Empirical strategy-proofness*

Rodrigo A. Velez† and Alexander L. Brown‡

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

January 28, 2020

Abstract

We study the plausibility of sub-optimal Nash equilibria of the direct revelation mechanism associated with a strategy-proof social choice function. By using the recently introduced empirical equilibrium analysis (Velez and Brown, 2019b), we determine that this behavior is plausible only when the social choice function violates a non-bossiness condition and information is not interior. Analysis of the accumulated experimental and empirical evidence on these games supports our findings.

JEL classification: C72, D47, D91.

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

1 Introduction

Strategy proofness, a coveted property in market design, requires that truthful reports be dominant strategies in the simultaneous direct revelation game associated with a social choice function (scf). Despite the theoretical appeal of this property, experimental and empirical evidence suggests that when an scf satisfying this property is operated, agents may persistently exhibit weakly dominated behavior (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; Andreoni et al., 2007; Hassidim et al., 2016; Rees-Jones, 2017; Li, 2017; Artemov et al., 2017; Chen and Pereyra, 2018). In this paper we study the plausibility of

*Thanks to James Andreoni, Antonio Cabrales, Yeon-Koo Che, Cary Deck, Huiyi Guo, Utku Unver and seminar participants in Boston College, NC State U., Ohio State U., UCSD, UT Dallas, SAET19, 7th TETC, and North American Meetings ESA 2019, 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

Nash equilibria of the direct revelation game of strategy-proof scfs. By doing so we identify the circumstances in which empirical distributions of play in these games may persistently exhibit weakly dominated actions that approximate a Nash equilibrium that produces sub-optimal outcomes.[1]

The conventional wisdom on plausibility of Nash equilibria offers no explanation on why weakly dominated behavior can be persistent in some dominant strategy games. Indeed, the most prominent theories either implicitly or explicitly assume that this behavior is not plausible (from the seminal tremble-based refinements of Selten (1975) and Myerson (1978), to their most recent forms in Milgrom and Mollner (2017, 2018) and Fudenberg and He (2018); see also Kohlberg and Mertens (1986) and van Damme (1991) for a survey up to the late 80's where this literature was most active).

In Velez and Brown (2019b) we attack the problem of plausibility of Nash equilibria with an alternative approach based on the following thought experiment. We imagine that we sample behavior in the game of our interest and construct a model of unobservables that explains the observed behavior.[2] For instance, we construct a randomly disturbed payoff model (Harsanyi, 1973; van Damme, 1991), a control cost model (van Damme, 1991), a structural QRE model (McKelvey and Palfrey, 1995), a regular QRE model (McKelvey and Palfrey, 1996; Goeree et al., 2005), etc. In order to bring our model to accepted standards of science we need to make sure it is *falsifiable*. We observe that in the most popular models for the analysis of experimental data, including the ones just mentioned, this has been done by requiring consistency with an a priori observable restriction for which there is empirical support, *weak payoff monotonicity*.[3] This property of the full profile of empirical distributions of play in a game requires that for each agent, differences in behavior 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. Finally, we proceed with our study and define a refinement of Nash equilibrium by means of "approachability" by behavior in our model à la Harsanyi (1973), van Damme (1991), and McKelvey and Palfrey (1996). That is, we label as implausible the Nash equilibria of our game that are not the limit of a sequence of behavior that can be generated by our model (in the whole range

[1] Given an scf we refer to a sub-optimal outcome as one that is different from the one intended by the scf for the true characteristics of the agents.

[2] Our benchmark is an experimental environment in which the researcher observes payoffs and samples frequencies of play. This observable payoffs framework is also a valuable benchmark for the foundation of Nash equilibrium (Harsanyi, 1973).

[3] Harsanyi (1973) does not explicitly impose weak payoff monotonicity in his randomly perturbed payoff models. The objective of his study is to show that certain properties hold for all randomly perturbed payoff models with vanishing perturbations for generic games. This makes it unnecessary to discipline the model with a priori restrictions. Requiring permutation invariance on Harsanyi (1973)'s models induces weak payoff monotonicity (van Damme, 1991).

in which unobservables are defined). If our model is well-specified, the equilibria that are ruled implausible by our refinement, will never be approached by observed behavior even when distributions of play approach mutual best responses.[4] Of course, we are not sure what the true model is. Our thought experiment was already fruitful, however. We learned that if we were able to construct the true model and our a priori restriction does not hinder its specification, the Nash equilibria that we would identify as implausible will necessarily contain those in the complement of the closure of weakly payoff monotone behavior. This leads us to the definition of *empirical equilibrium*, a Nash equilibrium for which there is a sequence of weakly payoff monotone distributions of play converging to it. The complement of this refinement (in the Nash equilibrium set), the *empirically implausible equilibria*, are the Nash equilibria that are determined implausible by each theory that is disciplined by weak payoff monotonicity.

We can considerably advance our understanding of the direct revelation game of a strategy-proof scf by calculating its empirical equilibria. On the one hand, suppose that we find that for a certain game each empirical equilibrium is truthful equivalent. Then, we learn that as long as empirical distributions of play are weakly payoff monotone, behavior will never approximate a sub-optimal Nash equilibrium. On the other hand, if we find that some empirical equilibria are not truthful equivalent, this alerts us about the possibility that we may plausibly observe persistent behavior that generates sub-optimal outcomes and approximates mutual best responses.

We present two main results. The first is that non-bosiness in welfare-outcome—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). The second is that the requirement that a strategy-proof scf have essentially unique dominant strategies, characterizes this form of robust implementation for type spaces with full support (Theorem 2). The sharp predictions of our theorems are consistent with experimental and empirical evidence on strategy-proof mechanisms (Sec. 6). 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 whose dominant strategies are unique. Deviations from truthful behavior are persistently observed when this mechanism is operated, but mainly for information structures for which agents' types are common information (Andreoni et al., 2007).

[4] We have in mind an unmodeled evolutionary process by which behavior approaches a Nash equilibrium. Thus, we are essentially interested in the situations in which eventually a game form is a good approximation of the strategic situation we model, as when perturbations vanish in Harsanyi (1973)'s approachability theory.

The remainder of the paper is organized as follows. Section 2 places our contribution in the context of the literature. Section 3 presents the intuition of our results illustrated for the Top Trading Cycles (TTC) mechanism and the second-price auction, two cornerstones of the market design literature. Section 4 introduces the model. Section 5 presents our main results. Section 6 contrasts our results with experimental and empirical evidence. Section 7 contrasts them with the characterizations of robust full implementation (Bergemann and Morris, 2011; Saijo et al., 2007; Adachi, 2014) with which one can draw an informative parallel, and discusses our restriction to direct revelation mechanisms. Section 8 concludes. The Appendix collects all proofs.

2 Related literature

The literature on strategy-proof mechanisms was initiated by Gibbard (1973) and Satterthwaite (1975) who proved that this property implies dictatorship when there are at least three outcomes and preferences are unrestricted. The theoretical literature that followed has shown that this property is also restrictive in economic environments, but can be achieved by reasonable scfs in restricted preference domains (see Barbera, 2010, for a survey). Among these are the VCG mechanisms for the choice of an outcome with transferable utility, which include the second-price auction of an object and the Pivotal mechanism for the selection of a public project (see Green and Laffont, 1977, and references therein); the TTC mechanism for the reallocation of indivisible goods (Shapley and Scarf, 1974); the Student Proposing Deferred Acceptance (SPDA) mechanism for the allocation of school seats based on priorities (Gale and Shapley, 1962; Abdulkadiroğlu and Sönmez, 2003); the median voting rules for the selection of an outcome in an interval with satiable preferences (Moulin, 1980); and the Uniform rule in the rationing of a good with satiable preferences (Benassy, 1982; Sprumont, 1983). Even though Gibbard (1973) is not convinced about the positive content of dominant strategy equilibrium, the theoretical literature that followed endorsed the view that strategy-proofness was providing a bulletproof form of implementation. Thus, when economics experiments were developed and gained popularity in the 1980s, the dominant strategy hypothesis became the center of attention of the experimental studies of strategy-proof mechanisms. Until recently the accepted wisdom was that behavior in a game with dominant strategies should be evaluated with respect to the benchmark of the dominant strategies hypothesis. The common finding in these experimental studies is a lack of support for this hypothesis (Sec. 6).[5]

[5]In a recently circulated paper, Masuda et al. (2019) present evidence that the rate of truthful reports in a second-price auction increases when agents are directly advised about the dominance strategy property of these reports (from 20% to 47%).

Economic theorists have been slowly reacting to the findings in laboratory experiments. The first attempt was made by Saijo et al. (2007) who looked for scfs that are implementable simultaneously in dominant strategies and Nash equilibrium in complete information environments. This form of implementation has a robustness property under multiple forms of incomplete information (Saijo et al., 2007; Adachi, 2014). Experiments have confirmed to some extent that their additional requirements actually improve the performance of mechanisms (Cason et al., 2006). More recently, Li (2017) looked for additional properties of mechanisms that induce agents to choose a dominant strategy in a dominant strategy game. Our study differs from Saijo et al. (2007) and Li (2017) in a similar form. We do not provide conditions guaranteeing that behavior will indeed quickly converge to a truthful equivalent Nash equilibrium. We characterize conditions that guarantee behavior will not accumulate around a sub-optimal Nash equilibrium. It is good news when we find an scf satisfies these properties. This means that an agent's sub-optimal choices will always continue to be disciplined by the choices of the other agents. A growing literature is now showing us that the higher aims of Saijo et al. (2007) and Li (2017) lead us to come up empty handed in many problems of interest (c.f. Bochet and Sakai, 2010; Fujinaka and Wakayama, 2011; Bade and Gonczarowski, 2017). Thus, studies like ours, which produce a better understanding of the incentives across all strategy-proof mechanisms, have significant value in face of these imposibilities.

Since its inception in private consumption environments by Satterthwaite and Sonnenschein (1981), non-bossiness has played a role in social choice literature. Essentially, this property fills a gap between axioms of individual and collective behavior (e.g., Barberà et al., 2016).[6] Our work differers from the social choice literature in that the, so to speak, left side of our characterizations, is a requirement on mechanisms based on a testable property of behavior, not on a property that one argues in favor of based on its normative content. In this sense our results are the first to establish a link between a non-bossiness condition and the empirical content of the Nash equilibrium prediction for the direct revelation mechanism of a strategy-proof scf.

Strategy-proof mechanisms have been operated for some time in the field. Empirical studies of such mechanisms have generally corraborated the observations from laboratory experiments (e.g. Hassidim et al., 2016; Rees-Jones, 2017; Artemov et al., 2017; Chen and

[6]See Thomson (2016) for a survey of the definition and the normative content of the different notions of non-bossiness that have been used in the social choice literature. The particular form of non-bossiness that emerges endogenously from our characterization has played a role in at least two previous studies. Bochet and Tumennassan (2017) find that non-bossiness in welfare-outcome is a necessary and sufficient condition for a strategy-proof game to have only truthful equivalent equilibria in complete information environments when truthful behavior is focal. Schummer and Velez (2019) show that non-bossiness in welfare-outcome is sufficient for a deterministic sequential direct revelation game associated with a strategy-proof scf to implement the scf itself in sequential equilibria for almost every prior.

