Mechanism Design with Limited Commitment*

Laura Doval† Vasiliki Skreta‡

December 10, 2021

Abstract

We develop a tool akin to the revelation principle for dynamic mechanism-selection games in which the designer can only commit to short-term mechanisms. We identify a *canonical* class of mechanisms rich enough to replicate the outcomes of any equilibrium in a mechanism-selection game between an uninformed designer and a privately informed agent. A cornerstone of our methodology is the idea that a mechanism should encode not only the rules that determine the allocation, but also the information the designer obtains from the interaction with the agent. Therefore, how much the designer learns, which is the key tension in design with limited commitment, becomes an explicit part of the design. Our result simplifies the search for the designer-optimal outcome by reducing the agent's behavior to a series of participation, truthtelling, and Bayes' plausibility constraints the mechanisms must satisfy.

KEYWORDS: *mechanism design, limited commitment, revelation principle, information design, short-term mechanisms*

JEL CLASSIFICATION: D84, D86

*We thank the Editor and four anonymous referees for excellent comments. We would also like to thank Rahul Deb, Françoise Forges, David Miller, Dan Quigley, Luciano Pomatto, Pablo Schenone, Omer Tamuz, and especially Michael Greinecker and Max Stinchcombe, as well as various audiences for thought-provoking questions and illuminating discussions. Alkis Georgiadis-Harris, Nathan Hancart, and Ignacio Núñez provided excellent research assistance. Vasiliki Skreta is grateful for generous financial support through the ERC consolidator grant 682417 "Frontiers in design." This research is supported by grants from the National Science Foundation (SES-1851744 and SES-1851729).

†Columbia University and CEPR. E-mail: laura.doval@columbia.edu

‡University of Texas at Austin, University College London, and CEPR. E-mail: vskreta@gmail.com

1 Introduction

The standard assumption in dynamic mechanism design is that the designer can commit to long-term contracts. This assumption is useful: it allows us to characterize the best possible payoff for the designer in the presence of adverse selection and/or moral hazard, and it is applicable in many settings. Often, however, this assumption is made for technical convenience. Indeed, when the designer can commit to long-term contracts, the mechanism-selection problem can be reduced to a constrained optimization problem thanks to the *revelation principle*.[1] However, as the literature starting with Freixas et al. (1985) and Laffont and Tirole (1988) shows, when the designer can commit only to short-term contracts, the tractability afforded by the revelation principle is lost. Indeed, mechanism design problems with limited commitment are difficult to analyze without imposing auxiliary assumptions either on the class of contracts available to the designer, as in Acharya and Ortner (2017) and Gerardi and Maestri (2020), or on the length of the horizon, as in Skreta (2006, 2015).

This paper provides a "revelation principle" for dynamic mechanism-selection games in which the designer can only commit to short-term mechanisms. We study a class of *mechanism-selection* games between an uninformed designer and an informed agent with persistent private information. Although the designer can commit within each period to the terms of the interaction–the current mechanism–he cannot commit to the terms the agent faces later on, namely, the mechanisms that are chosen in the continuation game. First, we identify a set of mechanisms, and hence *a* mechanism-selection game, that is sufficient to replicate any equilibrium outcome of any mechanism-selection game. Second, we identify a set of strategies for the designer and the agent that is sufficient to replicate any equilibrium outcome of the identified mechanism-selection game. We illustrate how our result can be used to characterize the designer's optimal outcome as the solution to a constrained optimization problem that only involves the designer. In this problem, the designer chooses amongst mechanisms that satisfy the usual truthtelling and participation constraints, and a third constraint, which captures the designer's sequential rationality.

The starting point of our analysis is the class of mechanisms we allow the designer to choose from. Following Myerson (1982) and Bester and Strausz (2007), we consider mechanisms as illustrated in Figure 1a:

[1] The "revelation principle" denotes a class of results in mechanism design; see, for instance, Myerson (1982).

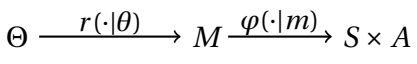

Figure (a) General Mechanisms

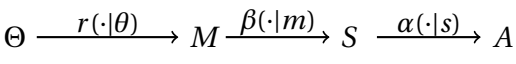

Figure (b) Canonical Mechanisms: $M = \Theta, S = \Delta(\Theta)$

Figure 1: Mechanisms

Having observed her private information (her type, $\theta \in \Theta$), the agent privately reports an *input* message, $m \in M$, to the mechanism, which then determines the distribution, $\varphi(\cdot|m)$, from which an *output* message, $s \in S$, and an allocation, $a \in A$, are drawn. The output message and the allocation are *publicly observable*: they constitute the contractible parts of the mechanism.

When the designer has commitment, the standard revelation principle implies that, without loss of generality, we can restrict attention to mechanisms satisfying three properties: (i) $M = \Theta$, (ii) $|M| = |S|$, and (iii) φ is such that by observing the output message, the designer learns the input message, in this case, the agent's type report. Moreover, the revelation principle implies we can restrict attention to equilibria in which the agent truthfully reports her type, which means the designer not only learns the agent's type report upon observing the output message but also learns the agent's true type.

Why restricting attention to mechanisms that satisfy properties (i)-(iii) and to truthtelling equilibria is with loss of generality under limited commitment is therefore clear: upon observing the output message, the designer learns the agent's type report and hence her type. Then, the agent may have an incentive to misreport if the designer cannot commit to not react to this information. This intuition is behind the main result in Bester and Strausz (2001), which is the first paper to provide a general analysis of optimal mechanism design with limited commitment. The authors restrict attention to mechanisms such that the cardinality of the set of input and output messages is the same, and φ is such that, by observing the output message, the designer learns the input message.[2] They show that to sustain payoffs in the Pareto frontier, mechanisms in which input messages are type reports are without loss of generality. However, focusing on truthtelling equilibria is with loss of generality. In a follow-up paper, Bester and Strausz (2007) lift the restrictions on the class of mechanisms (i.e., (ii) and (iii) above) and show in a one-period model that focusing on mechanisms in which input messages are type reports and on truthtelling equilibria is without loss of generality. The authors, however, do not characterize the set of output messages. Whether taking the set of input messages to be the set of type reports is without loss when the designer

[2]The class of mechanisms considered in Bester and Strausz (2001) encompasses the mechanisms considered by most papers in the literature on limited commitment starting from Laffont and Tirole (1988).

and the agent interact repeatedly is also unclear (see the discussion after Theorem 1).

The main contribution of this paper is to show that, under limited commitment, it is without loss of generality to take the set of output messages to be the set of the designer's posterior beliefs about the agent's type; that is, $S = \Delta(\Theta)$. Theorem 1 identifies a set of mechanisms, and hence a mechanism-selection game, that is enough to replicate any equilibrium outcome of any mechanism-selection game in which the designer chooses mechanisms as in Figure 1a. In this game, which we denote the *canonical game*, the designer can only offer mechanisms in which input messages are type reports and output messages are beliefs. Moreover, Theorem 1 shows that any equilibrium of the canonical game can be replicated by a *canonical equilibrium* in which the agent always participates in the mechanisms offered in equilibrium by the designer, and input and output messages have a *literal* meaning: the agent truthfully reports her type, and if the mechanism outputs a given posterior, this posterior coincides with the designer's equilibrium belief about the agent's type. Furthermore, in a canonical equilibrium, the designer only offers the agent *canonical mechanisms*, in which conditional on the output message, the allocation is drawn independently of the agent's type report (see Figure 1b). Thus, like the standard revelation principle, Theorem 1 implies that to characterize the distributions over types and allocations that can be achieved in some equilibrium in some mechanism-selection game, it is without loss of generality to restrict attention to the analysis of the canonical equilibria of the canonical game.

Theorem 1 provides researchers with a tractable way to analyze problems of mechanism design with limited commitment by making how much the designer learns about the agent an explicit part of the design. A major challenge in the received literature on limited commitment is how to keep track of how the agent's best response to the mechanism affects the information that the designer obtains from the interaction, which in turn affects the designer's incentives to offer the mechanism in the first place. Instead, our framework allows us to reduce the agent's best response to the designer's mechanism and its informational feedback to a familiar set of constraints that the mechanism must satisfy: the participation and incentive compatibility constraints for the agent, and the Bayes' plausibility constraint. This avoids having to consider complicated mixed strategies on the part of the agent (see Laffont and Tirole, 1988; Bester and Strausz, 2001) and transforms it instead into a program that combines elements of mechanism design and information design. Indeed, we exploit the information design elements to derive properties of the mechanisms (see Proposition 1 and our companion work, Doval and Skreta, 2020a, 2021).

We prove Theorem 1 under the assumption that the set of types is at most countable

and extend Theorem 1 to the case in which the agent's type is drawn from a continuum (a leading case in mechanism design) in Theorem 2. As we explain in Section 4.1, the results in Aumann (1961, 1964) imply that with a continuum type space we cannot rely on the usual formulation of an extensive-form game. This is the reason that we first conduct our analysis under the assumption that the set of types is at most countable, deriving Theorem 1 in the standard game-theoretic framework. Section 4.2 then develops a new framework that circumvents the issues raised by Aumann, while allowing us to extend Theorem 1 to continuum type spaces. The framework is based on the idea that any fixed sequence of mechanisms determines a well-defined extensive-form game for the agent. Like in the mechanism-selection game, Theorem 2 shows it is without loss of generality to assume the designer offers the agent sequences of canonical mechanisms and to restrict attention to canonical equilibria within the extensive-form game defined by the mechanisms.

We apply our results to a seemingly well-understood problem and show that our tools can shed new light on it.[3] As in Skreta (2006), Example 1 considers a seller, who owns one unit of a durable good, and interacts over two periods with a buyer with persistent and private information. In Example 1 the seller offers canonical mechanisms, whereas in Skreta (2006) the seller offers mechanisms in the class considered by Laffont and Tirole (1988) and Bester and Strausz (2001). In Section 3.1, we show how Theorem 1 can be used to reduce the characterization of the seller's maximum revenue to the characterization of the solution to a constrained optimization problem (see OPT). Furthermore, in Section 4.3, we argue that the solution to the program OPT also describes the seller's maximum revenue when the buyer's type is drawn from a continuum. We then show how to obtain the envelope representation of the agent's payoffs and hence, the dynamic virtual surplus representation of the seller's payoff. As we explain in Section 4.3, characterizing the solution to OPT is outside the scope of this paper. Instead, we use the virtual surplus representation of the seller's payoff to show that, in contrast to the main result in Skreta (2006), the seller can do strictly better than in the optimal posted-price mechanism.[4] Starting from the optimal posted-price mechanism, we show that the seller has a deviation to an alternative mechanism that combines posted prices with a form of *rationing*. This allows us to connect the mechanism design literature on the sale of a durable good with the work in theoretical industrial organization on alternative strategies for a durable goods monopolist, such as rationing (Denicolo and Garella,

[3] Theorem 1 also opens the door to the analysis of optimal mechanisms under limited commitment in infinite-horizon settings. See Doval and Skreta (2020a), where we solve an infinite-horizon binary-type version of the sale of a durable good.

