incentives and behavior in english, dutch and sealed-bid auctions. Economic Inquiry 18 (1), 1–22.

Davis, D. D., Holt, C. A., 1993. Experimental economics. Princeton university press.

de Clippel, G., October 2014. Behavioral implementation. American Economic Review 104 (10), 2975–3002.

de Clippel, G., Saran, R., Serrano, R., 06 2018. Level-k Mechanism Design. The Review of Economic Studies 86 (3), 1207–1227.

Dekel, E., Scotchmer, S., 1992. On the evolution of optimizing behavior. J Econ Theory 57 (2), 392 – 406.

Eliaz, K., 2002. Fault tolerant implementation. Review Econ Studies 69 (3), 589–610.

Esponda, I., Pouzo, D., 2016. Berk–nash equilibrium: A framework for modeling agents with misspecified models. Econometrica 84 (3), 1093–1130.

Fudenberg, D., He, K., 2021. Player-compatible learning and player-compatible equilibrium. J Econ Theory 194, 105238.

Fujinaka, Y., Wakayama, T., 2011. Secure implementation in Shapley-Scarf housing markets. Econ Theory 48 (1), 147–169.

Gale, D., Shapley, L. S., 1962. College admissions and the stability of marriage. American Math Monthly 69 (1), 9–15.

Gibbard, A., 1973. Manipulation of voting schemes: A general result. Econometrica 41 (4), 587–601.

Goeree, J. K., Holt, C. A., Palfrey, T. R., 2005. Regular quantal response equilibrium. Experimental Economics 8 (4), 347–367.

Goeree, J. K., Holt, C. A., Palfrey, T. R., 2016. Quantal Response Equilibrium: A Stochastic Theroy of Games. Princeton Univ. Press, Princeton, NJ.

Goeree, J. K., Louis, P., December 2021. M equilibrium: A theory of beliefs and choices in games. American Economic Review 111 (12), 4002–45.

Goeree, J. K., Louis, P., Zhang, J., 2018. Noisy introspection in the 11–20 game. The Economic Journal 128 (611), 1509–1530.

Green, J., Laffont, J.-J., 1977. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. Econometrica 45 (2), 427–438.

Haile, P. A., Hortaçsu, A., Kosenok, G., 2008. On the empirical content of quantal response equilibrium. Amer Econ Review 98 (1), 180–200.

Harsanyi, J. C., Dec 1973. Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points. International Journal of Game Theory 2 (1), 1–23.

Harstad, R. M., Dec 2000. Dominant strategy adoption and bidders' experience with pricing rules. Experimental Economics 3 (3), 261–280.

Hassidim, A., Romm, A., Shorrer, R. I., 2020. The limits of incentives in economic matching procedures. Management Sci 67 (2).

Healy, P. J., 2006. Learning dynamics for mechanism design: An experimental comparison of public goods mechanisms. J Econ Theory 129 (1), 114 – 149.

Jackson, M. O., 1991. Bayesian implementation. Econometrica 59 (2), 461–477.

Jackson, M. O., 1992. Implementation in undominated strategies: A look at bounded mechanisms. The Review of Economic Studies 59 (4), 757–775.

Kagel, J. H., Harstad, R. M., Levin, D., 1987. Information impact and allocation rules in auctions with affiliated private values: A laboratory study. Econometrica 55 (6), 1275–1304.

Kagel, J. H., Levin, D., 1993. Independent private value auctions: Bidder behaviour in first-, second- and third-price auctions with varying numbers of bidders. The Economic Journal 103 (419), 868–879.

Kawagoe, T., Mori, T., Aug 2001. Can the pivotal mechanism induce truth-telling? an experimental study. Public Choice 108 (3), 331–354.

Kneeland, T., 2022. Mechanism design with level-k types: theory and applications to bilateral trade, forthcoming J Econ Theory.

Kohlberg, E., Mertens, J.-F., 1986. On the strategic stability of equilibria. Econometrica 54 (5), 1003–1037.

Li, S., November 2017. Obviously strategy-proof mechanisms. Amer Econ Review 107 (11), 3257–87.