Pereyra, 2018; Shorrer and Sóvágó, 2019), in such high stakes environments as career choice (Roth, 1984) and school choice (Abdulkadiroğlu and Sönmez, 2003). Among these papers, Artemov et al. (2017) and Chen and Pereyra (2018), are the closest to ours. Besides presenting empirical evidence of persistent violations of the dominant strategies hypothesis, they propose theoretical explanations for it. They restrict to school choice environments in which a particular mechanism is used. Artemov et al. (2017) study a continuum model in which the SPDA mechanism is operated in a full-support incomplete information environment. They conclude that it is reasonable that one can observe equilibria in which agents make inconsequential mistakes. 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 (2018) study a finite school choice environment in which there is a unique ranking of students across all schools. They argue that only when information is not interior an agent can be expected to deviate from her truthful report based on the analysis of an ordinal form of equilibrium. Our study substantially differs in its scope with these two papers, because our results apply to all strategy-proof scfs. When applied to a school choice problem, our results are qualitatively in line with those in these two studies and thus provide a rational 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.

Our work can be related with a 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. They identify properties of mechanism that guarantee their convergence to desired allocations under certain dynamics. These conditions turn out to be strong. Indeed, they are violated by all the strategy-proof mechanisms we have mentioned above. Thus, their results do not inform us about the nature of arbitrary strategy-proof mechanisms. 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., 2017; Kneeland, 2017). Our work bridges these two approaches. It informs us about the performance of mechanisms when behavior approximates mutual best responses by some unrestricted process and at the same time satisfies a weak and testable form of rationality. We do pay a price for not modeling the convergence process, for our work is silent about the probability to observe approximate mutual best responses when a

mechanisms is operated. This is well justified, however. We can say with confidence that the conditions we identify guarantee a mechanism is unlikely to ever be documented consistently producing sub-optimal outcomes while agents do not have a significant incentive to change their behavior.

Finally, this paper is a part of the empirical equilibrium agenda, which consists on reevaluating game theory applications with the empirical equilibrium refinement. In Velez and Brown (2019b) we define empirical equilibrium and provide a foundation to it by means of the regular QRE model of McKelvey and Palfrey (1996) and Goeree et al. (2005). In Velez and Brown (2019c) we study the relationship of empirical equilibrium and the refinements obtained by means of approximation by the separable randomly perturbed payoff models of Harsanyi (1973) and McKelvey and Palfrey (1995). In Velez and Brown (2019a) we apply empirical equilibrium to the problem of full implementation. Finally, in Brown and Velez (2019) we test the comparative statics predicted by empirical equilibrium in a partnership dissolution problem in which dominant strategy mechanisms are not available.

3 The intuition: empirical plausibility of equilibria of TTC and second-price auction

Two mechanisms illustrate our main findings. The first is TTC for the reallocation of indivisible goods from individual endowments (Shapley and Scarf, 1974). The second is the popular second-price auction. For simplicity, let us consider two-agent stylized versions of these market design environments.

Suppose that two agents, say $\{A, B\}$, are to potentially trade the houses they own when each agent has strict preferences. TTC is the mechanism that operates as follows. Each agent is asked to point to the house that he or she prefers. Then, they trade if each agent points to the other agent's house and remain in their houses otherwise. It is well known that this mechanism is strategy-proof. That is, 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 an efficient trade. There are more Nash equilibria of the game that ensues when this mechanism is operated. Consider the strategy profile where each agent unconditionally points to his or her own house, regardless of information structure. This profile of strategies provides mutual best responses for expected utility maximizing agents, but does not necessarily produce the same outcomes as the truthful profile.

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 $\{A, B\}$ who may have a type $\theta_i \in$

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

$\{L, M, H\}$. 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. Then an agent with higher valuation receives the object and pays the seller the valuation of the other agent. Ties are decided uniformly at random. It is well known that this mechanism is also strategy-proof. 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. There are more Nash equilibria of the game that ensue when this mechanism is operated. In order to exhibit such equilibria let us suppose that agent A has type M, agent B has type H, and both agents have complete information of their types. Table 1 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.

Our quest is then to determine which, if any, of the sub-optimal equilibria of TTC, the second-price auction, and for that matter any strategy-proof mechanism, should concern a social planner who operates one of these mechanisms. In order to do so we calculate the empirical equilibria of the games induced by the operation of these mechanisms. It turns out that the Nash equilibria of the TTC and the second-price auction have a very different nature. No sub-optimal 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. This is surprising. 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 with the empirical plausibility of equilibria.

Consider the TTC game and a weakly payoff monotone distribution of play. Since revealing her true preference is dominant, each agent with each possible type will reveal her preferences with probability at least 1/2 in such a strategy. 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. Suppose that information is given by a common prior.[7] 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 1. Fix $\alpha \in [0, 1/2)$. For $\varepsilon > 0$, let $\sigma \equiv (\sigma_A, \sigma_B)$ be the pair of probability distributions on each agent's action space defined as follows: $\sigma_A(H) \equiv \varepsilon$, $\sigma_A(M) \equiv 1 - \alpha - \varepsilon$, $\sigma_A(L) \equiv \alpha$, $\sigma_B(H) \equiv 1 - 3\varepsilon$, $\sigma_B(M) \equiv 2\varepsilon$, $\sigma_B(L) \equiv \varepsilon$. One can easily see that when ε is small, σ is weakly payoff monotone. Indeed, for agent B action H weakly dominates M and this last action weakly dominantes L. Since σ_A is interior, σ_B is ordinally equivalent to the expected utility of actions for agent B given σ_A. Now, for agent A, action M is weakly dominant. Moreover, for small ε, $\sigma_B(H) \approx 1$, thus the expected utility of H for A is strictly less than that of L. Thus, σ_A is ordinally equivalent to the expected utility of actions for agent A given σ_B. Clearly, as $\varepsilon \to 0$, these distributions converge to a Nash equilibrium in which agent B plays H and agent A plays M with probability $1 - \alpha$ and plays L with probability α. Thus, the seller ends up selling for zero price with positive probability for some types whose minimum valuation is positive.

Empirical equilibrium allows us to draw a clear difference between TTC and the second-price auction. 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 in welfare-outcome, i.e., in the direct revelation game of the mechanism, an agent cannot change the outcome without changing her welfare (Theorem 1).

For the strategy-proof mechanisms that do violate non-bossiness, it is useful to examine which information structures produce undesirable empirical equilibria. It turns out that for a strategy-proof mechanism with essentially unique dominant strategies, like the second-price auction, this cannot happen for information structures with full support (Theorem 2).

[7] This can be relaxed to some extent. See Sec. 4.

Thus, in a sense, our example above with the second-price auction actually requires the type of information structure we used.

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. In Sec. 6 we review the relevant experimental and empirical literature and find that these predictions are consistent with it.

4 Model

A group of agents $N \equiv \{1, \ldots, n\}$ is to select an alternative in an arbitrary set 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 $\Theta \equiv \prod_{i \in N} \Theta_i$. 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}. Consistently, whenever convenient, 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 $p \in \Delta(\Theta)$.[8] 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.[9]

A social choice function (scf) selects a set of alternatives 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 (dominant strategy incentive compatible)* 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 in welfare-outcome* 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 *essentially unique dominant strategies* 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 $u_i(g(\tau_{-i}, \theta_i)|\theta_i) > u_i(g(\tau)|\theta_i)$.

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

[9] 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 much simpler and intuitive. 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 freedoms that games with non-common priors allow.

The first property is well-known. The second property 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, among other, by TTC, the Median Voting rule, and the Uniform rule. It is a strengthening of the non-bossiness condition of Satterthwaite and Sonnenschein (1981), which applies only to environments with private consumption. Non-bossiness in welfare-outcome is violated by the Pivotal mechanism, the second-price auction, and SPDA. 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 redundant. This is satisfied whenever true reports are the unique dominant strategies, as in the second-price auction. It is not necessary that dominant strategies be unique for this property to be satisfied. A student in a school choice environment with strict preferences, and in which SPDA is operated, may have multiple dominant strategies (think for instance of a student who is at the top of the ranking of each school). However, any misreport that is also a dominant strategy for this student, cannot change the outcome.[10]

A mechanism is a pair (M, φ) where $M \equiv (M_i)_{i \in N}$ is an unrestricted message space and $\varphi : M \to \Delta(X)$ is an outcome function. A finite mechanism is that for which each M_i is a finite set. Given the common prior p, (M, φ) determines a standard Bayesian game $\Gamma \equiv (M, \varphi, p)$. When the prior is degenerate, i.e., places probability one in 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 (behavior) 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 $\sigma_i(\cdot|\theta_i) \in \Delta(M_i)$.[11] We denote a profile of strategies by $\sigma \equiv (\sigma_i)_{i \in N}$. 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 $\sigma(\cdot|\theta)$.

[10] 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. See Fernandez (2018) for a related property of SPDA that guarantees students do not regret to lie when one also considers possible changes in the priorities of schools.

[11] All of our results refer to finite mechanisms. Thus, we avoid any formalism to account for strategies on infinite sets.

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

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

where the summation is over all $\theta_{-i} \in \theta_{-i}$ and $m \in M$. A profile of strategies σ is a *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 Bayesian Nash equilibria of Γ is $N(\Gamma)$. 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 $m_{-i} \in M_{-i}$, $u_i(m|\theta_i) \geq u_i(m_{-i}, r_i|\theta_i)$.

Our main basis for empirical plausibility of behavior is the following weak form of rationality.

Definition 1 (Velez and Brown, 2019b)**.** A profile of strategies for $\Gamma \equiv (M, \varphi, p)$, $\sigma \equiv (\sigma_i)_{i \in N}$, 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 empirically plausible behavior.

Definition 2 (Velez and Brown, 2019b)**.** An *empirical equilibrium* of $\Gamma \equiv (M, \varphi, p)$ is a Bayesian Nash equilibrium of Γ that is the limit of a sequence of weakly payoff monotone distributions for Γ.

In any finite game, proper equilibria (Myerson, 1978), firm equilibria and approachable equilibria (van Damme, 1991), and the limiting logistic equilibrium (McKelvey and Palfrey, 1995) are empirical equilibria. Thus, existence of empirical equilibrium holds for each finite game (Velez and Brown, 2019b).

5 Results

We start with a key lemma stating that, when available, weakly dominant actions will always be part of the support of each empirical equilibrium in a game.

Lemma 1. Let (M, φ) be a mechanism and p a common prior. Let $i \in N$ and $\theta_i \in \Theta_i$. Suppose that $m_i \in M_i$ is a weakly dominant action for agent i with type θ_i in (M, φ). Let σ be an empirical equilibrium of (M, φ, p). Then, m_i is in the support of $\sigma_i(\cdot|\theta_i)$.

The following theorem characterizes the strategy-proof scfs for which the empirical equilibria of its revelation game produce with certainty, for each common prior information structure, the truthful outcome.

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

1. For each common prior p and each empirical equilibrium of (Θ, g, p), say 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 $\sigma(\cdot|\theta)$, $g(\theta) = g(\tau)$.

2. g is strategy-proof and non-bossy in welfare-outcome.

We now discuss the proof of Theorem 1. Let us discuss first why a strategy-proof and non-bossy in welfare-outcome scf g has the robustness property in statement 1 in the theorem. Suppose that $\sigma \in N(\Theta, g, p)$, that the true type of the agents is θ, and that the agents end up reporting τ with positive probability under σ. Consider an arbitrary agent, say i. Since g is strategy-proof, τ_i can be a best response for agent i with type θ_i only if it gives the agent the same utility as reporting θ_i for each report of the other agents that agent i believes will be observed with positive probability. Thus, since there are rational expectations in a common prior game, report τ_i needs to give agent i the same utility as θ_i when the other agents report τ_{-i}. Since g is non-bossy in welfare-outcome, it has to be the case that $g(\tau_{-i}, \theta_i) = g(\tau)$. By Lemma 1, if σ is an empirical equilibrium of (Θ, g, p), agent i reports her true type with positive probability in σ. Thus, (τ_{-i}, θ_i) is played with positive probability in σ. Thus, we can iterate over the set of agents and conclude that $g(\theta) = g(\tau)$.