[4] Remark 2 discusses how canonical mechanisms differ from the mechanisms in Skreta (2006), which explains the difference in the results.

1999).

By highlighting the canonical role of beliefs as the signals employed by the mechanism, Theorem 1 underscores the importance of jointly determining the mechanism together with how information is used in the mechanism and transmitted across periods. In doing so, it marries information design, which studies the design of information structures in a given institution, with mechanism design, which generally studies institutional design within a given information structure.

Related Literature: The paper contributes to the literature on mechanism design with limited commitment with an informed agent with persistent private information, referenced throughout the introduction.[5] Following the seminal contribution of Bester and Strausz (2001), a body of work studies optimal mechanisms under limited commitment in finite-horizon settings with finitely many types (e.g., Bisin and Rampini, 2006; Hiriart et al., 2011; Fiocco and Strausz, 2015; Beccuti and Möller, 2018). However, the results in Bester and Strausz (2001) do not extend to settings with a continuum of types and/or infinite horizon. On the one hand, the proof strategy in Bester and Strausz (2001) relies on the assumption of finitely many types. On the other hand, their result only applies if the designer is earning his highest payoff consistent with the agent's payoff (see Lemma 1 in Bester and Strausz, 2001). Thus, implicit in their multistage extension is a restriction to equilibria of the mechanism-selection game that possess a Markov structure, which, as shown by Ausubel and Deneckere (1989), may not be enough to characterize the designer's best equilibrium payoff in infinite-horizon settings. As a consequence, there is a small body of work that studies mechanism-selection games with a continuum of types and finite horizon (Skreta, 2006, 2015; Deb and Said, 2015), or in infinite-horizon settings, imposing restrictions on the class of contracts that can be offered (e.g., Acharya and Ortner, 2017; Gerardi and Maestri, 2020), or on the solution concept (e.g., Acharya and Ortner, 2017).

Due to the difficulties with the revelation principle, a large body of work in public finance, political economy, and taxation considers optimal time-consistent policies in settings where private information is fully non-persistent (see Sleet and Yeltekin, 2008; Farhi et al., 2012; Golosov and Iovino, 2021). Moreover, a large literature studies the effect of limited commitment within a specific class of mechanisms: price dynamics in the durable goods literature (Bulow, 1982; Gul et al., 1986; Stokey, 1981) and reserve price dynamics in auction settings (McAfee and Vincent, 1997; Liu et al., 2019).

[5] A designer's lack of commitment can take various forms that are not considered in this paper but have been studied in others. See, for instance, McAdams and Schwarz (2007), Vartiainen (2013), and Akbarpour and Li (2020), in which the designer cannot commit even to obeying the rules of the *current* mechanism.

By highlighting the role that the designer's beliefs about the agent play in mechanism design with limited commitment, our paper also relates to Lipnowski and Ravid (2020) and Best and Quigley (2017), who study models of direct communication between an informed sender and an uninformed receiver.

Organization: The rest of the paper is organized as follows. Section 2 describes the model and notation and Section 3 introduces the main theorem for at most countable type spaces. Section 4 extends our analysis to continuum type spaces. Throughout the paper, we use a two-period version of the model in Skreta (2006) to illustrate our results. Section 5 concludes and discusses further directions. Omitted statements and all proofs are in the appendix (Appendices A-C) and the online appendix, Doval and Skreta (2020b) (Appendices D-E).

2 Model

Primitives: Two players, a principal (he) and an agent (she), interact over $T \leq \infty$ periods.[6] Before the game starts, the agent observes her type, $\theta \in \Theta$, which is distributed according to a full-support distribution μ_1. We initially assume Θ is at most countable and consider continuum type spaces in Section 4. Each period, as a result of the interaction between the principal and the agent, an allocation $a \in A$ is determined. Let A^T denote the set $\times_{t=1}^{T} A$. When the agent's type is θ and the allocation is $a^T \in A^T$, the principal and the agent's payoffs are given by $W(a^T, \theta)$ and $U(a^T, \theta)$, respectively.

We allow for the possibility that past allocations influence what the principal can offer the agent in the future. Thus, for each $t \geq 1$, a correspondence $\mathcal{A}_t : A^{t-1} \rightrightarrows A$ exists such that for every sequence of allocations up to period t, $a^{t-1} = (a_1, \ldots, a_{t-1})$, $\mathcal{A}_t(a^{t-1})$ describes the set of allocations the principal can offer in period t (with the convention that when $t = 1$, $a^0 = \{\emptyset\}$). Furthermore, we assume an allocation a^* exists such that a^* is always available. Below, allocation a^* plays the role of the agent's outside option. Given the general structure of payoffs, it is without loss of generality to assume that a^* is time-independent.

We impose some technical restrictions on our model.[7] The sets Θ and A are Polish, that is, completely metrizable, separable, topological spaces. They are endowed with their Borel

[6] To simultaneously analyze the cases of finite and infinite horizon, we abuse notation as follows. When $T = \infty$, and notation of the form $t = 1, \ldots, T$, $\sum_{t=1}^{T}$, or $\times_{t=1}^{T}$, appears, we take this to mean $t = 1, \ldots$, $\sum_{t \in \mathbb{N}}$, or $\times_{t \in \mathbb{N}}$, respectively.

[7] In what follows, we adopt the following notational conventions. First, all Polish spaces are endowed with their Borel σ-algebra. Second, product spaces are endowed with their product σ-algebra. Third, for a Polish space, Y, we let $\Delta(Y)$ denote the set of all Borel probability measures over Y, endowed with the weak* topology. Thus, $\Delta(Y)$ is also a Polish space (Aliprantis and Border, 2006). For any two measurable spaces, X and Y, a transition probability from X to Y is a measurable function $\zeta : X \mapsto \Delta(Y)$. When integrating under the measure $\zeta(x)$, we use the notation $\zeta(\cdot|x)$.

σ-algebra. We also assume Θ is compact. Endowing product sets with their product σ-algebra, we assume the principal and the agent's utility functions, W and U, are bounded measurable functions. Similarly, for each $t \geq 1$ and for each $a^{t-1} \in A^{t-1}$, the set $\mathcal{A}_t(a^{t-1})$ is a measurable set.

Mechanisms: In each period, the allocation is determined by a mechanism $\mathbf{M}_t = (M^{\mathbf{M}_t}, S^{\mathbf{M}_t}, \varphi^{\mathbf{M}_t})$, where $M^{\mathbf{M}_t}$ and $S^{\mathbf{M}_t}$ are the mechanism's input and output messages and $\varphi^{\mathbf{M}_t}$ assigns to each $m \in M^{\mathbf{M}_t}$ a distribution over $S^{\mathbf{M}_t} \times A$. We endow the principal with a collection $\{(M_i, S_i)\}_{i \in \mathcal{I}}$ of input and output message sets, such that (i) M_i, S_i are Polish spaces, (ii) $|\Theta| \leq |M_i|$, M_i is at most countable, and (iii) $|\Delta(\Theta)| \leq |S_i|$. Moreover, we assume $(\Theta, \Delta(\Theta))$ is an element in that collection. Let $\mathcal{M}_{\mathcal{I}}$ denote the set of all mechanisms with message sets $(M_i, S_i)_{i \in \mathcal{I}}$ that is, $\mathcal{M}_{\mathcal{I}} = \cup_{i,j \in \mathcal{I}} \{\varphi : M_i \mapsto \Delta(S_j \times A) : \varphi \text{ is measurable}\}$.

Three remarks are in order. First, we restrict the principal to choosing mechanisms in $\mathcal{M}_{\mathcal{I}}$. This restriction allows us to have a well-defined strategy space for the principal, thereby avoiding set-theoretic issues related to self-referential sets.[8] The analysis that follows shows the choice of the collection plays no further role in the analysis. Second, because each M_i is at most countable, the set of mechanisms $\mathcal{M}_{\mathcal{I}}$ is a Polish space. As we discuss in Section 4, this property is key to being able to define a mechanism-selection game for a given collection $\mathcal{I}$ (see also footnote 9). Third, we note all aspects of the environment, except the agent's type $\theta \in \Theta$, are common knowledge between the principal and the agent.

Mechanism-selection game(s): Each collection $\mathcal{I}$ induces a mechanism-selection game, which we denote by $G_{\mathcal{I}}$, and is defined as follows. At the beginning of each period, both players observe the realization of a public randomization device, $\omega \sim U[0,1]$. The principal then offers the agent a mechanism, $\mathbf{M}_t$, with the property that for all $m \in M^{\mathbf{M}_t}$, $\varphi^{\mathbf{M}_t}(S^{\mathbf{M}_t} \times \mathcal{A}_t(a^{t-1})|m) = 1$, where recall that a^{t-1} describes the allocations implemented through period $t-1$. Observing the mechanism, the agent decides whether to participate in the mechanism ($\pi = 1$) or not ($\pi = 0$). If she does not participate in the mechanism, a^* is implemented and the game proceeds to period $t+1$. Instead, if she chooses to participate, she sends a message $m \in M^{\mathbf{M}_t}$, which is unobserved by the principal. An output message and an allocation (s_t, a_t) are drawn according to $\varphi^{\mathbf{M}_t}(\cdot|m)$. The output message and the allocation are observed by both the principal and the agent, and the game proceeds to period $t+1$.

Histories: The game $G_{\mathcal{I}}$ has two types of histories: public and private. Public histories capture what the *principal* knows through period t: the past realizations of the public ran-

[8] To be precise, the set of all mechanisms is not constructible under Zermelo-Fraenkel's set theory, with the axiom of choice, since its axioms preclude Russell's paradox.

domization device, his past choices of mechanisms, the agent's participation decisions, and the realized output messages and allocations. We let h^t denote a public history through period t and let H^t denote the set of all such histories. Instead, private histories capture what the *agent* knows through period t. First, the agent knows the public history of the game and her input messages into the mechanism (henceforth, the agent history). Second, the agent also knows her private information. We let h_A^t denote an agent's history through period t and let $H_A^t(h^t)$ denote the set of agent histories consistent with public history h^t. Thus, $\Theta \times H_A^t(h^t)$ denotes the set of private histories consistent with public history h^t.