Mackenzie, A., 2020. A revelation principle for obviously strategy-proof implementation. Games Econ Behavior 124, 512–533.

Masuda, T., Mikami, R., Sakai, T., Serizaway, S., Wakayama, T., 2022. The net effect of advice on strategy-proof mechanisms: an experiment for the vickrey auction. Exp Econ.

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

McKelvey, R. D., Palfrey, T. R., 1996. A statistcial theory of equilibrium in games. Japanese Econ Review 47 (2), 186–209.

Melo, E., Pogorelskiy, K., Shum, M., 2019. Testing the quantal response hypothesis. International Economic Review 60 (1), 53–74.

Milgrom, P., Mollner, J., 2018. Equilibrium selection in auctions and high stakes games. Econometrica 86 (1), 219–261.

Milgrom, P., Mollner, J., 2021. Extended proper equilibrium. Journal of Economic Theory 194, 105258.

Moulin, H., 1980. On strategy-proofness and single peakedness. Public Choice 35 (4), 437–455.

Myerson, R. B., Jun 1978. Refinements of the nash equilibrium concept. International Journal of Game Theory 7 (2), 73–80.

Nachbar, J. H., Mar 1990. “evolutionary” selection dynamics in games: Convergence and limit properties. Int J Game Theory 19 (1), 59–89.

Plott, C. R., Zeiler, K., 2005. The willingness to pay-willingness to accept gap, the” endowment effect,” subject misconceptions, and experimental procedures for eliciting valuations. American Economic Review 95 (3), 530–545.

Pycia, M., Troyan, P., 2022. A theory of simplicity in games and mechanism design, mimeo.

Rees-Jones, A., 2017. sub-optimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ Behavior.

Repullo, R., 1985. Implementation in dominant strategies under complete and incomplete information. Review Econ Studies 52 (2), 223–229.

Rosenthal, R. W., Sep 1989. A bounded-rationality approach to the study of noncooperative games. Int J Game Theory 18 (3), 273–292.

Roth, A. E., 1984. The evolution of the labor market for medical interns and residents: A case study in game theory. J Political Econ 92 (6), 991–1016.

Saijo, T., Sjöström, T., Yamato, T., 2007. Secure implementation. Theor Econ 2 (3), 203–229.

Samuelson, L., 1992. Dominated strategies and common knowledge. Games Econ Behavior 4 (2), 284 – 313.

Satterthwaite, M. A., 1975. Strategy-proofness and arrow’s conditions: Existence and correspondence theorems for voting procedures and social welfare functions. J Econ Theory 10 (2), 187 – 217.

Satterthwaite, M. A., Sonnenschein, H., 1981. Strategy-proof allocation mechanisms at differentiable points. Review Econ Studies 48 (4), 587–597.

Selten, R., Mar 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. International Journal of Game Theory 4 (1), 25–55.

Shapley, L., Scarf, H., 1974. On cores and indivisibility. J Math Econ 1 (1), 23 – 37.

Shorrer, R., Sóvágó, S., 2022. Dominated choices in a strategically simple college admissions environment: The effect of admission selectivity, mimeo.

Smith, V. L., 1994. Economics in the laboratory. Journal of Economic Perspectives 8 (1), 113–131.

Sprumont, Y., 1983. The division problem with single-peaked preferences: A characterization of the uniform allocation rule. Econometrica 51, 939–954.

Thomson, W., Oct 2016. Non-bossiness. Soc Choice Welfare 47 (3), 665–696.

Troyan, P., 2019. Obviously strategy-proof implementation of top trading cycles. International Economic Review 60 (3), 1249–1261.

Tumennasan, N., 2013. To err is human: Implementation in quantal response equilibria. Games and Economic Behavior 77 (1), 138 – 152.

van Damme, E., 1991. Stability and Perfection of Nash Equilibria. Springer Berlin Heidelberg, Berlin, Heidelberg.