Let us discuss now the proof of the converse statement. First, we observe that it is well-known that the type of robust implementation in statement 1 of the theorem implies the scf is strategy-proof (Dasgupta et al., 1979; Bergemann and Morris, 2005). Thus, it is enough to prove that if g is strategy-proof and satisfies the robustness property, it has to be non-bossy in welfare-outcome. Our proof of this statement is by contradiction. We suppose to the contrary that for some type θ, an agent, say i, can change the outcome of g by reporting some alternative τ_i without changing her welfare. We then show that the complete information game (Θ, g, θ) has an empirical equilibrium in which (θ_{-i}, τ_i) is observed with positive probability. The subtlety of doing this resides in that our statement is free of details about the payoff environment in which it applies. We have an arbitrary number of agents and we know little about the structure of agents' preferences. If we

had additional information about the environment, as say for the second-price auction, the construction could be greatly simplified as in our illustrating example.

To solve this problem we design an operator that responds to four different types of signals, $\kappa^{\varepsilon,r,\eta,\lambda} : \Delta(\Theta_1) \times \cdots \times \Delta(\Theta_n) \to \Delta(\Theta_1) \times \cdots \times \Delta(\Theta_n)$, where $\{r, \lambda\} \subseteq \mathbb{N}$ and $\{\varepsilon, \eta\} \subseteq (0, 1)$. This operator has fixed points that are always weakly payoff monotone distributions for u. For a given ε, the operator restricts its search of distributions to those that place at least probability ε in each action for agent i. For a given η, the operator restricts its search of distributions to those that place at least probability η in each action for each agent $j \neq i$. If we take r to infinity, the operator looks for distributions in which agent i's frequency of play is almost a best response to the other agents' distribution (constrained by ε). If we take λ to infinity, the operator looks for distributions in which for each agent $j \neq i$, her frequency of play is almost a best response to the other agents' distribution (constrained by η). The proof is completed by proving that for the right sequence of signals, the operator will have fixed points that in the limit exhibit the required properties. To simplify our discussion without losing the core of the argument, let us suppose that each agent $j \neq i$ has a unique weakly dominant action for each type. Fix ε and r. Since we base the construction of our operator on continuous functions, one can prove that there is $\delta > 0$ such that for each fixed point of the operator, if the expected utility of reports θ_i and τ_i does not differ in more than δ, then agent i places probability almost the same on these two reports. Let $\eta > 0$. If each agent $j \neq i$ approximately places probability η in each action that is not weakly dominant and the rest in her dominant action, the utility of agent i from reports θ_i and τ_i will be almost the same when η is small. Thus, one can calibrate η for this difference to be less than $\delta > 0$. Let $\eta(\varepsilon, r, \delta)$ be this value. If we take λ to infinity keeping $\varepsilon, r, \eta(\varepsilon, r, \delta)$ constant, the distribution of each agent $j \neq i$ in each fixed point of the operator will place, approximately, probability $\eta(\varepsilon, r, \delta)$ in each action that is not weakly dominant. Thus, for large λ, $\kappa^{\varepsilon,r,\eta(\varepsilon,r,\delta),\lambda}$ has a fixed point in which agent i is playing θ_i and τ_i with almost the same probability and all other agents are playing their dominant strategy with almost certainty. We grab one of this distributions. It is the first point in our sequence, which we construct by repeating this argument starting from smaller εs and δs and larger rs.

Interestingly, the conclusions of Theorem 1 depend on our requirement that the empirical equilibria of the scf generate only truthful outcomes for type spaces in which an agent may know, with certainty, the payoff type of the other agents.

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

1. For each full-support prior p and each empirical equilibrium of (Θ, g, p), say σ, we have that for each pair $\{\theta, \tau\} \in \Theta$ where τ is in the support of $\sigma(\cdot|\theta)$, $g(\theta) = g(\tau)$.

2. g is strategy-proof and has essentially unique dominant strategies.

Lemma 1 and Theorems 1 and 2 give us a clear description of the weakly payoff monotone behavior that can be observed when a strategy-proof scf is operated. In the next section we contrast these predictions with experimental evidence on strategy-proof mechanisms.

6 Experimental and empirical evidence

6.1 Dominant strategy play

The performance of strategy-proof mechanisms in an experimental environment has attracted a fair amount of attention. Essentially, experiments have been run to test the hypothesis that dominant strategy equilibrium is a reasonable prediction for these games. The common finding is a lack of support for this hypothesis in most mechanisms. The only exceptions appear to be mechanisms for which dominant strategies are "obvious" (Li, 2017).

Our results provide an alternative theoretical framework from which one can reevaluate these experimental results. Theorems 1 and 2 state that as long as empirical distributions of play are weakly payoff monotone we should expect two features in data. First, we will never see agents' behavior approximate a Nash equilibrium that is not truthful equivalent in two situations: (i) the scf is strategy-proof and non-bossy in welfare-outcome; or (ii) each agent believes all other payoff types are possible and the scf has essentially unique dominant strategies. Second, one cannot rule out that sub-optimal equilibria are approximated by weakly payoff monotone behavior when the scf violates non-bossiness in welfare-outcome and information is complete.

It is informative to note that our first conclusion still holds if we only require, instead of weak payoff monotonicity, that there is a lower bound on the probability with which an agent reports truthfully, an easier hypothesis to test. Thus, in order to investigate whether a sub-optimal Nash equilibrium is approximated in situations (i) and (ii), it is enough to verify that truthful play is non-negligible and does not dissipate in experiments with multiple rounds. This is largely supported by data.

In Table 2, we survey the literature for experimental results with dominant strategy mechanisms. We find ten studies across a variety of mechanisms. In all of these studies we are able to determine, based on the number of pure strategies available to each player, how often a dominant strategy would be played if subjects uniformly played all pure strategies.[12] In every experiment, rates of dominant strategy play exceed this threshold.[13] A simple binomial test—treating each of these nine papers as a single observation—rejects any null

[12] Healy (2006) does not explicitly bound reports. We take as basis the range of submitted reports.
[13] These results are not different if one looks only at initial or late play in the experiments.

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.[14]

It is evident then that the accumulated experimental data supports the conclusion that under conditions (i) and (ii) agents' behavior is not likely to settle on a sub-optimal equilibrium. As long as agents are not choosing a best response, the behavior of the other agents will continue flagging their consequential deviations from truthful behavior as considerably inferior.[15]

6.2 Observing empirical equilibria

Among the experiments we surveyed, Cason et al. (2006), Healy (2006), and Andreoni et al. (2007) involve the operation of a strategy-proof mechanism that violates non-bossiness in welfare-outcome in an information environment in which information is not interior. These experiments offer us the chance to observe Nash equilibria attaining outcomes different from the truthful one with positive probability.

In Cason et al. (2006), two-agent groups (row and column) play eight to ten rounds in randomly rematched groups with the same pivotal mechanism payoff matrix over all rounds.[16] This experiment was designed to test "secure implementation" (Saijo et al., 2007). This theory obtains a characterization of scfs whose direct revelation game implements the scf itself both in dominant strategies and Nash equilibria for all complete information priors. By running experiments with the pivotal mechanism, which violates the secure implementation requirements, the authors illustrated that this may be compatible with the observation of equilibria that are not truthful equivalent. Indeed, these authors argue that even though deviations from dominant strategy play are arguably persistent in their

[14] Our benchmark of uniform bids is well defined in each finite environment. Thus it allows for a meaningful aggregation of the different studies. For the second-price auction, an alternative comparison is the rate of dominant strategy play in this mechanism and the frequency of bids that are equal to the agent's own value in the first-price auction. Among the experiments we survey, Andreoni et al. (2007) allows for this direct comparison in experimental sessions that differ only on the price rule. In this experiment, dominant strategy play in the second-price auction is 68.25%, 57.50%, 51.25%; and in the first-price auction the percentage of agents bidding their value is 6.17%, 11.92%, 19.48% for three corresponding information structures. Because there are only two sessions each under the two auction mechanisms, non-parametric tests cannot show these differences to be significant at the session level ($p = 1/3$). They are significantly different at the subject level for a variety of non-parametric and parametric tests ($p < 0.001$).

[15] Recall that our prediction is that under conditions (i) and (ii), behavior will not settle in a suboptimal equilibrium, not that behavior will necessarily converge to a truthful equilibrium.

[16] Agents are informed of their payoffs, but not of the payoff of the other agent. Each agent knows that the payoff of the other agent does not change across rounds, however. Thus, it is plausible that agents form beliefs about their opponents play that are not interior. Indeed, after some rounds, each agent has a small sample of the distribution of play of the other agent's fixed payoff type.

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 2: 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.

pivotal mechanism experiment, virtually all subjects are playing mutual best responses to the population of subjects by the end of the experiment (c.f., Figure 7, Cason et al., 2006).[17]

In Healy (2006), five-agent groups with fixed utility functions play fifty rounds in a mechanism that belongs to the VCG family to choose the level of provision of a public good. Agents have quasi-linear preferences and their utility for the public good is determined by two parameters.[18] Since it is central to his analysis, the author directly addresses the issue and concludes that "weakly dominated ε-Nash equilibria are observed, while the dominant strategy equilibrium is not" (Result 4 Healy, 2006).

In Andreoni et al. (2007), groups of four agents sequentially play three simultaneous games in each round for thirty rounds. Groups are rematched each round and play an auction game with the same values but increasing precision of information about the other players. The first game involves no information about the other players' valuations beyond the distribution from which they are drawn. The final game involves complete information. These authors run separate sessions with the first-price auction and the second-price auction.

Andreoni et al. (2007)'s main objective is to experimentally evaluate the effect of information structure on the first-price and second-price auctions. Their theoretical benchmark is the information-driven comparative statics developed by Kim and Che (2004) for the first-price auction, and the dominant strategy hypothesis, which implies there is no role of information structure, for the second-price auction. Thus, these authors designed and carried out an ideal experiment to evaluate the operation of a bossy strategy-proof scf that has unique dominant strategies, the second-price auction, in both full-support and complete information environments. In contrast to the dominant strategy hypothesis, empirical equilibrium analysis has sharp predictions for such a mechanism in these environments.

One can argue that frequencies of play in all treatments in Andreoni et al. (2007)'s experiment accumulate towards a Nash equilibrium. Fig. 1 shows the proportion of outcomes where all subjects in a group play best-responses (dark gray), where the subject with the highest valuation obtains it at the second highest valuation (light gray), and all subjects play dominant strategies (medium gray) under full-support incomplete information (left)

[17] Secure implementation is achieved by strategy-proof scfs that are non-bossy in welfare-outcome and satisfy a rectangularity condition we state in Theorem 3. Empirical equilibrium analysis reveals that Cason et al. (2006)'s experiment likely succeeded in exhibiting an undesirable equilibrium because the scf they chose violates non-bosiness in welfare-outcome in an information structure that is arguably not interior. Had these authors chosen an scf violating secure implementation but satisfying non-bossiness in welfare-outcome, like the TTC, it is unlikely that they would have observed behavior accumulating towards an untruthful Nash equilibrium (Sec. 6.1). See also our analysis of secure implementation in the context of robust implementation in Sec. 7.

[18] Again as in Cason et al. (2006), agents' types are fixed, but agents are not provided with the information of the payoff matrix of the other agents.