Belief system and strategies: Private histories capture what the principal does not know about the agent in period t: he is uncertain about both the agent's payoff-relevant type, θ, and the agent history, h_A^t. Thus, a belief for the principal in period t at public history h^t is a distribution $\mu_t(h^t) \in \Delta(\Theta \times H_A^t(h^t))$. The collection $(\mu_t)_{t=1}^T$ denotes the belief system.

A behavioral strategy for the principal is a collection of measurable mappings $(\sigma_{Pt})_{t=1}^T$, where for each period t and each public history h^t, $\sigma_{Pt}(h^t)$ describes the principal's (possibly random) choice of mechanism at h^t.[9] Similarly, a behavioral strategy for the agent is a collection of measurable mappings $(\sigma_{At})_{t=1}^T \equiv (\pi_t, r_t)_{t=1}^T$, where for each period t, each private history (θ, h_A^t), and each mechanism, $\mathbf{M}_t$, $\pi_t(\theta, h_A^t, \mathbf{M}_t)$ describes the agent's participation decision, whereas $r_t(\theta, h_A^t, \mathbf{M}_t)$ describes the agent's choice of input messages in the mechanism, conditional on participation.

The tuple $(\sigma_P, \sigma_A, \mu) \equiv (\sigma_{Pt}, \sigma_{At}, \mu_t)_{t=1}^T$ defines an *assessment*.

Equilibrium: For each collection $\mathcal{I}$, we study the equilibria of the mechanism-selection game $G_{\mathcal{I}}$. By equilibrium, we mean *Perfect Bayesian equilibrium* (henceforth, PBE), informally defined as follows. An assessment $(\sigma_P, \sigma_A, \mu)$ is a PBE if it is sequentially rational and the belief system satisfies Bayes' rule where possible. The formal statement is in Appendix A. For now, we note that if the principal's strategy space is finite, Θ is finite, and the mechanisms used by the principal have finite support, our definition of PBE coincides with that in Fudenberg and Tirole (1991).[10]

Equilibrium outcomes: The prior μ_1 together with a strategy profile (σ_P, σ_A) determine a distribution over the terminal nodes $\Theta \times H_A^{T+1}$. We are interested instead in the distribution they induce over the payoff-relevant outcomes, $\Theta \times A^T$. We say $\eta \in \Delta(\Theta \times A^T)$ is a PBE

[9] Because the set $\mathcal{M}_{\mathcal{I}}$ is Polish, the public and private histories are Polish, because they are the (at most) countable product of Polish spaces.

[10] The only difference between Bayes' rule where possible (Definition A.2) and consistency in sequential equilibrium is that under PBE, the principal can assign zero probability to a type and then, after the agent deviates, can assign positive probability to that same type.

outcome if a PBE of the mechanism-selection game exists that induces η. We denote by $\mathcal{O}^*_{\mathcal{I}}$ the set of PBE outcomes of $G_{\mathcal{I}}$.

Throughout, we use the following example to illustrate the concepts in the paper:

Example 1. A seller (the principal) and a buyer (the agent) interact over two periods; that is, $T = 2$. The seller owns one unit of a durable good and assigns value 0 to it. The buyer's value for the good, denoted by $\theta \in \Theta$, is her private information. We denote by μ_1 the seller's prior belief over Θ and by $\overline{\theta}$ the maximum element of Θ. An allocation is a pair $(q, x) \in \{0,1\} \times \mathbb{R} \equiv A$, where q indicates whether the good is sold ($q = 1$) or not ($q = 0$), and x is a payment from the buyer to the seller. If the good is sold in period 1, the game ends; that is, $\mathcal{A}_2((1,x)) = \{0\} \times \mathbb{R}$ and $\mathcal{A}_2((0,x)) = A$. Moreover, if the buyer rejects the mechanism, the good is not sold and no payments are made; that is, $a^* = (0,0)$. Payoffs are as follows. If the final allocation is $\{(q_t, x_t)\}_{t\in\{1,2\}}$, the buyer and the seller's payoffs are $U(\cdot,\theta) = \sum_{t=1}^{T} \delta^{t-1}\left(q_t\theta - x_t\right)$ and $W(\cdot,\theta) = \sum_{t=1}^{T} \delta^{t-1} x_t$, respectively, where $\delta \in (0,1)$ is a common discount factor.

2.1 Canonical mechanisms and assessments

Theorem 1 singles out one mechanism-selection game and a class of assessments that allows us to replicate any equilibrium outcome of $G_{\mathcal{I}}$, for any collection $\mathcal{I}$ of input and output messages. We dub this extensive form the canonical game and the class of assessments, canonical assessments, which we formally define next.

Canonical game: The canonical game is the mechanism-selection game in which $\mathcal{I} = \{(\Theta, \Delta(\Theta))\}$. We denote the set of equilibrium outcomes of the canonical game by $\mathcal{O}^*$.

Definition 1 (Canonical Mechanisms)**.** *A mechanism* $(\Theta, \Delta(\Theta), \varphi)$ *is* canonical *if the mapping* $\varphi : \Theta \mapsto \Delta(\Delta(\Theta) \times A)$ *can be obtained as the* composition *of two mappings,* $\beta : \Theta \mapsto \Delta(\Delta(\Theta))$ *and* $\alpha : \Delta(\Theta) \mapsto \Delta(A)$*. Formally, for each* θ *and each pair of measurable subsets,* $\widetilde{U} \subset \Delta(\Theta)$ *and* $\widetilde{A} \subset A$*,* $\varphi(\widetilde{U} \times \widetilde{A}|\theta) = \int_{\widetilde{U}} \alpha(\widetilde{A}|\mu)\beta(d\mu|\theta)$*.*

In a canonical mechanism, conditional on the output message, the allocation is drawn *independently* of the agent's type report. We refer to the mappings β and α as the mechanism's disclosure and allocation rules, respectively. Interpreted as a statistical experiment, β encodes how much information the principal learns about the agent's type. Instead, α describes the mechanism's (possibly randomized) allocation, given the information that the principal learns about the agent's type. We let $\mathcal{M}_C$ denote the set of canonical mechanisms.

Remark 1 (Comparison with direct revelation mechanisms)**.** *A direct revelation mechanism is a special case of a canonical mechanism. To see this, recall that a direct revelation mechanism is a map* $\tilde{\alpha}$ *assigning to each type* θ *a distribution over allocations; that is,* $\tilde{\alpha} : \Theta \mapsto \Delta(A)$*.*

A direct revelation mechanism then corresponds to a canonical mechanism (β, α)*, where* β *assigns* θ *with probability* 1 *to the Dirac measure on* θ*,* δ_θ*, and then sets* $\alpha(\cdot|\delta_\theta) = \tilde{\alpha}(\theta)$*.*

Canonical assessments: A canonical assessment specifies behavior for the principal and the agent that is, in a sense, simple. First, the principal always chooses canonical mechanisms. Second, the agent best responds to the principal's equilibrium choice of mechanisms by participating. Third, input and output messages have *literal* meaning: Conditional on participating, the agent truthfully reports her type, and if the mechanism outputs $\mu \in \Delta(\Theta)$, μ coincides with the principal's updated beliefs about the agent's type. Formally:

Definition 2 (Canonical assessments)**.** *An assessment* $(\sigma_P, \sigma_A, \mu)$ *of mechanism-selection game* $G_{\mathcal{I}}$ *is canonical if the following holds for all* $t \geq 1$ *and all public histories* h^t*:*

1. *The principal offers canonical mechanisms, that is,* $\sigma_{Pt}(h^t)(\mathcal{M}_C) = 1$*.*
2. *For all mechanisms* $\mathbf{M}_t$ *in the support of the principal's strategy at* h^t*,*
   (a) *For all types* θ *in the support of the principal's beliefs in period* t*,* $\mu_t(h^t)$*,* $\pi_t(\theta, h_A^t, \mathbf{M}_t) = 1$*,*
   (b) *For all types* θ *in* Θ*,* $r_t(\theta, h_A^t, \mathbf{M}_t) = \delta_\theta$*, and*
   (c) *The mechanism's output belief* μ *coincides with the principal's updated belief about the agent's type. Formally, for* $h^{t+1} = (h^t, \mathbf{M}_t, 1, \mu, \cdot)$*, the marginal of* $\mu_{t+1}(h^{t+1})$ *on* Θ*,* $\mu_{t+1\Theta}(h^{t+1})$*, coincides with* μ*.*[11]
3. *The agent's strategy depends only on her private type and the public history.*[12]

We let $\mathcal{O}^C$ denote the set of equilibrium outcomes of the canonical game that are induced by canonical PBE assessments (henceforth, canonical PBE).

2.2 Discussion

We now discuss three aspects of the model that are important for what follows: the principal may offer randomized allocations, the principal and the agent have access to public randomization, and output messages are public.

Randomized allocations: Randomized allocations are necessary to conclude that without loss of generality output messages coincide with the principal's posterior beliefs about

[11] The 1 in $(h^t, \mathbf{M}_t, 1, \mu, \cdot)$ denotes the agent's decision to participate in mechanism $\mathbf{M}_t$.

[12] Whereas items 2a and 2b of Definition 2 imply the agent's strategy depends only on her private type and the public history *on the path* of the equilibrium strategy starting at h^t, item 3 implies this property also holds *off the path* of the equilibrium strategy starting at h^t, e.g., when the principal deviates and offers a mechanism not in the support of $\sigma_{Pt}(h^t)$.

the agent's type. Indeed, the principal could use $S^{\mathbf{M}_t}$ to encode randomizations on the allocation; for example, two tuples, (s_t, a_t) and (s'_t, a'_t), may be associated with the same posterior belief. Because a canonical mechanism allows the principal to randomize on the allocation conditional on the posterior belief, we can collapse s_t and s'_t to one output message (the posterior belief).

Public randomization: Public randomization allows us to subsume two ways in which the principal may use the mechanism to coordinate continuation play in a PBE of the mechanism-selection game that would otherwise not be possible in a canonical PBE.