Velez, R. A., Brown, A. L., 2020a. Empirical bias of extreme-price auctions: analysis.
URL http://arxiv.org/abs/1905.08234

Velez, R. A., Brown, A. L., 2020b. Empirical equilibrium.
URL https://arxiv.org/abs/1804.07986

Appendix not for publication

Empirical strategy-proofness

Rodrigo A. Velez and Alexander L. Brown
Texas A&M University

May, 2022

Student Proposing Deferred Acceptance rule (SPDA) has essentially unique dominant strategies

The following discussion uses the standard language in school choice problems (c.f. Abdulkadiroğlu and Sönmez, 2003). Suppose that preferences are strict and starting from a profile in which student i is truthful, she changes her report but does not change the relative ranking of her assignment with respect to the other assignments. The SPDA assignment for the first profile, say m, is again stable for the second profile. Thus, for the new profile, each other agent is weakly better off. Agent i's allotment is the same in both markets because SPDA is strategy-proof. If another agent changes her allotment, it is because the new SPDA assignment was blocked in the original profile. Since the preferences of the other agents did not change, agent i needs to be in the blocking pair for the new assignment in the original market. However, this means she is in a blocking pair for the new assignment in the new market. Thus, with this type of lie, agent i cannot change the allotment of anybody else. If agent i changes the relative ranking of her allotment in the original market, she can be worse off with the lie. For instance, suppose that she moves m_j from her lower contour set at her allotment to the upper contour set. In the preference profile in which each agent different from i and j ranks top her allotment at m, and in which agent j ranks m_i top, agent i receives m_j in the SPDA assignment.

Robust Nash implementation

Proof of Theorem 3. We first prove that $1 \Rightarrow 4$. Suppose that Statement 1 is satisfied. Our argument in the proof of Lemma 4, taking σ as an equilibrium of (M, φ, p) for the interior p defined there, implies that g is strategy proof. We now prove that g is non-bossy and satisfies the outcome rectangular property. Our proof follows closely that of Adachi (Proposition 3, 2014). By Saijo et al. (Proposition 3, 2007), it is enough to prove that for each pair $\{\theta, \theta'\} \subseteq \Theta$, if for each $i \in N$, $u_i(g(\theta')|\theta_i) = u_i(g(\theta'_{-i}, \theta_i)|\theta_i)$, then $g(\theta) = g(\theta')$. Thus, let $\{\theta, \theta'\} \subseteq \Theta$, and suppose that for each $i \in N$,

$$u_i(g(\theta')|\theta_i) = u_i(g(\theta'_{-i}, \theta_i)|\theta_i). \tag{A.1}$$

Consider a prior p that places uniform probability on the set $\{(\theta'_{-i}, \mu_i) : i \in N, \mu_i \in \{\theta_i, \theta'_i\}\}$. Let σ be an equilibrium of (M, φ, p), which always exists because the mechanism is finite. Let $i \in N$, m_i in the support of $\sigma_i(\cdot|\theta_i)$, m'_i in the support of $\sigma_i(\cdot|\theta'_i)$, and $\hat{m}_{-i}$ in the support of $\sigma_{-i}(\cdot|\theta'_{-i})$. By Statement 1,

$$\varphi(\hat{m}_{-i}, m'_i) = g(\theta') \text{ and } \varphi(\hat{m}_{-i}, m_i) = g(\theta'_{-i}, \theta_i). \tag{A.2}$$

Thus, by (A.1),

$$\sum_{\hat{m}_{-i} \in M_{-i}} u_i(\varphi(\hat{m}_{-i}, m_i)|\theta_i)\sigma_{-i}(\hat{m}_{-i}|\theta'_{-i}) = \sum_{\hat{m}_{-i} \in M_{-i}} u_i(\varphi(\hat{m}_{-i}, m'_i)|\theta_i)\sigma_{-i}(\hat{m}_{-i}|\theta'_{-i}).$$

Since agent i knows the type of the other agents is θ'_{-i} when she draws type θ_i, equilibrium behavior implies that for each $\hat{m}_i \in M_i$,

$$\sum_{\hat{m}_{-i} \in M_{-i}} u_i(\varphi(\hat{m}_{-i}, m_i)|\theta_i)\sigma_{-i}(\hat{m}_{-i}|\theta'_{-i}) \geq \sum_{\hat{m}_{-i} \in M_{-i}} u_i(\varphi(\hat{m}_{-i}, \hat{m}_i)|\theta_i)\sigma_{-i}(\hat{m}_{-i}|\theta'_{-i}).$$