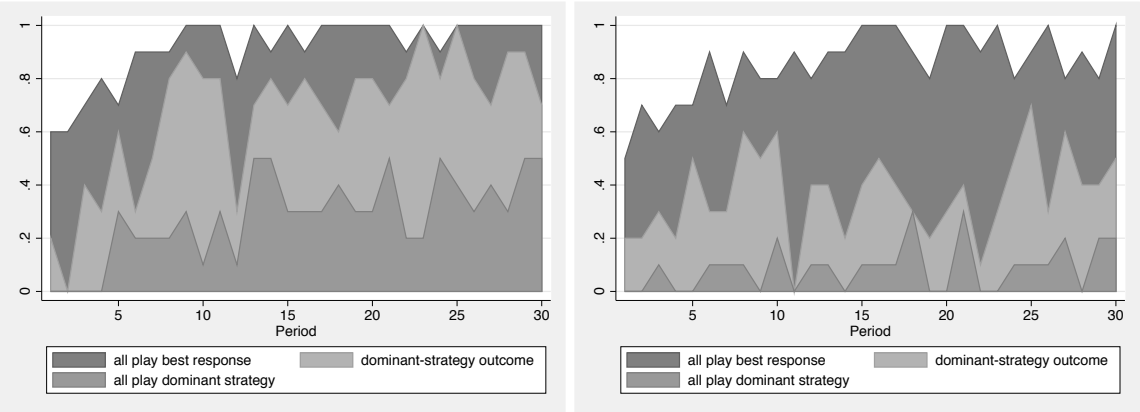

Figure 1: Group outcomes in 4-person, second-price auctions in Andreoni et al. (2007) under full-support incomplete (left) and complete (right) information. The dark gray area indicates the proportion of outcomes where all subjects play mutual best responses to the actions of all other group members. 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 across a group of 20 each period.

and complete information (right).[19] In both cases, virtually all subjects are playing mutual best responses to the population of subjects in the second half of the experiment. Note that frequencies of best response play plotted in Fig. 1 are the percentage of groups in which all four agents end up playing a best response to each other. Even when this percentage is 80%, individual rates of best response play is about 95%.

Empirical equilibrium analysis reveals that behavior that is weakly payoff monotone and approximates mutual best responses in this experiment will necessarily have certain characteristics. For the second-price auction if information is interior, as in the first information treatment, this type of behavior can *only* approximate a truthful equivalent Nash equilibrium. If information is complete, as in the last information treatment, this type of behavior *can* accumulate towards a Nash equilibrium in which the lower value agents randomize with positive probability. Both phenomena are supported by the data.

Fig. 2 allows us to understand behavior in both information structures. The figure standardizes bids to valuations (the highest valuation is assigned a value of 4, the second highest a value of 3, and so on) and shows the median bid and the range that contains the higher and lower 85% of bids for bids by each of the four ranked valuation types. In both treatments the median bid for any of the four types generally falls on its respective valuation, consistent with dominant strategy play.

[19] We concentrate our analysis on the extreme information structures in Andreoni et al. (2007) design for which Theorems 1 and 2 produce sharp predictions.

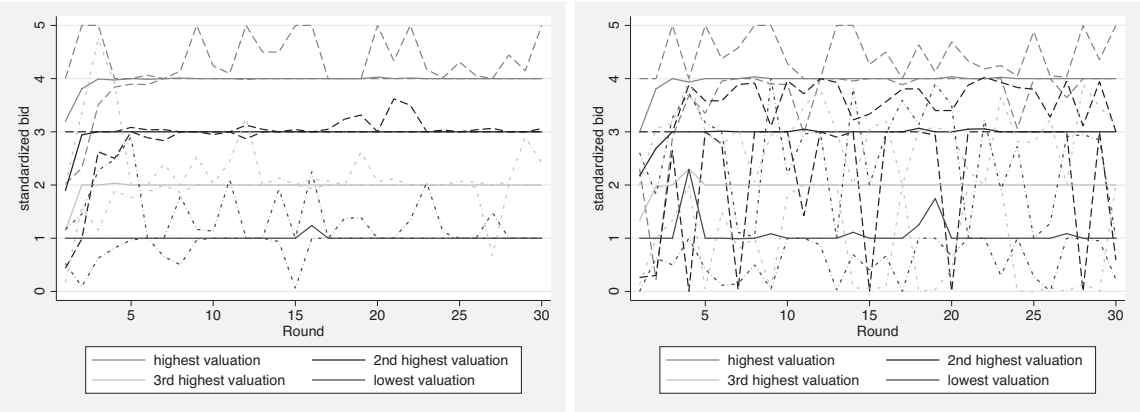

Figure 2: Median bid and 15th-85th percentile range by valuation type in 4-person, second-price auctions of Andreoni et al. (2007) under incomplete (left) and complete (right) information. Bids are standardized so that the valuation of the 1st-4th valuations in the specific auction are assigned values 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.

In the full-support incomplete information treatment, agents' deviations from their dominant strategies do not induce consequential deviations from the truthful equilibrium. After the initial five rounds, median bids are the agents' own values (Fig. 2 (left)). In the last twenty five rounds, 74.4% outcomes are truthful (Fig. 1 (left)); 97.2% outcomes are efficient, i.e., such that a highest valuation agent wins the auction (Fig. 3 (left)); in 94.4% of outcomes the price is determined by the bid of a second valuation agent; and on average the price paid by the winner differs in 1.188 points (average of the absolute value of differences) from the second highest valuation (Fig. 3 (right)). Thus, the mechanism is arguably achieving the social planner's objectives. It is virtually assigning the good to a highest valuation agent and it is essentially raising revenue equal to the second highest valuation.

In the complete information treatment, after five rounds median bids are also the agents' own values (Fig. 2 (right)). Differently from the incomplete information case, deviations from truthful behavior do not dissipate and are consequential. In the last twenty five rounds, 38.4% outcomes are truthful (Fig. 1 (right)); 91.6% outcomes are efficient, i.e., such that a highest valuation agent wins the auction (Fig. 3 (left)); in 68.4% of outcomes the price is determined by the bid of a second valuation agent; and on average the price paid by the winner differs in 8.704 points from the second highest valuation (Fig. 3 (right)).[20] Thus,

[20] Andreoni et al. (2007) only report two sessions under the second price auction. Each features a within-

even though the mechanism is assigning the good to the right agent, it is raising a revenue that is persistently away from the social planner's objective.

A simple reason explains the differences in behavior between treatments. Under incomplete information there is a penalty for a player to deviate too much from his/her dominant strategy. There is no corresponding penalty under complete information. As long as a lower valuation player does not outbid the first, the payoff of the lower valuation agent will be zero regardless. 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. 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. On the other hand, in the incomplete information treatment, for each bid, there is a positive probability that at least an agent draws that bid as 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. Thus, agents take into account a potential loss in utility, but only when there is an actual significant probability of it being realized.[21]

6.3 Payoff monotonicity

One of the advantages of empirical equilibrium analysis is that it is based on an observable property of behavior. That is, the conclusions of our theorems will hold whenever empirical distributions are weakly payoff monotone. Thus, evaluating the extent to which agents frequencies of play satisfy this property allows us to understand better the positive content of our theory.

Evaluating weak payoff monotonicity is an elusive task, however. In realistic games as those in the experiments we surveyed, action spaces and type spaces are large (e.g., Attiyeh et al., 2000 has 2001 actions). This makes the data requirements for fully testing payoff monotonicity unrealistic. It is plausible that data can point to differences on frequencies of play between two given actions for a certain agent type. In order to test that this

session comparison of these two information structures. Because there are only two paired comparisons at the session level, non-parametric tests cannot show these differences to be significant ($p = 0.5$). At the subject level, they are significantly different for a variety of non-parametric and parametric tests ($p < 0.001$).

[21] 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 what the other are doing is very low. Our analysis goes beyond this observation by showing that as predicted by empirical equilibrium analysis, the degree to which these deviations are consequential is linked to the non-bossiness properties of the scf and the information structure.

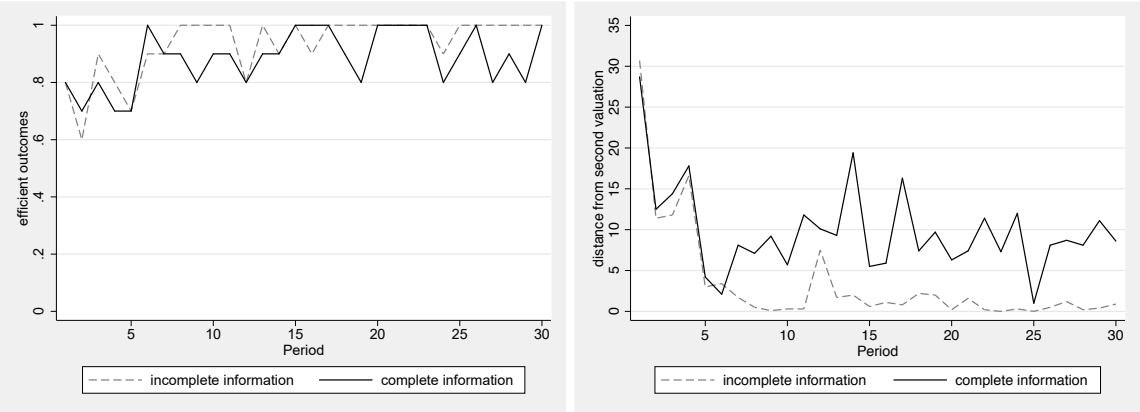

Figure 3: Frequency of efficient outcomes (left) and average distance (conditional on efficient outcome) between the price and a second valuation (right) in the second-price auction experiments of Andreoni et al. (2007) in the full-support incomplete and complete information treatments.

is consistent with weak payoff monotonicity one would need to verify that the expected payoffs of these actions given what the other agents are doing are ranked in accordance to the frequencies of play of these actions. Doing so requires, in most cases, that one has a good estimate of the *whole* distribution of play for all agent types.

Even though fully testing weak payoff monotonicity is not feasible with realistic data sets, 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 association 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 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.[22] 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 ($p < 0.001$).[23]

Not all features of data are in line with weak payoff monotonicity, however. We are aware of three of these. 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

[22] 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.

[23] 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.

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 suggest the later 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 an agent's 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, which as Andreoni et al. (2007) argue, may have origin in spiteful behavior of the low value agents.[24] Finally, a simple behavioral regularity as rounding to multiples of five, can easily induce violations of weak payoff monotonicity (such patterns are present in the auction data of Andreoni et al., 2007; Brown and Velez, 2019; Li, 2017, for instance).

In order to evaluate the positive content of empirical equilibrium analysis, it is necessary to understand the consequences for our analysis of these and other possible violations of weak payoff monotonicity. One avenue is to reconsider our construction and restart from a more basic principle than weak payoff monotonicity. Observe that this property can be stated in its contrapositive form as follows: If between two actions, say a and b, an agent's expected utility of a given what the other are doing is greater than or equal to that of b, then the frequency with which the agent plays a should be no less than the frequency with which the agent plays b. Stated in this form this property can be naturally weakened as follows. One can require the existence of some constant $\alpha \in (0, 1)$ such that for any two actions available to an agent, say a and b, if the expected utility of a given what the other are doing is greater than or equal than that of b, then the frequency with which the agent plays a should be no less than α times the frequency with which the agent plays b. One can determine that all our results follow through if we take as basis for plausibility this weaker property. It is interesting in itself to see that such a weak property still provides empirical

[24] There is a commonly accepted folk wisdom within experimental economics literature that supports the idea that private rather than common information of values may be beneficial for market outcomes (see 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. Our theory does not require nor utilize this type of behavior to justify the differences in predicted plausible equilibria between incomplete and complete information. In this particular instance, the "spiteful behavior" noted in the complete information treatment of Andreoni et al. (2007) is not present in the full-support incomplete information treatment, which makes it difficult to reconcile with any model of other regarding preferences. Thus, at least in this game, other regarding preferences play a role only when individual incentives for truthful revelation are negligible.

restrictions on data. Moreover, the main message of empirical equilibrium analysis in other applications, like full implementation, is also preserved under this generalization (Velez and Brown, 2019b).