First, in the mechanism-selection game, the principal could use $S^{\mathbf{M}_t}$ to coordinate continuation play; for example, two tuples $(s_t, a_t), (s'_t, a_t)$ may be associated with two different continuation equilibria, but with the same posterior belief. This is one place where the requirement that in a canonical PBE output messages must coincide with the principal's updated beliefs may be more restrictive than allowing for arbitrary PBE assessments: beliefs are not a rich enough language to encode both the principal's updated beliefs and the suggested continuation play.[13] As we explain after the statement of Theorem 1, the public randomization device in the canonical game allows us to subsume this *coordination* role of the output messages, which, in turn, allows us to conclude that without loss of generality, output messages coincide with the principal's posterior beliefs about the agent's type.[14]

Second, in the mechanism-selection game, the principal could use the "name" of the mechanism to coordinate continuation play. To be concrete, consider Figure 2a: In $t = 1$, the principal randomizes between two mechanisms, $\mathbf{M}$ and $\mathbf{M}'$, each of which is followed by different continuation play, denoted by continuation($\mathbf{M}$) and continuation($\mathbf{M}'$), respectively. To replicate this in a canonical PBE, we need to determine for each mechanism the distribution over $t = 1$-allocations induced by the mechanism together with the agent's strategy. Suppose when coupled with the agent's strategy both $\mathbf{M}$ and $\mathbf{M}'$ lead to the *same* canonical mechanism, $\mathbf{M}^C$. This is another way in which the language of canonical mechanisms is *coarser* than that of the mechanisms in $\mathcal{M}_{\mathcal{I}}$. In the mechanism-selection game, different indirect mechanisms can lead to different continuation equilibria, but to the same canonical

[13] Note this is not a matter of cardinality, but a consequence of the restriction that output beliefs must coincide with equilibrium beliefs. Ultimately, by Kuratowski's theorem (Srivastava, 2008), $\Delta(\Theta)$ is in bijection with $\Delta(\Theta) \times [0,1]$, so the number of messages is sufficient to encode both the beliefs and the suggested continuation play.

[14] The same idea arises in the literature on Bayesian persuasion. Implicit in the result in Kamenica and Gentzkow (2011) that any experiment can be written as a distribution over posteriors is the assumption that the receiver breaks ties in favor of the sender. Unlike in Bayesian persuasion, it is not clear that the players may be indifferent between two continuation equilibria, so the public randomization device does not generally reduce to simple tie-breaking.

mechanism. It is natural to conclude that, similar to the use of the public randomization device in the previous paragraph, we could replicate play in the mechanism-selection game via a canonical PBE in the canonical game as illustrated in Figure 2b. In the canonical game, the principal offers mechanism $\mathbf{M}^C$ in $t = 1$, and then we use the public randomization device in $t = 2$ to replicate the continuation play associated with $\mathbf{M}$ and $\mathbf{M}'$ in the original assessment.

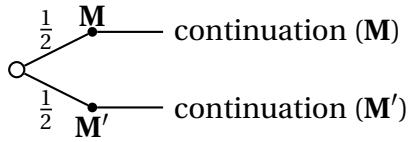

Figure (a) Mechanism-selection game

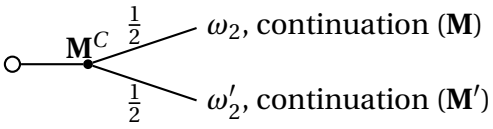

Figure (b) Canonical game: $t = 2$-public randomization

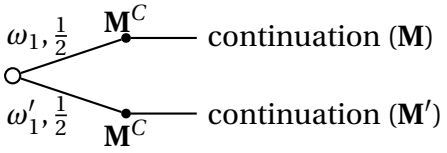

Figure (c) Canonical game: $t = 1$-public randomization

Figure 2: Different mechanisms lead to the same canonical mechanism

However, the construction in Figure 2b may not be enough to replicate the original outcome distribution. To see this, suppose in the mechanism-selection game, both $\mathbf{M}$ and $\mathbf{M}'$ are accepted with probability 1, so the principal's beliefs after non-participation are not pinned down by Bayes' rule. Furthermore, assume different off-path beliefs and different continuation equilibria are associated with the rejection of $\mathbf{M}$ and $\mathbf{M}'$. However, we can only assign *one* belief and *one* continuation equilibrium to the event in which the agent rejects $\mathbf{M}^C$ in Figure 2b. It may not be possible to find one off-path belief and one continuation equilibrium that simultaneously make it optimal for the agent to participate in $\mathbf{M}^C$ and sequentially rational for the principal to follow the prescribed continuation play.

Instead, we can replicate the original outcome distribution using the construction in Figure 2c: We use the public randomization device in $t = 1$ to encode the indirect mechanism that led to $\mathbf{M}^C$ in the mechanism-selection game. It follows that in order to replicate the outcome distributions via canonical PBE public randomization may be needed even before play begins.

While the proof of Theorem 1 deals explicitly with the coordination role of the output message, it does not deal explicitly with the the issue illustrated in Figure 2, which can only arise if the principal is using mixed strategies. Instead, we first show in Lemma D.1 that it is without loss of generality to assume that the principal plays a pure strategy in the mechanism-selection game. Similar to the construction in Figure 2c, we use the public randomization

device in the mechanism-selection game to encode the mechanisms over which the principal is randomizing, effectively purifying the principal's strategy.

Public output messages: Because in the mechanism-selection game the output message is public, the agent is the only player with private information, which is the leading informational setting in the literature on mechanism design with limited commitment and short-term contracts referenced in the introduction. Indeed, our only point of departure from this literature is the class of mechanisms we endow the principal with.

If, instead, output messages were observed only by the principal, he would become *endogenously* privately informed. Having an informed principal adds a new friction even in a one-shot interaction: the mere choice of a mechanism serves as a signal of the principal's private information. Little is known about dynamic and *exogenously* informed principal problems under commitment, let alone under limited commitment. For these reasons, the comparison of the equilibrium outcomes of the mechanism-selection game with public and private output messages is an open question.

3 Revelation principle

Section 3 presents the main result of the paper: To characterize the set of equilibrium outcomes that can arise in some mechanism-selection game, it is enough to characterize the canonical PBE outcomes of the canonical game. Formally,

Theorem 1 (Revelation principle)**.** *For any PBE outcome of any mechanism-selection game $G_{\mathcal{I}}$, an outcome-equivalent canonical PBE of the canonical game exists. That is,*

$$\bigcup_{\mathcal{I}} \mathcal{O}^*_{\mathcal{I}} = \mathcal{O}^* = \mathcal{O}^C.$$

Theorem 1 plays the same role in mechanism design with limited commitment as the revelation principle does in the commitment case. First, it identifies a well-defined set of mechanisms, $\mathcal{M}_C$, that, without loss of generality, the principal uses to implement any equilibrium outcome. Second, it simplifies the analysis of the behavior of the agent in the game induced by the mechanisms chosen by the principal: we can always restrict attention to assessments in which the agent participates and truthfully reports her type. As we illustrate through the application in Example 1, this restriction allows us to reduce the agent's behavior to a set of constraints that the mechanism must satisfy, as in the case of commitment.

The proof of Theorem 1 follows from two observations. The first one is easy: because the canonical game is a mechanism-selection game, any PBE outcome of the canonical game

is a PBE outcome of some mechanism-selection game; that is, $\mathcal{O}^* \subseteq \cup_{\mathcal{I}} \mathcal{O}^*_{\mathcal{I}}$. The second one constitutes the bulk of the proof, which we overview below: we show that for any collection $\mathcal{I}$ and for any PBE assessment of the mechanism-selection game $G_{\mathcal{I}}$, an outcome-equivalent canonical PBE assessment of $G_{\mathcal{I}}$ exists. Because in the canonical game the principal has fewer deviations than in $G_{\mathcal{I}}$ *and* in a canonical PBE assessment the principal plays a strategy that is available in the canonical game, it follows that for any PBE assessment of the mechanism-selection game, an outcome-equivalent canonical PBE assessment of the canonical game exists; that is $\mathcal{O}^*_{\mathcal{I}} \subseteq \mathcal{O}^C$. Because $\mathcal{O}^C \subseteq \mathcal{O}^*$, this concludes the proof.

We now review the steps involved in the proof that any PBE outcome of the mechanism-selection game $G_{\mathcal{I}}$ can be achieved in a canonical PBE of the canonical game.[15] To simplify the presentation, we rely on the following construction.[16] Given a mechanism $\mathbf{M}$ and the agent's participation and reporting strategies (π, r), we can extend the principal's mechanism and the agent's reporting strategy as follows. Extend the set of input messages so as to include a non-participation message, $M^{\mathbf{M}}_{\emptyset} = M^{\mathbf{M}} \cup \{\emptyset\}$. Similarly, extend the set of output messages, $S^{\mathbf{M}}_{\emptyset} = S^{\mathbf{M}} \cup \{\emptyset\}$ and define $\mathbf{M}_\pi = (M^{\mathbf{M}}_{\emptyset}, S^{\mathbf{M}}_{\emptyset}, \varphi^{\mathbf{M}}_\pi)$ as follows: $\varphi^{\mathbf{M}}_\pi$ coincides with $\varphi^{\mathbf{M}}$ on $M^{\mathbf{M}}$ and assigns probability 1 to $\{(\emptyset, a^*)\}$ for $m = \emptyset$. Finally, define r_π as follows: r_π sends message $\emptyset$ with probability 1-π, and sends message $m \in M^{\mathbf{M}}$ with probability $\pi r(m)$.

Input messages as type reports: To fix ideas, consider the proof for the standard revelation principle in static settings. The extended mechanism, $\mathbf{M}_\pi$, together with the agent's extended reporting strategy, induce a mapping from Θ to distributions over $S^{\mathbf{M}}_{\emptyset} \times A$. This allows us to conclude we can replace the set of input messages with the set of type reports, as illustrated in Figure 3a.

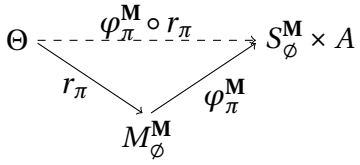

Figure (a) Static revelation principle

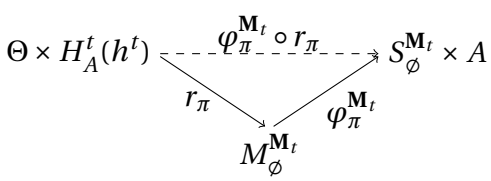

Figure (b) Mechanism-selection game

Figure 3: Type reports as input messages

In the dynamic setting, however, this argument would only allow us to conclude we can rewrite the mechanism as a mapping from $\Theta \times H^t_A(h^t)$ to distributions over $S^{\mathbf{M}_t}_{\emptyset} \times A$, as illus-

[15] Appendix B proves Theorem 1 under the assumption of finitely many types and that the principal only offers mechanisms $\mathbf{M}$ such that for all input messages $m \in M^{\mathbf{M}}$, the support of $\varphi^{\mathbf{M}}(\cdot|m)$ is finite. This simple proof mirrors the complete one in Online Appendix D.1, but it is technically simpler and more accessible.