By the last two displayed equations, for each $\hat{m}_i \in M_i$,

$$\sum_{\hat{m}_{-i} \in M_{-i}} u_i(\varphi(\hat{m}_{-i}, m'_i)|\theta_i)\sigma_{-i}(\hat{m}_{-i}|\theta'_{-i}) \geq \sum_{\hat{m}_{-i} \in M_{-i}} u_i(\varphi(\hat{m}_{-i}, \hat{m}_i)|\theta_i)\sigma_{-i}(\hat{m}_{-i}|\theta'_{-i}).$$

Thus, if μ is a behavior strategy such that $\mu(\cdot|\theta) = \sigma(\cdot|\theta')$, for each $\hat{m}_i \in M_i$,

$$\sum_{\hat{m}_{-i} \in M_{-i}} u_i(\varphi(\hat{m}_{-i}, m'_i)|\theta_i)\mu_{-i}(\hat{m}_{-i}|\theta_{-i}) \geq \sum_{\hat{m}_{-i} \in M_{-i}} u_i(\varphi(\hat{m}_{-i}, \hat{m}_i)|\theta_i)\mu_{-i}(\hat{m}_{-i}|\theta_{-i}).$$

Thus, μ is a Nash equilibrium of (M, φ, θ). By Statement 1, $\varphi(m') = g(\theta)$. Thus, $g(\theta) = g(\theta')$.

The argument that $2 \Rightarrow 4$ in Saijo et al. (2007) applies to our environment unmodified. This argument involves only dominant strategy and pure strategy equilibria in complete information structures.

We now prove that $4 \Rightarrow 3$. Suppose that 4 is satisfied. Let σ be an equilibrium of (Θ, g, p) for some common prior p. Let θ in the support of p and τ be in the support of $\sigma(\cdot|\theta)$. Observe since σ is an equilibrium, Statement 4 implies that Assumptions (i) and (ii) in the proof of Lemma 2 are satisfied. Thus, for each $i \in N$, $g(\tau_{-i}, \theta_i) = g(\tau)$. Then, by the outcome rectangular property, we have that $g(\tau) = g(\theta)$.

Finally, we observe that trivially $3 \Rightarrow 1$, and by Lemma 4, 3 implies g is strategy-proof. Thus, $3 \Rightarrow 2$.

□

Social choice correspondences

We now show that our results depend on our restriction to social choice functions. That is, our requirement that the social planner's objective be summarized on a function that selects a unique determinate outcome for each social state. Since mixed strategy equilibria are essential in our analysis, a generalization of our model requires that we first reconsider the role of mixed strategies in Bayesian implementation. Indeed, in some environments, almost all pure strategy equilibria of a mechanism may be completely wiped out by the empirical equilibrium refinement, while a continuum of mixed strategy equilibria survive (Velez and Brown, 2020a).

An alternative that we find appealing as a starting point is to study typical Bayesian implementation (Jackson, 1991) in a finitely generated model in which the social planner selects probability measures on outcomes for each social state. More precisely, for a finite outcome space X let Θ be a payoff type space as defined in our model. A (random) social choice function associates with each type profile a probability distribution on X, i.e., $g : \Theta \to \Delta(X)$. A mechanism (M, φ) is defined as usual, but allowing for randomization, i.e., $\varphi : M \to \Delta(X)$. A (random) social choice set G is a subset of social choice functions. Then one can determine the success of a mechanism from the point of view of a mechanism designer who identifies G as desirable by comparing the equilibria of (M, φ, p) with the elements of G.

The following example shows that strategy-proofness is not necessary to obtain a meaningful form of robust implementation in empirical equilibrium when one allows for multi-valued objectives. That is, one can construct a finite X and a payoff-type space Θ that admits a social choice set G that contains no strategy-proof scf and for which there is a finite mechanism (M, φ) such that for each common prior p and each empirical equilibrium of (M, φ, p), say σ, there is an element of G that coincides with the induced conditional measures $\theta \mapsto \varphi(\sigma(\cdot|\theta))$ in the support of p.