We prefer to maintain the analysis based on weak payoff monotonicity because it strikes a balance between the regularity it provides while being challenged only by phenomena that (i) do not seem universally relevant, and (ii) seem to induce only continuous violations of this principle. Agents may round their bids, may be attracted by labels attached to certain actions, may exhibit other regarding preferences in certain contexts, and so on. At the end, what matters for empirical equilibrium analysis is that these features of behavior will be part of a bigger scheme in which agents are to a significant extent trying to hit their best payoffs given what the other agents are doing. By analyzing what happens when these less well understood effects are absent, we obtain a powerful benchmark producing policy relevant comparative statics.

Finally, an emerging empirical literature concerning strategy-proof mechanisms presents evidence in line with the predictions of empirical equilibrium. In empirical data, in which payoff types are not observable, it is of course challenging to determine what a deviation from truthful behavior is. However, in some instances the researcher is able to identify dominated actions, as an agent refusing to apply for financial support when this does not influence her acceptance to an academic position (Hassidim et al., 2016) or by means of ex-post surveys (Rees-Jones, 2017). The common finding is that these types of reports are observed with positive probability. However, in line with our results, they are more common among the agents for whom they are less likely to be consequential (Hassidim et al., 2016; Rees-Jones, 2017; Artemov et al., 2017; Chen and Pereyra, 2018; Shorrer and Sóvágó, 2019).

7 Robust mechanism design and revelation principle

One can draw an informative parallel between our results and the robust full implementation of scfs (Bergemann and Morris, 2005). This literature articulates the idea that the designer should look for mechanisms that operate well independently of informational assumptions. Of course one's judgement about this depends on the prediction that one uses. Here are the news if one considers the Nash equilibrium prediction.[25]

Theorem 3. Let g be an scf. The following are equivalent.

[25] One can even go further and require this type of robustness for all realizations of agents' types for type spaces with no rational expectations a la Bergemann and Morris (2005). In a private values model without imposing common prior discipline, very little can be done (Bergemann and Morris, 2011; Adachi, 2014). On the other hand, if one aims at obtaining the right outcomes at least when agents consider themselves mutually possible, which covers each possible realization in each common prior payoff-type space, the mechanisms characterized in Theorem 3 still do the job (Adachi, 2014).

1. There is a finite mechanism (M, φ) such that for each possible common prior p, each Bayesian Nash equilibrium σ of (M, φ, p), each possible $\theta \in \Theta$ in the support of p, and each message m in the support of $\sigma(\cdot|\theta)$, $\varphi(m) = g(\theta)$.

2. (i) g is strategy-proof and non-bossy in welfare-outcome, 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) ($1 \Rightarrow 2$) and Adachi (2014) ($2 \Rightarrow 1$) 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.

Theorem 3 allows us to make a precise comparison of Theorems 1 and 2 with the literature on robust implementation. As mentioned in the introduction, the conditions in Theorem 3 are quite restrictive (c.f. Saijo et al., 2007; Bochet and Sakai, 2010; Fujinaka and Wakayama, 2011). The outcome rectangular property is responsible for large part of these restrictions (Table 3). Thus, the aim of designing mechanisms that produce only the desired outcomes, in all Nash equilibria for all information structures, may be unnecessarily pessimistic. None of the mechanisms in Table 3 pass the test. However, if one already believes that a Nash equilibrium will be a good prediction when the mechanism is operated, it is enough to be concerned only with the Nash equilibria that is plausible will be observed. By Theorem 1, TTC, Uniform rule, and median voting pass the more realistic test for all common prior type spaces. By Theorem 2, the second-price auction, Pivotal mechanism, and SPDA pass the test for all full-support common prior type spaces.

It is worth noting that statement 1 in Theorem 3 is quantified over all finite mechanisms, while statement 1 in Theorem 1 only refers to the direct revelation game of the scf. It turns out that whenever statement 1 in Theorem 3 is satisfied by some mechanism for an scf, it is also satisfied by the scf's direct revelation mechanism (Saijo et al., 2007). This means that a "revelation principle" holds for this type of implementation.

It is not clear that a revelation principle holds when empirical equilibrium is one's prediction in these games. That is, we do not know whether there is a strategy-proof scf that violates non-bossiness in welfare-outcome for which there is a mechanism that has the properties in statement 1 of Theorem 1. The issue is very interesting and subtle.

It is known that the restriction to direct revelation mechanisms is not without loss of

scf	Strategy proofness	Essentially unique dominant strategies	Non-bossiness in welfare-outcome	outcome rectan-gular property
TTC	+	+	+	−
Uniform rule	+	+	+	−
Median voting	+	+	+	−
Second price auction	+	+	−	−
Pivotal	+	+	−	−
SPDA	+	+	−	−

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.

generality for full implementation. That is, dominant strategy full implementation may require richer message spaces than the payoff-type spaces (Dasgupta et al., 1979; Repullo, 1985). Strikingly, Repullo (1985) constructs a finite social choice environment that admits a strategy-proof social choice function whose direct revelation game for certain type has a dominant strategy equilibrium that Pareto dominates the outcome selected by the scf for that type. Moreover, the social choice environment in this example also admits a mechanism that implements in dominant strategies the social choice function.

By Lemma 1 we know that a dominant strategy profile in a game will always be observed with positive probability in each empirical equilibrium of the game.[26] Thus, Repullo (1985)'s concern that undesirable outcomes —in this case dominant strategy equilibrium outcomes— of a direct revelation game for a strategy-proof scf may be empirically plausible, is well founded. As Repullo (1985) proves, it is possible to enlarge the message spaces and tighten the incentives for the selection of a particular outcome in a way that the desired outcome is the only dominant strategy outcome. It turns out that this type of message space enlargement, i.e., those that retain the existence of dominant strategies, will not resolve the issue.

Theorem 4 (Revelation principle for dominant strategy finite mechanisms)**.** Let g be an scf. The following are equivalent.

1. There is a finite mechanism (M, φ) for which each agent type has at least a weakly dominant action, and such that for each possible common prior p, each empirical equilibrium σ of (M, φ, p), each possible type $\theta \in \Theta$ in the support of p, and each message m in the support of $\sigma(\cdot|\theta)$, $\varphi(m) = g(\theta)$.

2. For each common prior p and each empirical equilibrium of (Θ, g, p), say σ, we have

[26] Observe also that by Theorem 1, Repullo (1985)'s scf necessarily violates non-bossiness in welfare-outcome.

that for each pair $\{\theta, \tau\} \subseteq \Theta$ where θ is in the support of p and τ is in the support of $\sigma(\cdot|\theta)$, $g(\theta) = g(\tau)$.

3. g is strategy-proof and non-bossy in welfare-outcome.

Theorem 4 implies that it is impossible to obtain robust implementation in empirical equilibrium of a social choice function that violates non-bossiness in welfare-outcome by a dominant strategies mechanism. It is worth noting that enlarging the message space on the direct revelation game of a strategy-proof scf that violates non-bossiness in welfare-outcome may have a meaningful effect on the performance of the mechanism, even when one preserves the existence of dominant strategies.

Example 1. Consider an environment with two agents $N \equiv \{1, 2\}$ whose payoff-type spaces are $\Theta_1 \equiv \{\theta_1\}$ and $\Theta_2 \equiv \{\theta_2, \theta_2'\}$. There are two possible outcomes $\{a, b\}$; and $u_1(a|\theta_1) > u_1(b|\theta_1)$, $u_2(a|\theta_2) = u_1(b|\theta_2)$, and $u_2(a|\theta_2') < u_2(b|\theta_2')$. Suppose that a social planner desires to implement the efficient dictatorship in which agent 2 gets her top choice. One can easily see that for any common prior p, for each empirical equilibrium of (Θ, g, p), say σ, agent 2 with payoff type θ_2 uniformly randomizes in Θ_2. Thus, in each empirical equilibrium of (Θ, g, p), agent 2 always achieves her top choice and agent 1 receives her top choice with $1/2$ probability when this does not conflict with agent 2's preferences. Suppose now that the social planner uses mechanism (M, φ) defined as follows: $M_1 \equiv \{\theta_1\}$, $M_2 \equiv \{\theta_2', m_2^1, ..., m_2^k\}$ where $k \in \mathbb{N}$, $\varphi(\theta_1, \theta_2') = b$, and for each $l = 1, ..., k$, $\varphi(\theta_1, m_2^l) = a$. One can see easily that in each empirical equilibrium of (M, φ, p), agent 2 always achieves her top choice and agent 1 receives her top choice with $k/(k+1)$ probability when this does not conflict with agent 2's preferences. □

Finally, it is well known that the restriction to social choice *functions* is not without loss of generality in robust implementation. Indeed, 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 2 in the Appendix).

8 Conclusion

We have presented theoretical evidence that strategy-proof mechanisms are not all the same. Our analysis is based on empirical equilibrium, a refinement of Nash equilibrium that is based only on observables. It selects all the Nash equilibria that are not rejected as implausible by some model that is disciplined by weak payoff monotonicity. We draw

two main conclusions under the hypothesis that observable behavior satisfies this property. First, behavior from the operation of a strategy-proof and non-bossy in welfare-outcome scf will never approximate a sub-optimal Nash equilibrium. Second, if the mechanism violates the non-bossiness condition but has essentially unique dominant strategies, then behavior can approximate a sub-optimal equilibrium only if information is not interior. These predictions are supported by experimental data on multiple mechanisms. The weak payoff monotonicity hypothesis fares well in data, but violations of it can be spotted in particular environments. These violations do not hinder the main conclusions of our study, however.

Our results can be interpreted as positive developments in the theory of mechanism design. Existence of strategy-proof mechanisms is difficult on itself. Many of them do not pass the higher bar set by other approaches (e.g. Saijo et al., 2007; Li, 2017). Instead of trying to redesign strategy-proof mechanisms, we tried to understand them better. Our results then allowed us to come to terms with the experimental data that is against the dominant strategy hypothesis. Essentially, we learned that even though behavior in strategy-proof mechanisms may not quickly converge to a truthful equilibrium, many of these mechanisms (the non-bossy in welfare-outcome) will likely never get stuck in a sub-optimal self-enforcing state, and most of these mechanisms (the ones with essentially unique dominant strategies) will have this problem only for corner information structures.