[16] This construction is the standard one whereby the non-participation decision is included as an option in the mechanism. At the end of the overview, we address why we model participation as a separate decision.

trated in Figure 3b. Indeed, to replicate the agent's reporting strategy, the mechanism needs to obtain all the information on which the agent conditions her strategy, which potentially is (θ, h_A^t).

However, we show that given a PBE in which the agent conditions her strategy on the payoff-irrelevant part of her private history at some public history h^t, another outcome-equivalent PBE exists in which she does not (see Proposition B.1).[17] Thus, conditional on the public history h^t, the mechanism, together with the agent's reporting strategy, induces a mapping from Θ to distributions over $S_{\emptyset}^{\mathbf{M}_t} \times A$, so we can always take the set of input messages to be the set of type reports. This result relies on two observations. First, because input messages are payoff irrelevant and unobserved by the principal, if the agent chooses different strategies at (θ, h_A^t) and $(\theta, \tilde{h}_A^t)$ with $h_A^t, \tilde{h}_A^t \in H_A^t(h^t)$, she is indifferent between these two strategies. However, the principal may not be indifferent between these two strategies. Second, we build an alternative strategy for the agent that conditions only on (θ, h^t) and yields the same outcome distribution, and hence the same payoff for the principal.

Two implications follow from this step. First, given a public history h^t and the principal's choice of mechanism at h^t, $\mathbf{M}_t$, the *auxiliary* mapping, $\varphi_\pi^{\mathbf{M}_t} \circ r_\pi$, only takes Θ as an input. The mapping $\varphi_\pi^{\mathbf{M}_t} \circ r_\pi$ is a *direct* mechanism, where the set of input messages are type reports, and for which the agent finds it optimal to truthfully report her type. Second, the relevant part of the principal's beliefs in the game are about the agent's type, θ, and not the agent's private history, h_A^t. This finding is important because although the auxiliary mapping allows us to replicate the distribution over period-t outcomes induced by the agent's strategy and the mechanism $\mathbf{M}_t$, in a dynamic game, we also need to replicate the distribution over continuation play, and the principal's beliefs are an important component of continuation play.

Output messages as beliefs and canonical mechanisms: As discussed in Section 2.2, the principal may have other uses for the output messages beyond encoding information about the agent: Conditional on the same posterior belief, (i) he may encode randomizations over the allocation, and (ii) he may use the output message to coordinate continuation play. As the proof of Theorem 1 shows, randomized allocations and public randomization allow us to subsume these two roles of the output messages.

Now, the potential challenge in using the public randomization device to subsume the

[17] This result is useful also in applications. It states that in our game, the set of PBE payoffs coincides with the set of public PBE payoffs (Athey and Bagwell, 2008). In games with time-separable payoffs, public PBE payoffs are amenable to self-generation techniques, as in Abreu et al. (1990). See Doval and Skreta (2020a).

second role of the output message is that, by definition, the use of the public randomization device in the canonical game can only depend on publicly available information. Instead, because in the auxiliary mapping the agent is reporting truthfully, the output message in the mechanism-selection game is drawn as a function of the agent's type, which allows the principal to coordinate future play above and beyond what he would be able to do by solely relying on the public randomization device in the canonical game.

To overcome this challenge, we leverage here that canonical mechanisms use *beliefs* as output messages. Note that beliefs are a sufficient statistic for the information about the agent's type that is encoded in the output messages of the mechanism-selection game. Thus, conditional on the induced belief and the allocation, the selection of continuation play contains no further information about the agent's type, which allows us to *decompose* the mechanism in the mechanism-selection game into a canonical mechanism in the canonical game that uses beliefs as output messages and a public randomization device.

The above argument also explains why, in a canonical mechanism, conditional on the output message, the allocation can be drawn independently of the agent's type (report). Ultimately, conditional on the induced belief, the allocation contains no further information about the agent's type.

Briefly, the proof of this result proceeds as follows (see Proposition B.2). Suppose that the principal offers $\mathbf{M}_t$ in period t. The principal's belief about the agent's type, $\mu_t(h^t)$, together with the auxiliary mapping, $\varphi_\pi^{\mathbf{M}_t} \circ r_\pi$, induces a joint distribution over $\Theta \times S_\emptyset^{\mathbf{M}_t} \times A \times \Delta(\Theta) \times \Omega$, where $\Omega = [0,1]$ denotes the public randomization device in the mechanism-selection game. Because conditional on the induced posterior, $(s_t, a_t) \in S_\emptyset^{\mathbf{M}_t} \times A$ carries no further information about the agent's type, this allows us to "split" the mechanism into a transition probability β from Θ to $\Delta(\Theta)$, a transition probability α from $\Delta(\Theta)$ to A, and a transition probability ω from $\Delta(\Theta) \times A$ to $S_\emptyset^{\mathbf{M}_t} \times \Omega$. The transition probability α plays the first role of the output message and highlights the importance of allowing the principal to offer randomized allocations. The transition probability ω corresponds to the public randomization device: by Kuratowski's theorem, we can always embed $S_\emptyset^{\mathbf{M}_t} \times \Omega$ into $[0,1]$ (see Srivastava, 2008).

Three conceptual insights arise from this result. First, when the mechanism is canonical, the principal can separate the design of the information that the mechanism encodes about the agent's type from the design of the allocation. Second, the allocation has to be measurable with respect to the information generated by the mechanism: The more the principal desires to tailor the allocation to the agent's type, the more he has to learn about the agent's

type through the mechanism.[18] Third, it highlights the *coordination* role of the mechanism, which is subsumed by the public randomization device: beyond determining today's allocation and the information that is carried forward in the interaction, the mechanism allows the principal to coordinate future play.

Bayes' rule, truthtelling, and participation: Underlying the previous step is the assumption that the beliefs associated with the output messages are determined via Bayes' rule. In particular, the principal is never surprised by any output message he observes. To ensure that beliefs are pinned down by Bayes' rule, Proposition B.2 shows we can "eliminate" from the mechanism all input messages that are used only by types to whom the principal assigns 0 probability. Eliminating these input messages, however, may change the participation decision for types not in the support of the principal's beliefs, which is why a canonical PBE assessment does not require that these types participate in the mechanism.[19]

Instead, it should be intuitive that the agent participates in the mechanism whenever her type is in the support of the principal's beliefs: By relying on the map between output messages and posterior beliefs and the public randomization device, we guarantee that, conditional on participation, the agent faces the same period-t allocation and distribution over continuations as when she did not participate. The map between output messages and beliefs allows us to identify which output message one should associate with the types that chose not to participate: the one that corresponds to the principal's updated belief conditional on non-participation.[20] The map between output messages and the public randomization device allows us to replicate the distribution over continuations the agent faces in the PBE of the mechanism-selection game for those types that found it optimal to randomize between participating and not participating.

That beliefs in the support of the mechanism are determined via Bayes' rule has one practical implication that we exploit throughout our analysis of Example 1 and in our concurrent work (Doval and Skreta, 2020a, 2021): the mechanism's disclosure rule together with

[18] Contrast this case with the one in which the principal has commitment, where we write a mechanism as a menu of options, one for each type of the agent. We do so even if the optimal mechanism offers the same allocation to a set of agent types. When the principal has commitment, whether the allocation reveals more information beyond the set of types that receive that allocation is irrelevant, because additional information can always be ignored. Under limited commitment, however, this is not the case, and the principal in general trades off tailoring the allocation to the agent's type and the information that is learned through this.

[19] If an agent's type has zero probability at a specific public history, she can only reach it through a deviation from σ_A. This change to the mechanism actually makes the deviation less attractive, and hence, it "disincentivizes" the agent from deviating in the first place.

[20] Starting from an equilibrium in which the mechanism is rejected with positive probability, this belief is also determined via Bayes' rule.

the principal's belief about the agent's type induce a Bayes' plausible distribution over posteriors. As a consequence, we can apply tools from information design to derive qualitative properties of the principal's problem (see Proposition 1). This is where modeling participation as a decision that happens outside the mechanism as opposed to as an input message that locks the outside option as in Figure 3b is important: because non-participation is a zero-probability event, the principal's beliefs after non-participation are not pinned down via Bayes' rule, and hence cannot be necessarily obtained from a Bayes' plausible distribution over posteriors. At the same time, these beliefs cannot be ignored in the analysis, because they determine the continuations after non-participation, and hence, the agent's incentives to participate in the first place. By modeling participation as a decision that happens outside of the mechanism, we can rely on the tools of information design to design the mechanism's disclosure rule, while, as we illustrate using the application in Example 1, the participation decision is summarized by a participation constraint (see OPT).

3.1 The revelation principle at work

We now illustrate the simplifications afforded by Theorem 1 within the context of the application in Example 1. In particular, we show that in order to characterize the seller's revenue maximizing PBE outcome it is enough to characterize the solution to a constrained optimization problem, denoted OPT, that only involves the seller. This is already in stark contrast to the existing work in mechanism design with limited commitment, which needs to keep track of how the buyer's best response to the seller's mechanism determines the information that the seller obtains from the interaction, which, in turn, affects the seller's incentives to offer the mechanism in the first place.