Example 1. Consider the following modification of Bergemann and Morris (2005, Example 2): $\Theta_1 := \{\theta_1, \theta_1', \theta_1''\}$, $\Theta_2 := \{\theta_2, \theta_2'\}$, $X := \Delta(\{a, b, c, d, a', b', c', d'\})$,

u_1	a	b	c	d	a'	b'	c'	d'
θ_1	1	-1	$1/2-\varepsilon$	-1	-1	1	-1	$1/2-\varepsilon$
θ_1'	0	0	1	0	0	0	1	0
θ_1''	0	0	0	1	0	0	0	1

and

u_2	a	b	c	d	a'	b'	c'	d'
θ_2	ε	1	0	0	0	$1-\varepsilon$	-1	-1
θ_2'	$1-\varepsilon$	0	-1	-1	1	ε	0	0

Let F be the correspondence that assigns to each type profile the set of probability distributions on outcomes in the following table.

	θ_2	θ_2'
θ_1	$\Delta(\{a,b\})$	$\Delta(\{a',b'\})$
θ_1'	$\{c\}$	$\{c'\}$
θ_1''	$\{d\}$	$\{d'\}$

Let G be the social choice set of all scfs g such that for each θ, $g(\theta) \in F(\theta)$.

An argument as that in Bergemann and Morris (2005) shows that if $\varepsilon < (9-\sqrt{65})/8$, there is no strategy-proof scf g such that for each $\theta \in \Theta$, $g(\theta) \in F(\theta)$. Thus, there is no strategy-proof scf in G.

Finally, let (M,φ) be the mechanism where $M_1 := \{m_1^1, m_1^2, m_1^3, m_1^4\}$, $M_2 := \{m_2^1, m_2^2\}$, and φ is given by:

	m_1^1	m_1^2	m_1^3	m_1^4
m_2^1	a	b	c	d
m_2^2	a'	b'	c'	d'

Consider a common prior p. Observe that m_2^1 is strictly dominant for payoff type θ_2 and m_2^2 is strictly dominant for payoff type θ_2'. Thus, in each Nash equilibrium of (M,φ,p) these payoff types play these strategies with probability one. Now, consider agent 1 with type θ_1. Clearly, m_1^1 weakly dominates m_1^3 and m_1^2 weakly dominates m_1^4. Moreover, if the expected value of m_1^1 is the same as that for m_1^3, we have that the expected value of m_1^2 is greater than that of m_1^4. Thus, agent 1 with type θ_1 will never play m_3^1 nor m_1^4 in an equilibrium of (M,φ,p). Note also that agent 1 with types θ_1' and θ_1'' has strictly dominant actions m_1^3 and m_1^4, respectively. Thus, for each p, each empirical equilibrium of (M,φ,p), say σ, and each realization of payoff types $\theta \in \Theta$, $\sigma(\cdot|\theta)$ induces a measure on X that belongs to $F(\theta)$. □