Appendix

Proof of Lemma 1. Let $\Gamma \equiv (M, \varphi, p)$ and $\sigma \in N(\Gamma)$ be as in the statement of the lemma. 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 $\theta_i \in T_i$, as $\lambda \to \infty$, $\sigma^\lambda(\cdot|\theta_i) \to \sigma(\cdot|\theta_i)$. Let $\lambda \in \mathbb{N}$ and $m_{-i} \in M_{-i}$. Since m_i is a weakly dominant action for agent i with type θ_i in (M, φ), for each $r_i \in M_i$,

$$u_i(\varphi(m_{-i}, m_i)|\theta_i) \geq u_i(\varphi(m_{-i}, r_i)|\theta_i).$$

Thus,

$$U_\varphi(\sigma^\lambda_{-i}, \delta_{m_i}|p, \theta_i) \geq U_\varphi(\sigma^\lambda_{-i}, \delta_{r_i}|p, \theta_i).$$

Since σ is weakly payoff monotone for Γ, we have that for each $r_i \in M_i$,

$$\sigma^\lambda_i(m_i|\theta_i) \geq \sigma^\lambda_i(r_i|\theta_i).$$

Convergence implies that

$$\sigma_i(m_i|\theta_i) \geq \sigma_i(r_i|\theta_i).$$

Thus, m_i is in the support of $\sigma_i(\cdot|\theta_i)$. $\square$

Proof of Theorems 1 and 4. We prove Theorem 4, which implies Theorem 1. We first prove that statement 3 in the theorem implies statement 2. Suppose that g is strategy-proof and non-bossy in welfare-outcome. Let p be a common prior and σ an empirical equilibrium of (Θ, g, p). Let $\theta \in \Theta$ be in the support of p. Thus, for each $i \in N$, $p(\theta_{-i}|\theta_i) > 0$. Let $\tau \in \Theta$ be in the support of $\sigma(\cdot|\theta)$, i.e., τ is a report that is observed with positive probability when the true types are θ. Let $i \in N$. Since $\sigma \in N(\Theta, g, p)$, we have that

$$U_g(\sigma_{-i}, \delta_{\tau_i}|p, \theta_i) \geq U_g(\sigma_{-i}, \delta_{\theta_i}|p, \theta_i).$$

Since g is strategy-proof, the integrand of the expression on the right dominates point-wise the integrand of the expression on the left. Thus, the integrands are equal on the support of the common integrating measure. Notice that since $p(\theta_{-i}|\theta_i) > 0$ and τ is in the support of $\sigma(\cdot|\theta)$, agent i assigns positive probability that the other agents profile of reports is τ_{-i}. Thus,

$$u_i(g(\tau)|\theta_i) = u_i(g(\tau_{-i}, \theta_i)|\theta_i).$$

Since g is non-bossy in welfare-outcome,

$$g(\tau) = g(\tau_{-i}, \theta_i). \tag{1}$$

By Lemma 1, θ_i is in the support of $\sigma_i(\cdot|\theta_i)$. Thus, (τ_{-i}, θ_i) is in the support of $\sigma(\cdot|\theta)$. Thus, the recursive argument shows that $g(\tau) = g(\theta)$.

We now prove that statement 2 implies statement 1. Since each empirical equilibrium is a Bayesian Nash equilibrium, statement 2 implies that for each common prior p there is a Bayesian Nash equilibrium of (Θ, g, p) that obtains for each $\theta \in \Theta$, $g(\theta)$ with probability one. It is well known that this implies g is strategy-proof (Dasgupta et al., 1979; Bergemann and Morris, 2005).[27] Thus, (Θ, g) is a dominant strategies mechanism that satisfies the conditions in statement 1 of the theorem.

We now prove that statement 1 implies statement 3. Suppose that statement 1 is satisfied. That is, there is a finite mechanism (M, φ) for which each agent type has at least a dominant strategy, and such that for each common prior p, each empirical equilibrium σ of (M, φ, p), each possible type $\theta \in \Theta$ in the support of p, and each message m in the support of $\sigma(\cdot|\theta)$, $\varphi(m) = g(\theta)$.

We prove that g is strategy-proof. For each $i \in N$ and $\theta_i \in \Theta_i$, let $m_i(\theta_i)$ be a weakly

[27] This can be easily seen by analyzing for each $\theta \in \Theta$ and $\tau_i \in \Theta_i$, the common prior $p = (1/2)\delta_\theta + (1/2)\delta_{(\theta_{-i}, \tau_i)}$. See Theorem 2 for the explicit proof of a slightly stronger result where this is obtained for interior common priors.

dominant action for i with type θ_i in (M, φ). Let p be a full-support common prior and σ an empirical equilibrium of (M, φ, p). By Lemma 1, for each $i \in N$ and $\theta_i \in \Theta_i$, $m_i(\theta_i)$ is in the support of $\sigma(\cdot|\theta_i)$. By statement 1, for each $\theta \in \Theta$, $\varphi(m(\theta)) = g(\theta)$ (this means that (M, φ) fully implements g in dominant strategy equilibria, which by the usual revelation principle argument, which we spell out next, implies g is strategy-proof). Let $\theta \in \Theta$, $i \in N$ and $\tau_i \in \Theta_i$. Since $m_i(\theta_i)$ is a weakly dominant action for i with type θ_i in (M, φ), we have that $u_i(\varphi(m(\theta)|\theta_i) \geq u_i(\varphi(m_{-i}(\theta_{-i}), m_i(\tau_i))|\theta_i)$. Thus, $u_i(g(\theta)|\theta_i) \geq u_i(g(\theta_{-i}, \tau_i)|\theta_i)$. Thus, g is strategy-proof.

We prove that g is non-bossy in welfare-outcome. Suppose by contradiction that there is $\theta \in \Theta$, $i \in N$, and $\tau_i \in \Theta_i$ such that $u_i(g(\theta)|\theta_i) = u_i(g(\theta_{-i}, \tau_i)|\theta_i)$ and $g(\theta) \neq g(\theta_{-i}, \tau_i)$. Suppose without loss of generality that this agent is $i = 1$. Let $a \equiv g(\theta)$ and $b \equiv g(\theta_{-1}, \tau_1)$. Again, for each $i \in N$ and $\theta_i \in \Theta_i$, let $m_i(\theta_i)$ be a weakly dominant action for i with type θ_i in (M, φ). We claim that $m_1(\tau_1)$ is a best response to $m_{-1}(\theta_{-1})$ for agent i with type θ_1, i.e., for each $m_1' \in M_1$,

$$u_1(\varphi(m_{-1}(\theta_{-1}), m_1(\tau_1))|\theta_1) \geq u_1(\varphi(m_{-1}(\theta_{-1}), m_1')|\theta_1). \tag{2}$$

By Lemma 1, in each empirical equilibrium of $(M, \varphi, (\theta_{-1}, \tau_1))$, $(m_{-1}(\theta_{-1}), m_1(\tau_1))$ is played with positive probability and in each empirical equilibrium of (M, φ, θ), $(m_{-1}(\theta_{-1}), m_1(\theta_1))$ is played with positive probability. Since (M, φ) satisfies statement 1, $\varphi(m_{-1}(\theta_{-1}), m_1(\tau_1)) = b$ and $\varphi(m_{-1}(\theta_{-1}), m_1(\theta)) = a$. Since $m_1(\theta_1)$ is a dominant strategy for agent 1 with type θ_1, we have that for each $m_1' \in M_1$, $u_1(\varphi(m_{-1}(\theta_{-1}), m_1(\tau_1))|\theta_1) = u_1(\varphi(m_{-1}(\theta_{-1}), m_1(\theta_1))|\theta_1) \geq u_1(\varphi(m_{-1}(\theta_{-1}), m_1')|\theta_1)$. This is (2).

Consider the complete information game (M, φ, θ). Let σ^* be the profile of strategies in (M, φ, θ) defined as follows. For each agent $j \neq 1$, σ_j^* uniformly randomizes among j's weakly dominant actions; agent 1 uniformly randomizes among her best responses to σ_{-1}^*. (Recall that in a complete information game we do not condition strategies on agents' types, i.e., σ_i^* is the strategy of agent i with type θ_i.) Clearly, σ^* is a Bayesian Nash equilibrium of (M, φ, θ). Since $m_{-1}(\theta_{-1})$ in (2) is an arbitrary profile of weakly dominant strategies for agents $N \setminus \{i\}$ with type θ_{-1} in (M, φ), we have that for agent i with type θ_1, $m_1(\tau_1)$ is a best response to σ_{-1}^* in (M, φ, θ), i.e., for each $m_1' \in M_1$,

$$U_\varphi(\sigma_{-i}^*, \delta_{m_1(\tau_1)}|\delta_{\theta_{-1}}, \theta_i) \geq U_\varphi(\sigma_{-i}^*, \delta_{m_1'}|\delta_{\theta_{-1}}, \theta_i).$$

Thus, $(m_{-1}(\theta_{-1}), m_1(\tau_1))$ is in the support of σ^*. Thus, σ^* is a Bayesian Nash equilibrium of (M, φ, θ) that obtains with positive probability outcome $b = \varphi(m_{-1}(\theta_{-1}), m_1(\tau_1))$ (when the agents' type is θ, which is the only element in the support of the prior).

The proof concludes by showing that σ^* is an empirical equilibrium of (M, φ, θ), which contradicts statement 1 because $b \neq a$. We follow the intuition that we presented in Sec. 5 for the direct revelation mechanism (Θ, g).

We will make use of the so-called Quantal Response Equilibria (McKelvey and Palfrey, 1995), which are weakly payoff monotone distributions. A quantal response function for agent i is a continuous function $Q_i : \mathbb{R}^{M_i} \to \Delta(M_i)$. For each $m_i \in M_i$, $Q_{im_i}(x)$ denotes the value assigned to m_i by $Q_i(x)$. We refer to the list $Q \equiv (Q_i)_{i \in N}$ simply as a quantal response function. Agent i's quantal response function Q_i is *monotone* if for each $x \in \mathbb{R}^{M_i}$, and each pair $\{m_i, m_i'\} \subseteq M_i$ such that $x_{m_i} > x_{m_i'}$, $Q_{im_i}(x) > Q_{im_i'}(x)$ (Goeree et al., 2005). The *logistic* quantal response function with parameter $\lambda \geq 0$, denoted by l^λ, assigns to each $m \in M_i$ and each $x \in \mathbb{R}^{M_i}$ the value,

$$l^\lambda_{im}(x) \equiv \frac{e^{\lambda x_m}}{\sum_{t \in M_i} e^{\lambda x_t}}. \tag{3}$$

It can easily be checked that for each $\lambda \geq 0$, the corresponding logistic quantal response function is continuous and monotone (McKelvey and Palfrey, 1995). A *quantal response equilibrium* of $\Gamma \equiv (M, \varphi, \theta)$ with respect to quantal response function Q is a fixed point of the composition of Q and the expected payoff operator in Γ (McKelvey and Palfrey, 1995), i.e., a strategy profile for (M, φ, θ), $\sigma \equiv (\sigma_i)_{i \in N}$, such that for each $i \in N$, $\sigma_i = Q_i(U_\varphi(\sigma_{-i}, \delta_{m_i} | \delta_{\theta_{-i}}, \theta_i)_{m_i \in M_i})$. Brouwer's fixed point theorem guarantees that for each continuous quantal response function there is a quantal response equilibrium associated with it (McKelvey and Palfrey, 1995). One can easily see that if the quantal response function is monotone, each of its quantal response equilibria are weakly payoff monotone.