To arrive at the program that characterizes the seller's maximum revenue, we appeal to Theorem 1. First, in what follows, we restrict attention to the canonical game and to assessments in which the seller offers canonical mechanisms. Second, without loss of generality, we can consider assessments where the buyer's strategy does not depend on the payoff-irrelevant part of the private history. In particular, in $t = 2$, the seller's optimal mechanism only needs to elicit the buyer's payoff relevant type, θ. Let μ_2 denote the seller's *posterior* belief in $t = 2$. The optimal mechanism in $t = 2$ is a posted price *regardless* of the properties of μ_2 (see Proposition 2 in Skreta, 2006). For each belief the seller may have in $t = 2$, μ_2, we let $p_2(\mu_2)$ denote a selection from the set of optimal prices in $t = 2$ when his belief is μ_2.[21]

[21]The seller may be indifferent in $t = 2$ among several prices. In that case, as in the literature on Bayesian persuasion, we determine the tie-breaking rule as a solution to the seller's problem in $t = 1$ (see OPT). Note that because of public randomization the selection from the set of optimal prices in $t = 2$ at belief μ_2 may be

Third, it is without loss of generality to consider assessments in which (i) the buyer's best response to the seller's equilibrium choice of mechanism in $t = 1$ is to participate and truthfully report her type with probability 1, and (ii) when the output message is μ_2, the seller updates his belief to μ_2. Moreover, the assumption of quasilinearity implies that, without loss of generality, the seller does not randomize on the transfers: below $x(\mu_2)$, denotes the expected payment conditional on μ_2, and $q(\mu_2) \in [0,1]$ denotes the probability with which the good is sold. Thus, we can write the seller's problem as follows:

$$\max_{(q,\beta,x,p_2)} \mathbb{E}_{\mu_1}\left[\int_{\Delta(\Theta)} \left(x(\mu_2) + (1-q(\mu_2))\delta\, p_2(\mu_2)\mathbb{1}[\theta \geq p_2(\mu_2)]\right)\beta(d\mu_2|\theta)\right] \tag{OPT}$$

$$\text{s.t.} \quad (\forall \theta \in \Theta) U(\theta) = \int_{\Delta(\Theta)} \left(\theta q(\mu_2) - x(\mu_2) + (1-q(\mu_2))\delta\, u^*(\theta,\mu_2)\right)\beta(d\mu_2|\theta) \geq 0 \tag{PC}$$

$$(\forall \theta, \theta' \in \Theta) U(\theta) \geq \int_{\Delta(\Theta)} \left(\theta q(\mu_2) - x(\mu_2) + (1-q(\mu_2))\delta\, u^*(\theta,\mu_2)\right)\beta(d\mu_2|\theta') \tag{IC}$$

$$(\forall \theta \in \Theta)(\forall \widetilde{U} \subset \Delta(\Theta)) \int_{\widetilde{U}} \mu_2(\theta)\left(\mathbb{E}_{\tilde{\theta}}\left[\beta(d\mu_2|\tilde{\theta})\mu_1(\tilde{\theta})\right]\right) = \beta(\widetilde{U}|\theta)\mu_1(\theta). \tag{BP}$$

That the seller's belief about the buyer's type updates to μ_2 when the output message is μ_2 appears twice in the above expression: First, in Equation BP, which is the Bayes' plausibility constraint and, second, in the objective function, where the seller's payoff in $t = 2$ when the agent's type is θ and his belief is μ_2 corresponds to whether θ buys the good at a price of $p_2(\mu_2)$. The latter affords an important simplification: instead of writing the seller's program as one in which the seller chooses a mechanism for period 1 and one for period 2 subject to the constraint that the period 2 mechanism is optimal given the seller's belief in $t = 2$, the program OPT has the seller maximize over one-period mechanisms by replacing the seller's best response in $t = 2$ in the seller's objective function.

The two remaining constraints in OPT are the buyer's participation and incentive compatibility constraints (Equations PC and IC). The buyer's payoff in the mechanism, $U(\theta)$, is determined as follows. For each μ_2 in the support of $\beta(\cdot|\theta)$, she receives the good with probability $q(\mu_2)$ and makes a payment of $x(\mu_2)$; with the remaining probability, no trade occurs, and she obtains a continuation payoff, $u^*(\theta,\mu_2)$, which describes her optimal decision of whether to buy the good at $p_2(\mu_2)$. The participation constraint states that the buyer has to earn a payoff of at least 0 by participating. Indeed, because non-participation is a 0 probability event, we can specify that upon rejection of the mechanism, the seller believes the buyer's valuation is $\overline{\theta}$, so that in $t = 1$, the seller chooses a price of $\overline{\theta}$ when the buyer chooses not to participate. The incentive compatibility constraint states that when her type is θ, the

random. However, our notation does not account for this explicitly to simplify the presentation.

buyer cannot obtain a higher payoff by reporting that her type is $\theta' \neq \theta$. When the buyer reports θ', she obtains a different distribution over output messages $\beta(\cdot|\theta')$; however, in $t = 2$, she still chooses optimally whether to buy the good, which explains the term $u^*(\theta, \mu_2)$.

The three constraints in OPT provide us with a tractable representation of both the buyer's behavior and its impact on the mechanism offered in $t = 2$ via the information that is generated about the buyer's type in $t = 1$. Thus, instead of having to consider complicated mixed strategies on the part of the agent (see Laffont and Tirole, 1988; Bester and Strausz, 2001), we have reduced the problem of characterizing the seller-optimal PBE outcome to the solution of a program OPT that combines elements of information design and mechanism design. Indeed, the solution to OPT can be leveraged to fully specify the PBE assessment that implements the seller's maximum revenue.[22]

Furthermore, as we show below in Proposition 1, when Θ is finite, OPT can be further simplified: without loss of generality we can assume the seller employs mechanisms such that the support of β is finite for all $\theta \in \Theta$.

Proposition 1. *Suppose Θ is finite. Fix $\mathcal{I}$ and let $(\sigma_P, \sigma_A, \mu)$ denote a canonical PBE assessment of $G_{\mathcal{I}}$. Then, a payoff-equivalent canonical PBE assessment exists such that for all $t \geq 1$ and all h^t, the principal's choice of mechanism at h^t, $\mathbf{M}_t$, satisfies that for all $\theta \in \Theta$, the support of $\beta^{\mathbf{M}_t}(\cdot|\theta)$ is finite.*

The proof of Proposition 1 highlights that Equation BP implies that the mechanism's disclosure rule, β, induces a Bayes' plausible distribution over posteriors. Like in the literature on Bayesian persuasion, we can then rely on Carathéodory's theorem (Rockafellar, 1970) to ensure that the principal and the agent's payoffs remain the same should the principal use a mechanism that employs finitely many posteriors.

Most of the existing analysis of the model in Example 1 is performed for continuum type spaces; we thus revisit OPT when Θ is a continuum in Section 4.3. Before doing so, we first explain why it is impossible to endow the set of mechanisms with a measure structure so that the game is well-defined and, then, develop a framework in Section 4.2 that is suitable to study mechanism design under limited commitment with continuum types spaces.

4 Continuum type spaces

Section 4 considers the case in which the set of types is an uncountable compact Polish space. This extension is important because much of the standard toolkit of mechanism de-

[22] For an illustration of this, see our companion work, Doval and Skreta (2020a).

sign has been developed for continuum (and convex) type spaces, where the representation of incentive compatible mechanisms can be obtained using the envelope theorem. Section 4.1 reviews the issues raised by Aumann (1961), and hence the difficulties with having a well-defined mechanism-selection game when Θ is uncountable. In particular, we explain why the usual solution to this problem, namely, restricting the principal to choosing mechanisms in a suitably defined set, is not enough for the purpose of deriving a revelation principle. With little loss of continuity, the reader can skip to Section 4.2, where we propose a framework that allows us to sidestep the aforementioned issues and obtain the analogue of Theorem 1 when Θ is uncountable. We then apply our results to Example 1.

4.1 Choosing functions at random

To define the mechanism-selection game, a measurable structure on $\mathcal{M}_{\mathcal{I}}$ is needed to define (i) the principal's mixed strategies, (ii) the principal and the agent's expected payoffs from those mixed strategies, and (iii) the principal and the agent's strategies as measurable functions of the histories, which include the past choices of the mechanisms. Focusing on (i) and (ii), Aumann (1961, Theorem D) implies no suitable measure structure on $\mathcal{M}_{\mathcal{I}}$ exists when the set of input messages is uncountable. Instead, Aumann (1964) circumvents the issue of defining a measurable structure on the set $\mathcal{M}_{\mathcal{I}}$ to define mixed strategies by relying on randomization devices. However, the construction in Aumann (1964) is insufficient for our purposes because the mechanisms chosen by the principal through period $t-1$ are part of the public histories. Thus, to define the principal and the agent's strategies as measurable functions of the histories, we again face the issue of defining a measurable structure on the set of mechanisms and with the negative answers in Aumann (1961).

For this reason, the literature on competing principals (see, e.g. Attar et al., 2021a,b) follows a different approach: Theorem D in Aumann (1961) implies that the issues raised above would be mute if we restrict the principal to choosing mechanisms from a subset $\mathcal{M}'$ of $\mathcal{M}_{\mathcal{I}}$, such that $\mathcal{M}'$ is of bounded Borel class.[23] For the purposes of deriving a revelation principle, this approach is again insufficient: Borel classes are not always closed under composition (Srivastava, 2008) and we obtain a canonical mechanism by composing the agent's strategy with the mechanism the principal employs in the mechanism-selection game. Unless the agent's strategy is continuous in her type, the induced canonical mechanism may be of a Borel class strictly larger than that of the original mechanism. Furthermore, different equilibria of the mechanism-selection game may necessitate canonical mechanisms of

[23] For uncountable M, $\{\varphi : M \rightarrowtail \Delta(S \times A) : \varphi \text{ is measurable}\}$ is not of bounded Borel class (Aumann, 1961).

different Borel classes, which makes the task of defining the canonical game pointless: one would have to potentially consider a different canonical game for each equilibrium of each mechanism-selection game. Lastly, the restriction to a set of mechanisms of bounded Borel class is difficult to work with in applications: only payoffs from deviations to mechanisms within that class are well-defined and, in practice, verifying that only deviations to those mechanisms are being contemplated is difficult.

Motivated by the importance of continuum type spaces, Section 4.2 proposes an approach to model mechanism-selection games that circumvents the above issues.

4.2 PBE-feasible outcomes

We develop a framework to characterize the outcomes that can be sustained under limited commitment, which we dub PBE-feasible outcomes and formally define below (Definition 5). By analogy with the mechanism-selection game, we keep the notation $\mathcal{O}^*_{\mathcal{I}}$ to denote PBE-feasible outcomes. Contrary to the mechanism-selection game, $\mathcal{O}^*_{\mathcal{I}}$ is now a correspondence describing the set of PBE-feasible outcomes for each period $t \geq 1$, each principal's belief μ_t and each sequence of allocations up to period t, a^{t-1}, $\mathcal{O}^*_{\mathcal{I}}(\mu_t, a^{t-1})$. The reason is that the definition of the set of PBE-feasible outcomes is recursive, and what is PBE-feasible in period t naturally depends on what is PBE-feasible in period $t+1$.