Summary of dominant strategy play

scf	% Dominant Strategy	no. of available pure strategies	% Dominant if strategies played at random	do payoffs of played strategies exceed non-played?[1]	Description/Source
2nd-Price Auction	50.0	50 (mean)	2.0	N.A.	- 6 sessions with number of rounds from 10 to 24; Coppinger et al. (1980, Table 8).
	27.0, 32.5	> 2830	< 0.4	N.A.	- Two sessions with 24 and 35 rounds; totals for experiments with groups of 5 and 10 agents respectively; dominant strategies classified as +/-0.05 from true value.;Kagel and Levin (1993, Table 2).
	68.2, 57.5, 51.2	201	0.5	Y, Y, Y	- 30 rounds; totals correspond to incomplete info, partial info, and perfect info, respectively; four-agent groups randomly drawn each period; Andreoni et al. (2007).* In the referenced paper, dominant strategies are classified as +/-0.01 from true value, producing slightly different numbers.
	44.5	1,000,000	0.0	N.A.	- 20 rounds; Percentages pooled over all sessions with different information; two-agent groups randomly drawn each period; Cooper and Fang (2008).
	17.8	601	0.2	Y	- 10 rounds; four-agent groups; Li (2017).*
+X Variant	20.4	601	0.2	Y	- 10 rounds; four-agent groups; Li (2017).*
Pivotal	10.5, 8.25	2001	0.0	Y, Y	- 10 rounds; totals for experiments with groups of 5 and 10 agents respectively; 2001 actions available to each agent; Attiyeh et al. (2000).*
	17, 14, 47	51	2.0	N.A.	- 10 rounds; total for experiments with three alternative description of mechanism; Kawagoe and Mori (2001).
	73.3	25	4.0	Y	- 8 to 10 rounds; two-agent groups; each agent has two weakly dominant actions in each game; Cason et al. (2006).*
cVCG	54	$> 505,000$	0.0	N.A.	- Public good provision with quasi-linear preferences; utility for public good has two parameters; unbounded reports; 4 sessions of 50 rounds (Healy, 2006).
Student Optimal Deferred Acceptance	72.2, 50	5040	0.0	N.A.	- 1 round; totals for uniformly random and correlated priority structures; Chen and Sönmez (2006).
Top Trading Cycles	55.6, 43.1	5040	0.0	N.A.	- 1 round; totals for uniformly random and correlated priority structures; Chen and Sönmez (2006).
Random Serial Priority	71.0	24	4.2	Y	- 10 rounds; four-agent groups; Li (2017).*

Table A.1: Frequency of dominant strategy play in strategy-proof mechanisms; *denotes statistics calculated directly from data, not reported by authors.

[1] None of the "Y"s in the table would change if this analysis were performed excluding any decisions where subjects chose the dominant strategy.

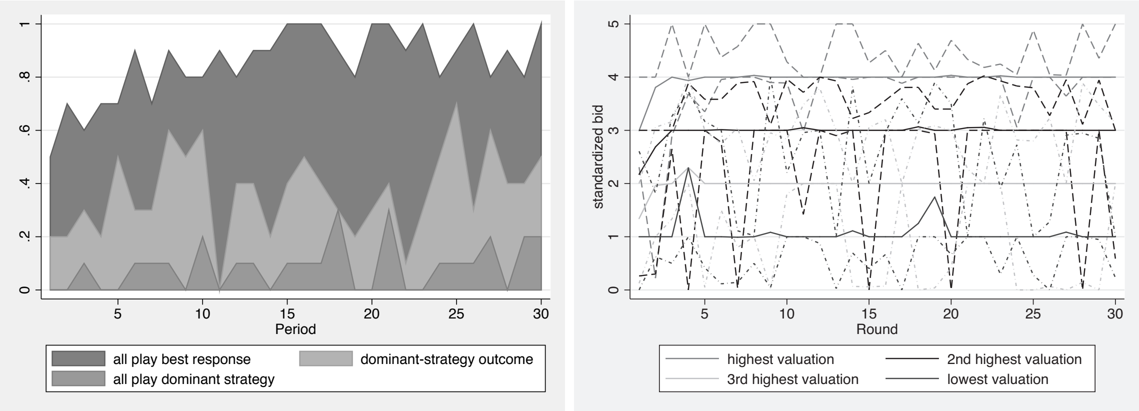

Figure A.1: Second-price auction in Andreoni et al. (2007) under complete information. (left) The dark gray area indicates the proportion of outcomes where all subjects play mutual best responses to the actions of all other group members. The light gray area indicates outcomes where the transaction associated with the dominant strategy outcome occurs, that is, the subject with the highest valuation obtains the item and pays the amount of the second-highest valuation. The medium gray area indicates the percentage of group outcomes where all subjects play a dominant strategy. Note that each level necessarily contains the subsequent level. Subjects are rematched randomly, in four-agent groups, across a group of 20 each period. (Right) Median bid and 15th-85th percentile range by valuation type. Bids are standardized so that the valuation of the 1st-4th valuations in the specific auction are assigned values 4–1, respectively. Bids of 100 (the highest possible valuation) and 200 (the highest possible bid) are assigned values of 5 and 6, respectively. If two valuation types have the same value, valuation order is randomly assigned. Bids between two valuations are standardized by $(bid - valuation_j)/(valuation_i - valuation_j)$ where i is the highest valuation a bid exceeds and j is the next highest valuation. Bids below the lowest valuation are standardized on the interval between 0 and the lowest valuation. Bids above the highest valuation are standardized either on the interval between the highest valuation and 100 (values of 4–5), or 100 and 200 (values of 5–6). For example, for the four valuations 80, 40, 25, 10, bids of 150, 40, 30, and 5 would be 5.5, 3, 2.33, and 0.5, respectively.