For each $j \in N$, $\varepsilon \in (0,1)$, and $\lambda \in \mathbb{N}$ let $n_j \equiv |M_j|$ and $\kappa_j^{\varepsilon,\lambda}$ the quantal response function that for each $x \in \mathbb{R}^{M_j}$,

$$\kappa_j^{\varepsilon,\lambda}(x) \equiv \varepsilon / n_j + (1 - \varepsilon) l^\lambda(x).$$

Since l^λ is continuous and monotone, so is $\kappa_j^{\varepsilon,\lambda}$. Fix $\varepsilon > 0$, $\delta > 0$, and $r \in \mathbb{N}$. By continuity of $\kappa_1^{\varepsilon,r}$ and the expected utility operator, as $\eta \to 0$,

$$\kappa_1^{\varepsilon,r}(U_\varphi((\eta/n_j + (1-\eta)\sigma_j^*)_{i \in N \setminus \{1\}}, \delta_{m_1} | \delta_{\theta_{-1}}, \theta_1)_{m_1 \in M_1}) \to \kappa_1^{\varepsilon,r}(U_\varphi(\sigma_{-1}^*, \delta_{m_1} |, \theta_1)_{m_1 \in M_1}).$$

By monotonicity of $\kappa_1^{\varepsilon,r}$, $\kappa_1^{\varepsilon,r}(U_\varphi(\sigma_{-1}^*, \delta_{m_1} |, \theta_1)_{m_1 \in M_1})$ places maximal probability on the best responses for agent 1 to σ_{-1}^*. Thus there is $\eta(\varepsilon, r, \delta) < \delta$ such that for each $m_1^* \in M_1$

that is a best response for agent 1 to σ^*_{-1}, the distance between

$$\kappa^{\varepsilon,r}_{1m^*_1}(U_\varphi((\eta(\varepsilon,r,\delta)/n_j + (1-\eta(\varepsilon,r,\delta))\sigma^*_j)_{i\in N\setminus\{1\}}, \delta_{m_1}|\delta_{\theta_{-1}},\theta_1)_{m_1\in M_1}))$$

and

$$\kappa^{\varepsilon,r}_{1m_1(\theta_1)}(U_\varphi((\eta(\varepsilon,r,\delta)/n_j + (1-\eta(\varepsilon,r,\delta))\sigma^*_j)_{i\in N\setminus\{1\}}, \delta_{m_1}|\delta_{\theta_{-1}},\theta_1)_{m_1\in M_1})),$$

is at most $\delta/2$. Fix such a $\eta(\varepsilon,r,\delta)$. Consider a sequence of quantal response equilibria for the sequence of quantal response functions

$$\{(\kappa^{\varepsilon,r}_1, \kappa^{\eta(\varepsilon,r,\delta),t}_2, ..., \kappa^{\eta(\varepsilon,r,\delta),t}_n)\}_{t\in\mathbb{N}}.$$

Let $\{\sigma^t\}_{t\in\mathbb{N}}$ be this sequence. Compactness of the simplex of probabilities implies that there is a convergent subsequence. Without loss of generality we assume then that $\{\sigma^t\}_{t\in\mathbb{N}}$ is convergent and its limit as $t\to\infty$ is, say σ. Since each agent places in each action a probability that is at least the minimum between ε/n_1 and $\min\{\eta(\varepsilon,r,\delta)/n_j : j\in N\setminus\{1\}\}$, σ is interior. Now, observe that for each $t\in\mathbb{N}$, each $j\in N\setminus\{1\}$, and each $m'_j\in M_j$,

$$\frac{l^t_{m'_j}(U_\varphi(\sigma^t_{-j},\delta_{m_j}|\delta_{\theta_{-j}},\theta_j)_{m_j\in M_j})}{l^t_{m_j(\theta_j)}(U_\varphi(\sigma^t_{-j},\delta_{m_j}|\delta_{\theta_{-j}},\theta_j)_{m_j\in M_j})} = e^{t(U_\varphi(\sigma^t_{-j},\delta_{m_j}|\delta_{\theta_{-j}},\theta_j)-U_\varphi(\sigma^t_{-j},\delta_{m_j(\theta_j)}|\delta_{\theta_{-j}},\theta_j))}.$$

Suppose that m'_j is not a dominant action for j in (M,φ). Since σ_{-j} is interior, we have that

$$U_\varphi(\sigma_{-j},\delta_{m_j}|\delta_{\theta_{-j}},\theta_j) - U_\varphi(\sigma_{-j},\delta_{m_j(\theta)}|\delta_{\theta_{-j}},\theta_j) < 0.$$

Since as $t\to\infty$, $\sigma^t\to\sigma$, we also have that as $t\to\infty$,

$$\frac{l^t_{m'_j}(U_\varphi(\sigma^t_{-j},\delta_{m_j}|\delta_{\theta_{-j}},\theta_j)_{m_j\in M_j})}{l^t_{m_j(\theta_j)}(U_\varphi(\sigma^t_{-j},\delta_{m_j}|\delta_{\theta_{-j}},\theta_j)_{m_j\in M_j})}\to 0. \tag{4}$$

By monotonicity of l^t, $l^t(U_\varphi(\sigma^t_{-j},\delta_{m_j}|,\theta_j)_{m_j\in M_1})$ places maximal probability on the best responses for agent j to σ^t_{-j}. Thus, it places maximal probability on $m_j(\theta_j)$. Since as $t\to\infty$, $\sigma^t\to\sigma$, the expressions in the numerator and denominator of (4) form convergent sequences. Thus, $\lim_{t\to\infty} l^t_{m_j(\theta_j)}(U_\varphi(\sigma^t_{-j},\delta_{m_j}|\delta_{\theta_{-j}},\theta_j)_{m_j\in M_j}) \geq 1/n_j > 0$. By (4), as $t\to\infty$, $l^t_{m'_j}(U_\varphi(\sigma^t_{-j},\delta_{m_j}|\delta_{\theta_{-j}},\theta_j)_{m_j\in M_j})\to 0$. Thus, $\sigma_j = \eta(\varepsilon,r,\delta)/n + (1-\eta(\varepsilon,r,\delta))\sigma^*_j$.

Now, for agent 1, since both parameters in her quantal response function are fixed in the sequence,

$$\sigma_1 = \kappa^{\varepsilon,r}_1(U_\varphi((\eta(\varepsilon,r,\delta)/n_j + (1-\eta(\varepsilon,r,\delta))\sigma^*_j)_{i\in N\setminus\{1\}}, \delta_{m_1}|\delta_{\theta_{-1}},\theta_1)_{m_1\in M_1}).$$

Thus, there is $t > r$ such that the max distance, between σ^t and σ is $\delta/4$. Let $\gamma^{\varepsilon,r,\delta} = \sigma^t$ for such a t. By our choice of $\eta(\varepsilon, r, \delta)$, for each $m_1^* \in M_1$ that is a best response to σ_{-1}^* for agent 1 with type θ_1, the distance between $\kappa_{1m_1^*}^{\varepsilon,r}(U_\varphi(\gamma_{-1}^r, \delta_{m_1}|\delta_{\theta_{-1}}, \theta_1)_{m_1 \in M_1})$ and $\kappa_{1m_1(\theta_1)}^{\varepsilon,r}(U_\varphi(\gamma_{-i}^r, \delta_{m_1}|\delta_{\theta_{-1}}, \theta_1)_{m_1 \in M_1})$ is at most δ.

For each $r \in \mathbb{N}$, let $\varepsilon(r) \equiv 1/r$ and $\delta(r) \equiv 1/r$. Let $\eta(r) \equiv \eta(\varepsilon(r), r, \delta(r))$ and $\gamma^r \equiv \gamma^{\varepsilon(r),r,\delta(r)}$ be constructed as above. By passing to a subsequence if necessary we can suppose without loss of generality that $\{\gamma^r\}_{r\in\mathbb{N}}$ is convergent. Since $0 < \eta(r) < 1/r$, we have that as $r \to \infty$, $\eta(r) \to 0$. Let $j \neq 1$. By our construction, the maximum distance between γ_j^r and $\eta(r)/n + (1 - \eta(r))\sigma_j^*$ is at most $\delta(r)/4$. Thus, as $r \to \infty$, $\gamma_j^r \to \sigma_j^*$.

Let μ_1 be the limit as $r \to \infty$ of γ_1^r. For each $m_1^* \in M_1$ that is a best response for θ_1 to σ_{-1}^*, we have that $|\gamma_1^r(m_1^*) - \gamma_1^r(m_1(\theta_1)| \leq \delta(r)$. Thus, $\mu_1(m_1^*) = \mu_1(m_1(\theta_1))$. Since $\kappa_1^{\varepsilon,r}$ is monotone, $\mu_1(m_1(\theta_1)) > 0$. Now, observe that for each $r \in \mathbb{N}$, and each $m_1' \in M_1$,

$$\frac{l_{m_1'}^r(U_\varphi(\gamma_{-1}^r, \delta_{m_1}|\delta_{\theta_{-1}}, \theta_1)_{m_1\in M_1})}{l_{m_1(\theta_1)}^r(U_\varphi(\gamma_{-1}^r, \delta_{m_1}|\delta_{\theta_{-1}}, \theta_1)_{m_1\in M_1})} = e^{r(U_\varphi(\gamma_{-1}^r, \delta_{m_1}|\delta_{\theta_{-1}}, \theta_1) - U_\varphi(\gamma_{-1}^r, \delta_{m_1(\theta_1)}|\delta_{\theta_{-1}}, \theta_1))}.$$

If $m_1' \in M_1$ is not a best response to σ_{-1}^*,

$$U_\varphi(\sigma_{-1}^*, \delta_{m_1}|\delta_{\theta_{-1}}, \theta_1) < U_\varphi(\sigma_{-1}^*, \delta_{m_1(\theta_1)}|\delta_{\theta_{-1}}, \theta_1).$$

Thus, $\mu_1(m_1')/\mu_1(m_1(\theta_1)) = 0$ and $\mu_1(m_1') = 0$. Thus, $\mu_1 = \sigma_1^*$. Since each γ^r is weakly payoff monotone and as $r \to \infty$, $\gamma^r \to \sigma^*$, we have that σ^* is an empirical equilibrium of (M, φ, θ). □

Proof of Theorem 2. Suppose that statement 1 is satisfied. 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 has full-support. Let σ be a Bayesian Nash equilibrium of (Θ, g, p) such 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 essentially unique dominant strategies. 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.

Suppose now that g is strategy-proof and has essentially unique dominant strategies. Let p have full support and σ be an empirical equilibrium of (Θ, g, p). Let $\theta \in \Theta$. We prove that $\sigma(\cdot|\theta)$ obtains $g(\theta)$ with probability one. Let $i \in N$. Suppose that τ_i is in the support of $\sigma_i(\cdot|\theta_i)$. We first prove that for each $\tau_{-i} \in \Theta_{-i}$, $g(\tau_{-i}, \theta_i) = g(\tau_{-i}, \tau_i)$. Since σ is a Bayesian Nash equilibrium

$$U_g(\delta_{\tau_i}, \sigma_{-i}|p, \theta_i) \geq U_g(\delta_{\theta_i}, \sigma_{-i}|p, \theta_i).$$

Since g is strategy-proof, the integrand of the expression on the right dominates point-wise the integrand of the expression on the left. Thus, the integrands are equal on the support of the common integrating measure. Since $p(\tau_{-i}|\theta_i) > 0$ and since by Lemma 1 the probability with which τ_{-i} is realized for $\sigma_{-i}(\cdot|\tau_{-i})$ is positive, we have that

$$u_i(g(\tau_{-i}, \tau_i)|\theta_i) = u_i(g(\tau_{-i}, \theta_i)|\theta_i). \tag{5}$$

We claim that $g(\tau_{-i}, \tau_i) = g(\tau_{-i}, \theta_i)$. Suppose by contradiction that $g(\tau_{-i}, \tau_i) \neq g(\tau_{-i}, \theta_i)$. This means that $\theta_i \neq \tau_i$. Since g has essentially unique dominant strategies, there is $\mu_{-i} \in \Theta_{-i}$ such that $u_i(g(\mu_{-i}, \theta_i)|\theta_i) > u_i(g(\mu_{-i}, \tau_i)|\theta_i)$. This contradicts (5), which holds for arbitrary $\tau_{-i} \in \Theta_{-i}$.

Let $\tau \in \Theta$ be in the support of $\sigma(\cdot|\theta)$. Then $g(\tau) = g(\tau_{-i}, \theta_i)$. By Lemma 1, θ_i is in the support of $\sigma_i(\cdot|\theta_i)$. Thus, (τ_{-i}, θ_i) is also in the support of $\sigma(\cdot|\theta)$. By iterating for the other agents we get that $g(\tau) = g(\theta)$. □

We finally 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, 2019a).