The discussion in Section 4.1 implies that the framework must sidestep the need to define the principal and the agent's "strategies" as measurable functions of the principal's past choices of mechanisms. There are (at least) two important roles measurability plays in the mechanism-selection game. First, it allows us to describe the agent's behavior along the path of the principal's strategy, which in turn allows us to evaluate the principal's payoff from a given strategy. Second, it allows us to describe how the agent's behavior changes when the principal deviates from the prescribed strategy, which in turn allows us to evaluate the principal's payoff from a deviation from the prescribed strategy. The comparison between these two payoffs determines whether the principal's strategy is sequentially rational.

Informally, the framework circumvents the aforementioned measurability issues as follows. The starting point of the analysis is that we no longer model the principal as a player. Instead, we describe the analogue of a principal's strategy in the mechanism-selection game via an extensive-form game for the agent, which describes the sequence of mechanisms the agent faces as a function of her participation decisions and the outcomes of the mechanisms (Definition 3). Importantly, in this extensive-form game the agent's strategy only depends on her type, her participation decisions, her input messages, and the outcomes of

the mechanisms, but not on the mechanisms themselves. The principal's belief about the agent's type together with the agent's strategy define a distribution over the terminal nodes in this extensive form, which allows us to evaluate the principal's payoff from a particular sequence of mechanisms. The final component of the framework is how we define that a particular extensive form is *sequentially rational* for the principal; in other words, that the principal does not wish to revise the mechanisms that define the extensive form at any point in time (Definition 4). Here we rely on two ideas. First, to determine whether the principal wishes to revise the mechanisms that define the extensive form, we only need to know the outcome distributions the principal expects he will face in the event of a deviation (Equation 1). Second, these outcome distributions can be defined without reference to a strategy of the agent that conditions on the sequence of mechanisms that has been offered so far.

We now formally define the set of PBE-feasible outcomes, $\mathcal{O}^*_{\mathcal{I}}(\mu_t, a^{t-1})$, for each period $t \geq 1$, and pair $(\mu_t, a^{t-1}) \in \Delta(\Theta) \times A^{t-1}$ (Definition 5). Because the definition is recursive, we fix a period $t \geq 1$, and a pair (μ_t, a^{t-1}) throughout. Definition 5 consists of three components, which we introduce first: (i) the sequence of mechanisms offered by the principal (Definition 3), (ii) optimal behavior by the agent within those mechanisms, and (iii) the outcome distributions the principal anticipates upon a deviation (Equation 1). For simplicity, we assume the principal has one set of input and output messages, that is, $\mathcal{I} = \{(M, S)\}$, and we use the shorthand notation $SA_\emptyset$ to denote the set $(S \times A) \cup \{(\emptyset, a^*)\}$.

Dynamic mechanisms: Instead of having the principal be a player in a game, we describe the analogue of the principal's strategy via a *dynamic mechanism*, defined as follows:

Definition 3 (Dynamic mechanisms)**.** *For $t \geq 1$ and $a^{t-1} \in A^{t-1}$, a dynamic mechanism given a^{t-1}, $(\varphi_\tau)_{\tau \geq t}$, is a sequence of measurable mappings $\varphi_\tau : (SA_\emptyset \times \Omega)^{\tau - t} \times M \mapsto \Delta(S \times A)$, such that for all $\tau \geq t$ and all $(s^{\tau-t}, a^{\tau-t}, \omega^{\tau-t})$,*

1. *$\varphi_\tau(s^{\tau-t}, a^{\tau-t}, \omega^{\tau-t}, \cdot) : M \mapsto \Delta(S \times A)$ is a measurable function, and*
2. *for all $m \in M$, $\varphi_\tau(s^{\tau-t}, a^{\tau-t}, \omega^{\tau-t})\left(\mathcal{A}_t(a^{t-1}, a^{\tau-t}) | m\right) = 1$.*

When $t = 1$, a dynamic mechanism describes the mechanism the agent faces in period 1, φ_1, the mechanism the agent faces in period 2 as a function of the agent's participation decision (i.e., whether $(s_1, a_1) \neq (\emptyset, a^*)$), and the realization of the public randomization device, $\varphi_2(s_1, a_1, \omega_2)$, and so on. Consider now $t > 1$ and suppose the allocation so far is a^{t-1}. Then, we require that the dynamic mechanism only implements allocations that are feasible given a^{t-1}.

As we explain next, a dynamic mechanism defines an extensive-form game for the agent:

Agent-extensive form: Given (μ_t, a^{t-1}), a dynamic mechanism $(\varphi_\tau)_{\tau\geq t}$ defines an extensive-form game for the agent, $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau\geq t})$, as follows. First, nature draws the agent's type according to μ_t. Having observed her type, suppose that in stage $\tau - t$, the public history is $h_t^\tau = (s^{\tau-t}, a^{\tau-t}, \omega^{\tau-t})$. Then, faced with $\varphi_\tau(h_t^\tau)$, the agent decides whether to participate and, conditional on participating, her reporting strategy. If the agent rejects φ_τ, the "output message" is $\emptyset$ and the allocation is a^*. Instead, if she accepts $\varphi_\tau(h_t^\tau)$, she chooses an input message $m \in M$ that determines the distribution from which the output message and the allocation are drawn, $\varphi_\tau(h_t^\tau)(\cdot|m)$. In both cases, we proceed to stage $\tau + 1 - t$.

In the agent-extensive form $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau\geq t})$, there are two types of histories. The public history h_t^τ encodes the agent's participation in the mechanism, the realized output messages and allocations, and the realizations of the public randomization device. The private histories encode everything the agent knows: her payoff-relevant type θ, the public history h_t^τ, and her past input messages. In a slight abuse of notation, we denote by $H_{At}^\tau(h_t^\tau)$ the set of agent histories consistent with h_t^τ.

Importantly, we do not need to encode the mechanisms defining $(\varphi_\tau)_{\tau\geq t}$ in the histories of the agent-extensive form, because the mechanisms are *akin* to a move by nature in the extensive form $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau\geq t})$. It thus follows that in the agent-extensive form we can define the agent's strategy, σ_A, as a measurable function of the private histories.

Finally, we note that the agent evaluates the payoffs of a strategy in $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau\geq t})$ using $U(a^{t-1}, \cdot, \theta)$. We are now ready to define optimal play by the agent in $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau\geq t})$:

Agent-PBE: Together with the agent strategy σ_A, we can also define a system of beliefs $\mu \equiv (\mu_\tau)_{\tau\geq t}^T$, which describes for each period $\tau \geq t$ and for each public history h_t^τ, the principal's beliefs over the private histories, $\mu_\tau(h_t^\tau) \in \Delta\left(\Theta \times H_{At}^\tau(h_t^\tau)\right)$.

We say that (σ_A, μ) is an agent-PBE of the agent-extensive form $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau\geq t})$ if the agent's strategy is sequentially rational (under payoffs $U(a^{t-1}, \cdot)$) and the belief system satisfies Bayes' rule where possible. Although the belief system is not needed to test whether the agent's strategy is optimal in the extensive-form game, it is needed to test the optimality of the principal's choice of mechanism.

Proposition B.1 applies verbatim, allowing us to conclude that for every agent-PBE (σ_A, μ) of $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau\geq t})$, an outcome equivalent (σ'_A, μ') exists, in which the agent's strategy only conditions on her type and the public history. This property is responsible for the recursive nature of the set of PBE-feasible outcomes here and also in the mechanism-selection game. Hereafter, when we say agent-PBE, we mean one that satisfies the above property.

(Continuation) Outcome distributions: An agent-PBE (σ_A, μ) of $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau \geq t})$ defines a distribution over $\Theta \times A^T$, $\eta^{(\varphi_\tau)_{\tau \geq t}, \sigma_A}$, that satisfies two conditions. First, the marginal on Θ is μ_t. Second, the distribution is supported on those tuples (θ, a^T) such that a^T coincides with a^{t-1} through $t-1$. (Online Appendix E.1 contains the formal definition.)

Furthermore, at any history h_t^τ, the belief assessment together with the dynamic mechanisms and the agent's strategy, defines a *continuation* outcome, $\eta^{(\varphi_\tau)_{\tau \geq t}, \sigma_A | h_t^\tau}$, whose marginal on Θ coincides with $\mu_\tau(h_t^\tau)$ and assigns positive probability to those allocations that are consistent with a^{t-1} and h_t^τ.

Principal's sequential rationality: Fix a dynamic mechanism given a^{t-1}, $(\varphi_\tau)_{\tau \geq t}$, and an agent-PBE (σ_A, μ) of $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau \geq t})$. Suppose that the principal considers offering mechanism φ'_t instead of φ_t. In order to determine whether the principal wishes to deviate to φ'_t, we need to determine the outcome distributions that can follow φ'_t. We denote this set by $\mathsf{D}_{\mathcal{O}^*_{\mathcal{I}}}(\mu_t, a^{t-1}, \varphi'_t)$ and is defined as follows:

$$\mathsf{D}_{\mathcal{O}^*_{\mathcal{I}}}(\mu_t, a^{t-1}, \varphi'_t) = \left\{ \begin{array}{ll} \eta' \in \Delta(\Theta \times A^T) & : \eta' = \eta^{(\varphi'_\tau)_{\tau \geq t}, \sigma'_A} \text{ where} (\varphi'_\tau)_{\tau \geq t} = (\varphi'_t, (\varphi'_\tau)_{\tau \geq t+1}) \text{ and} \\ & \text{(i)} (\varphi'_\tau)_{\tau \geq t} \text{ is a dynamic mechanism given } a^{t-1} \\ & \text{(ii)} (\sigma'_A, \mu') \text{ is an agent-PBE of } \Gamma(\mu_t, a^{t-1}, (\varphi'_\tau)_{\tau \geq t}) \\ & \text{(iii)} (\forall (s', a', \omega') \in SA_\emptyset \times \Omega) \eta^{(\varphi'_\tau)_{\tau \geq t}, \sigma'_A | s', a', \omega'} \in \mathcal{O}^*_{\mathcal{I}}(\mu_{t+1}(s', a', \omega'), a^{t-1}, a') \end{array} \right\}. \tag{1}$$

In words, an outcome η' is in $\mathsf{D}_{\mathcal{O}^*_{\mathcal{I}}}(\cdot, \varphi'_t)$ if it satisfies two properties. First, $\eta' = \eta^{(\varphi'_\tau)_{\tau \geq t}, \sigma'_A}$ for a dynamic mechanism $(\varphi'_\tau)_{\tau \geq t}$ such that φ'_t is the period t-mechanism and (σ'_A, μ') is an agent-PBE given $(\varphi'_\tau)_{\tau \geq t}$. Second, *continuation* outcomes are PBE-feasible. The reason that we are able to require that continuation outcomes are PBE-feasible is that whenever the agent does not condition her strategy on the payoff-irrelevant part of the private history, the following holds: If (σ'_A, μ') is an agent-PBE of $\Gamma(\mu_t, a^{t-1}, (\varphi'_\tau)_{\tau \geq t})$, then for all $h_t^{t+1} = (s', a', \omega')$, $(\sigma'_A, \mu')|_{h_t^{t+1}}$ is an agent-PBE of $\Gamma(\mu'_{t+1,\Theta}(h_t^{t+1}), (a^{t-1}, a'), (\varphi'_\tau(h_t^{t+1}, \cdot))_{\tau \geq t+1})$.