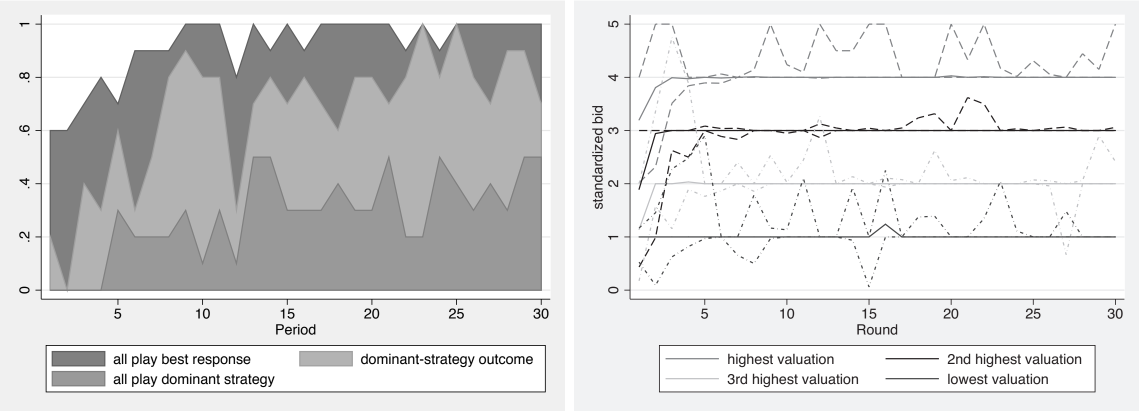

Figure A.2: Second-price auction experiment in Andreoni et al. (2007) under interior incomplete information; (left) percentages of best responses, dominant strategy-outcomes and dominant strategy play; (Right) normalized median, 15th and 85th percentiles, for bids (same normalization as in Fig. A.1)

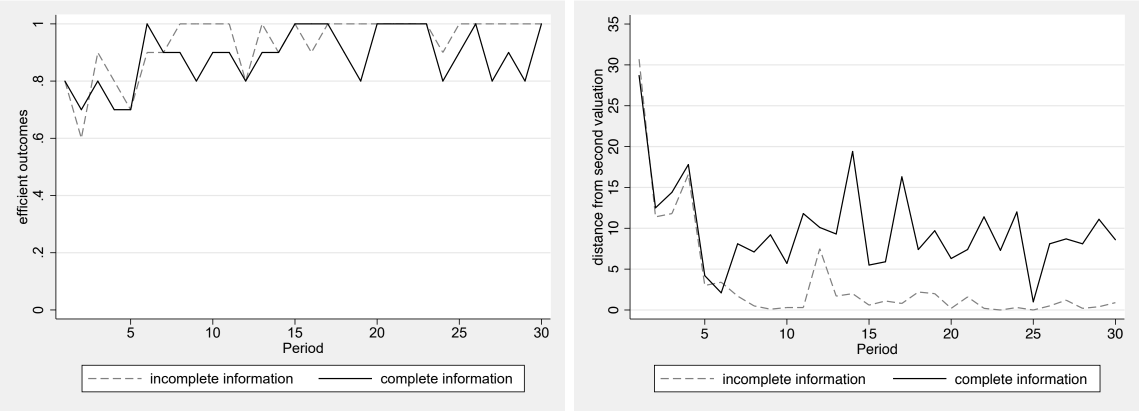

Figure A.3: Frequency of efficient outcomes (left) and average distance (conditional on efficient outcome) between the price and a second valuation (right) in the second-price auction experiments of Andreoni et al. (2007) in the interior incomplete and complete information treatments.