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

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 $M_1 \equiv \{m_1^1, m_1^2, m_1^3, m_1^4\}$, $M_2 \equiv \{m_2^1, m_2^2\}$, 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 a Bayesian Nash 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 $F(\theta)$. $\square$

References

Abdulkadiroğlu, A., Sönmez, T., June 2003. School choice: A mechanism design approach. Amer Econ Review 93 (3), 729–747.
URL https://doi.org/10.1257/000282803322157061

Adachi, T., 2014. Robust and secure implementation: equivalence theorems. Games Econ Behavior 86 (0), 96 – 101.
URL http://dx.doi.org/10.1016/j.geb.2014.03.015

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.
URL http://dx.doi.org/10.1016/j.geb.2006.09.003

Artemov, G., Che, Y.-K., He, Y., 2017. Strategic 'mistakes': Implications for market design research, Mimeo.

Attiyeh, G., Franciosi, R., Isaac, R. M., Jan 2000. Experiments with the pivot process for providing public goods. Public Choice 102 (1), 93–112.
URL https://doi.org/10.1023/A:1005025416722

Bade, S., Gonczarowski, Y. A., 2017. Gibbard-satterthwaite success stories and obvious strategyproofness.
URL https://arxiv.org/abs/1610.04873

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.

Barberà, S., Berga, D., Moreno, B., April 2016. Group strategy-proofness in private good economies. Amer Econ Review 106 (4), 1073–99.
URL https://doi.org/10.1257/aer.20141727

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.
URL http://www.jstor.org/stable/3598751

Bergemann, D., Morris, S., 2011. Robust implementation in general mechanisms. Games and Economic Behavior 71 (2), 261 – 281.
URL http://dx.doi.org/10.1016/j.geb.2010.05.001

Bochet, O., Sakai, T., 2010. Secure implementation in allotment economies. Games Econ Behavior 68 (1), 35 – 49.
URL http://dx.doi.org/10.1016/j.geb.2009.04.023

Bochet, O., Tumennassan, N., 2017. One truth and a thousand lies: Focal points in mechanism design, mimeo.

Brown, A. L., Velez, R. A., 2019. 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.
URL http://www.sciencedirect.com/science/article/pii/S1094202599900820

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.
URL http://dx.doi.org/10.1016/j.geb.2005.12.007

Chen, L., Pereyra, J. S., 2018. Self selection in school choice, Mimeo.

Chen, Y., Sönmez, T., 2006. School choice: an experimental study. Journal of Economic Theory 127 (1), 202 – 231.
URL http://www.sciencedirect.com/science/article/pii/S0022053104002418

Cooper, D. J., Fang, H., 2008. Understanding overbidding in second price auctions: An experimental study*. The Economic Journal 118 (532), 1572–1595.
URL https://doi.org/10.1111/j.1468-0297.2008.02181.x

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.
URL https://doi.org/10.1111/j.1465-7295.1980.tb00556.x

Dasgupta, P., Hammond, P., Maskin, E., 1979. The implementation of social choice rules: Some general results on incentive compatibility. Review Econ Studies 46 (2), 185–216.
URL http://www.jstor.org/stable/2297045

de Clippel, G., October 2014. Behavioral implementation. American Economic Review 104 (10), 2975–3002.
URL https://doi.org/10.1257/aer.104.10.2975

de Clippel, G., Saran, R., Serrano, R., 2017. Level-k mechanism design, mimeo.

Eliaz, K., 2002. Fault tolerant implementation. The Review of Econ Stud 69 (3), 589–610.
URL http://www.jstor.org/stable/1556711

Fernandez, M. A., 2018. Deferred acceptance and regret-free truthtelling: A characterization result, ph.D. thesis, California Institute of Technology.

Fudenberg, D., He, K., 2018. Player-compatible equilibrium, mimeo, Accessed on October 4th, 2018.
URL http://economics.mit.edu/files/15442

Fujinaka, Y., Wakayama, T., 2011. Secure implementation in Shapley-Scarf housing markets. Econ Theory 48 (1), 147–169.
URL http://dx.doi.org/10.1007/s00199-010-0538-x

Gale, D., Shapley, L. S., 1962. College admissions and the stability of marriage. American Math Monthly 69 (1), 9–15.
URL http://www.jstor.org/stable/2312726

Gibbard, A., 1973. Manipulation of voting schemes: A general result. Econometrica 41 (4), 587–601.
URL http://www.jstor.org/stable/1914083

Goeree, J. K., Holt, C. A., Palfrey, T. R., 2005. Regular quantal response equilibrium. Experimental Economics 8 (4), 347–367.
URL http://dx.doi.org/10.1007/s10683-005-5374-7

Green, J., Laffont, J.-J., 1977. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. Econometrica 45 (2), 427–438.
URL http://www.jstor.org/stable/1911219

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.
URL https://doi.org/10.1007/BF01737554

Harstad, R. M., Dec 2000. Dominant strategy adoption and bidders' experience with pricing rules. Experimental Economics 3 (3), 261–280.
URL https://doi.org/10.1007/BF01669775

Hassidim, A., Romm, A., Shorrer, R. I., May 26 2016. 'strategic' behavior in a strategy-proof environment.
URL https://ssrn.com/abstract=2784659

Healy, P. J., 2006. Learning dynamics for mechanism design: An experimental comparison of public goods mechanisms. J Econ Theory 129 (1), 114 – 149.
URL https://doi.org/10.1016/j.jet.2005.03.002

Jackson, M. O., 1991. Bayesian implementation. Econometrica 59 (2), 461–477.
URL http://www.jstor.org/stable/2938265

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.
URL http://www.jstor.org/stable/1913557

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.
URL http://www.jstor.org/stable/2234706

Kawagoe, T., Mori, T., Aug 2001. Can the pivotal mechanism induce truth-telling? an experimental study. Public Choice 108 (3), 331–354.
URL https://doi.org/10.1023/A:1017542406848

Kim, J., Che, Y.-K., 2004. Asymmetric information about rivals' types in standard auctions. Games Econ Behavior 46 (2), 383–397.
URL https://doi.org/10.1016/S0899-8256(03)00126-X

Kneeland, T., 2017. Mechanism design with level-k types: theory and applications to bilateral trade, wBZ Discussion paper SPII 2017-303.

Kohlberg, E., Mertens, J.-F., 1986. On the strategic stability of equilibria. Econometrica 54 (5), 1003–1037.
URL http://www.jstor.org/stable/1912320

Li, S., November 2017. Obviously strategy-proof mechanisms. Amer Econ Review 107 (11), 3257–87.
URL http://dx.doi.org/10.1257/aer.20160425

Masuda, T., Sakai, T., Serizaway, S., Wakayama, T., 2019. A strategy-proof mechanism should be announced to be strategy-proof: An experiment for the Vrickey auction, Discussion paper No. 1048, The Institute of Social and Economic Research, Osaka University.

McKelvey, R. D., Palfrey, T. R., 1995. Quantal response equilibria for normal form games. Games and Economic Behavior 10 (1), 6–38.
URL http://dx.doi.org/10.1006/game.1995.1023

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

Milgrom, P., Mollner, J., 2017. Extended proper equilibrium, mimeo.
URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035565

Milgrom, P., Mollner, J., 2018. Equilibrium selection in auctions and high stakes games. Econometrica 86 (1), 219–261.
URL https://doi.org/10.3982/ECTA12536

Moulin, H., 1980. On strategy-proofness and single peakedness. Public Choice 35 (4), 437–455.
URL http://www.jstor.org/stable/30023824

Myerson, R. B., Jun 1978. Refinements of the nash equilibrium concept. International Journal of Game Theory 7 (2), 73–80.
URL https://doi.org/10.1007/BF01753236

Rees-Jones, A., 2017. sub-optimal behavior in strategy-proof mechanisms: Evidence from the residency match. Games Econ Behavior.
URL http://www.sciencedirect.com/science/article/pii/S0899825617300751

Repullo, R., 1985. Implementation in dominant strategies under complete and incomplete information. Review Econ Studies 52 (2), 223–229.
URL http://www.jstor.org/stable/2297618

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.
URL https://doi.org/10.1086/261272

Saijo, T., Sjöström, T., Yamato, T., 2007. Secure implementation. Theor Econ 2 (3), 203–229.
URL http://econtheory.org/ojs/index.php/te/article/view/20070203/0

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.
URL https://doi.org/10.1016/0022-0531(75)90050-2

Satterthwaite, M. A., Sonnenschein, H., 1981. Strategy-proof allocation mechanisms at differentiable points. Review Econ Studies 48 (4), 587–597.
URL http://www.jstor.org/stable/2297198

Schummer, J., Velez, R. A., 2019. Sequential preference revelation in incomplete information settings, Forthcoming Americal Economic Journal: Microeconomics.
URL https://sites.google.com/site/rodrigoavelezswebpage/home

Selten, R., Mar 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. International Journal of Game Theory 4 (1), 25–55.
URL https://doi.org/10.1007/BF01766400

Shapley, L., Scarf, H., 1974. On cores and indivisibility. J Math Econ 1 (1), 23 – 37.
URL http://dx.doi.org/10.1016/0304-4068(74)90033-0

Shorrer, R., Sóvágó, S., 2019. Obvious Mistakes in a Strategically Simple College Admissions Environment: Causes and Consequences, mimeo.
URL http://rshorrer.weebly.com/uploads/2/4/4/5/24450164/shs.pdf

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.
URL http://www.jstor.org/stable/2938268

Thomson, W., Oct 2016. Non-bossiness. Soc Choice Welfare 47 (3), 665–696.
URL https://doi.org/10.1007/s00355-016-0987-7

Tumennasan, N., 2013. To err is human: Implementation in quantal response equilibria. Games and Economic Behavior 77 (1), 138 – 152.
URL http://www.sciencedirect.com/science/article/pii/S0899825612001522

van Damme, E., 1991. Stability and Perfection of Nash Equilibria. Springer Berlin Heidelberg, Berlin, Heidelberg.
URL https://link.springer.com/book/10.1007/978-3-642-58242-4

Velez, R. A., Brown, A. L., 2019a. Empirical bias of extreme-price auctions: analysis.
URL http://arxiv.org/abs/1905.08234

Velez, R. A., Brown, A. L., 2019b. Empirical equilibrium.
URL https://arxiv.org/abs/1804.07986

Velez, R. A., Brown, A. L., 2019c. The paradox of monotone structural qre.
URL https://arxiv.org/abs/1905.05814

Appendix not for publication

Empirical strategy-proofness

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

January 8th, 2020

Proof of Theorem 3. Suppose that statement 1 is satisfied. Our argument in the proof of Theorem 2, taking σ as a Bayesian Nash 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 in welfare-outcome 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 $i \in N$,

$$u_i(g(\theta')|\theta_i) = u_i(g(\theta'_{-i}, \theta_i)|\theta_i). \tag{6}$$

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 a Bayesian Nash 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,

$$\varphi(\hat{m}_{-i}, m'_i) = g(\theta') \text{ and } \varphi(\hat{m}_{-i}, m_i) = g(\theta'_{-i}, \theta_i). \tag{7}$$

Thus, by (6),

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

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 $\hat{m}_i \in M_i$,

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

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

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

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

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

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

Finally, we show that statement 1 follows from statement 2. Let σ be a Bayesian Nash 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 that equation (1) in our proof of Theorem 1 holds when g is *strategy-proof* and *non-bossy in welfare-outcome*. Thus, for each $i \in N$, $g(\tau_{-i}, \theta_i) = g(\tau)$. Then, by the outcome rectangular property, we have that $g(\tau) = g(\theta)$. □