While we can use the set $\mathsf{D}_{\mathcal{O}^*_{\mathcal{I}}}(\mu_t, a^{t-1}, \cdot)$ to test whether the principal has a deviation from $(\varphi_\tau)_{\tau \geq t}$ at the root of $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau \geq t})$, Definition 4 also describes how we test for sequential rationality at later points in the agent-extensive form:

Definition 4 (Sequential rationality)**.** *Fix $t \geq 1$, (μ_t, a^{t-1}), a dynamic mechanism $(\varphi_\tau)_{\tau \geq t}$ given a^{t-1}, and an agent-PBE (σ_A, μ) of $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau \geq t})$. $(\varphi_\tau)_{\tau \geq t}$ is* sequentially rational *given (σ_A, μ) if the following hold:*

1. *For all $\varphi'_t : M \mapsto \Delta(S \times A)$, a distribution $\eta' \in \mathsf{D}_{\mathcal{O}^*_{\mathcal{I}}}(\cdot, \varphi'_t)$ exists such that the principal prefers $\eta^{(\varphi_\tau)_{\tau \geq t}, \sigma_A}$ to η'; that is, $\int_{\Theta \times A^T} W(a^{t-1}, \cdot, \theta) d\eta^{(\varphi_\tau)_{\tau \geq t}, \sigma_A} \geq \int_{\Theta \times A^T} W(a^{t-1}, \cdot, \theta) d\eta'$.*

2. *For all* $h_t^{t+1} = (s', a', \omega') \in (SA_\emptyset \times \Omega)$, $\eta^{(\varphi_\tau)_{\tau \geq t}, \sigma_A | h_t^{t+1}} \in \mathcal{O}^*_{\mathcal{I}}(\mu_{t+1}(h_t^{t+1}), a^{t-1}, a')$.

The first part of Definition 4 states that the principal has no deviations in period t. The second part says that the principal has no deviations in periods $\tau \geq t+1$: the continuation outcome distribution induced by $(\varphi_\tau)_{\tau \geq t}$ and (σ_A, μ) is PBE - feasible in the continuation.

Definition 4 resembles the sequential rationality conditions in the mechanism-selection game, except for one important aspect: when we consider the outcomes the principal faces in the event of a deviation, we do not require that the agent's strategy is measurable in any way with respect to the history of mechanisms so far.[24] By contrast, underlying the definition of the mechanism-selection game is the ability to *measurably* select as a function of φ'_t (in fact, as a function of the whole sequence of mechanisms that has been offered through period $t-1$) the outcomes that the principal faces in the event of a deviation.

We are now ready to define the set of PBE feasible outcomes at (μ_t, a^{t-1}), $\mathcal{O}^*_{\mathcal{I}}(\mu_t, a^{t-1})$:

Definition 5 (PBE-feasible outcomes)**.** *Fix* $t \geq 1$, $(\mu_t, a^{t-1}) \in \Delta(\Theta) \times A^{t-1}$. *The distribution* $\eta \in \Delta(\Theta \times A^T)$ *is PBE-feasible at* (μ_t, a^{t-1}) *if a dynamic mechanism* $(\varphi_\tau)_{\tau \geq t}$ *given* a^{t-1} *and an agent-PBE* (σ_A, μ) *of* $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau \geq t})$ *exist such that*

1. η *is the outcome distribution induced by* (σ_A, μ) *in* $\Gamma(\mu_t, a^{t-1}, (\varphi_\tau)_{\tau \geq t})$, $\eta^{(\varphi_\tau)_{\tau \geq t}, \sigma_A}$,

2. $(\varphi_\tau)_{\tau \geq t}$ *is sequentially rational given* (σ_A, μ).

$\mathcal{O}^*_{\mathcal{I}}(\mu_t, a^{t-1})$ *denotes the set of PBE-feasible outcomes at* (μ_t, a^{t-1}).

By varying $\mathcal{I}$, we can define the set of PBE-feasible outcomes when the principal can offer mechanisms whose input and output messages are $(\Theta, \Delta(\Theta))$. Like in Theorem 1, our interest is in the canonical-PBE-feasible outcomes, that is, those outcomes that are induced by canonical dynamic mechanisms $\left(\varphi^C_\tau\right)_{\tau \geq t}$ (Definition 1) and canonical-agent PBE of the extensive-form game $\Gamma(\mu_t, a^{t-1}, \left(\varphi^C_\tau\right)_{\tau \geq t})$. In a slight abuse of notation, let $\mathcal{O}^C(\cdot)$ denote the correspondence of canonical-PBE-feasible outcomes when $\mathcal{I} = \{(\Theta, \Delta(\Theta))\}$.

Theorem 2. *For all* $t \geq 1$ *and pairs* $(\mu_t, a^{t-1}) \in \Delta(\Theta) \times A^{t-1}$, $\mathcal{O}^*_{\mathcal{I}}(\mu_t, a^{t-1}) = \mathcal{O}^C(\mu_t, a^{t-1})$.

The proof of Theorem 2, which can be found in Online Appendix E.4, follows almost immediately from that in Theorem 1. The only difference is that the proof of Theorem 2 must account for the explicit recursive structure of the sets $\mathcal{O}^*_{\mathcal{I}}$ and $\mathcal{O}^C$. In particular, we are implicitly defining $\mathcal{O}^C$ by requiring that the continuation outcomes are drawn from the (continuation) set $\mathcal{O}^C$. However, in accounting for this difference, we illustrate how both sets can

[24] Indeed, we only need the principal's belief μ_t and the allocations so far a^{t-1} to describe what is PBE-feasible from period t onward. In a sense, the past mechanisms and output messages are bygones.

be defined in terms of an operator similar to that considered in Abreu et al. (1990), which in turn could be used in an application to characterize these sets.

We conclude Section 4 by illustrating the framework in Section 4.2 within Example 1:

4.3 *Example 1 revisited*

We now consider Example 1 under the assumption that the set of types is a continuum. Formally, let $\Theta = [\underline{\theta}, \overline{\theta}] \subseteq \mathbb{R}$, for some finite $\underline{\theta} < \overline{\theta}$. In what follows, we use the standard cdf notation F_1 and F_2 to denote the seller's prior and posterior beliefs, instead of μ_1 and μ_2 as in Section 3.1.

Seller's program: We first argue that the program OPT continues to represent the seller's maximum revenue under a PBE-feasible outcome distribution. By Theorem 2, it continues to be without loss of generality to assume the seller chooses dynamic canonical mechanisms and to analyze the canonical agent-PBE of the game induced by the canonical dynamic mechanism. The only difference between the program OPT and the framework in Section 4.2 is that in OPT the seller only chooses the $t = 1$ mechanism, instead of a dynamic mechanism. However, this distinction is inconsequential: One of the conditions in Definition 5 is that given the seller's belief F_2, the $t = 2$-outcome distribution is PBE-feasible in $t = 2$. It follows that the $t = 2$-mechanism chosen by the seller must maximize his revenue given his posterior belief about the buyer. Proposition 2 in Skreta (2006) implies the $t = 2$-mechanism is a posted price as a function of F_2, $p_2(F_2)$.

Virtual surplus representation: The assumption of a continuum of types allows us to apply standard mechanism design tools to represent the seller's payoff as a function of the allocation rule and the distribution over posteriors implied by the mechanism. To see this, let $J(\theta, F_1) = \theta - (1 - F_1(\theta))/f_1(\theta)$ denote the buyer's virtual value.

As we show in Appendix C, the incentive constraints deliver the envelope representation of the buyer's payoffs, so we can replace the transfers out of the seller's payoff and reduce OPT to the following program. In period 1, the seller chooses a distribution over posteriors, $P_{\Delta(\Theta)}$, and for each posterior he induces, a probability of trade $q(F_2)$ to solve

$$\max_{P_{\Delta(\Theta)}, q} \int_{\Delta(\Theta)} \left[q(F_2) \int_{\underline{\theta}}^{\overline{\theta}} J(\theta, F_1) F_2(d\theta) + (1 - q(F_2))\delta \int_{p_2(F_2)}^{\overline{\theta}} J(\theta, F_1) F_2(d\theta) \right] P_{\Delta(\Theta)}(dF_2) \qquad (2)$$

subject to (i) $P_{\Delta(\Theta)}$ must be Bayes' plausible given F_1 and (ii) a monotonicity condition, which states that, in expectation, higher types must trade with higher probability (see Equation C.5 in Appendix C). Equation 2 describes the seller's payoff in terms of the distribu-

tion over posteriors induced by the mechanism. If at posterior F_2 the seller sells the good ($q(F_2) = 1$), he obtains the expected virtual surplus, where the expectation is calculated using F_2, but the virtual values are calculated using F_1. This reflects that the probability with which the seller pays rents to a buyer of type θ is measured by the probability $F_1(\theta)$ that buyer types below θ receive the good. Instead, if at F_2 the seller does not sell the good ($q(F_2) = 0$) he obtains the (discounted) expected virtual surplus of selling the good at price $p_2(F_2)$. Although the posted price in period 2 is optimal with respect to the *posterior* virtual values $J(\theta, F_2)$, it may not be for the *prior* virtual values. This reflects the conflict between the period 1 and period 2 sellers: if they hold different beliefs about the buyer's type, they pay rents with different probabilities, and therefore may prefer different mechanisms.

Although the solution to the problem in Equation 2 is beyond the scope of this paper,[25] the virtual surplus representation of the seller's problem allows us to expand on what is known about the sale of a durable good under limited commitment. Skreta (2006) shows that, among the mechanisms in Bester and Strausz (2001), posted-prices are the optimal mechanism for the seller. Instead, when endowed with canonical mechanisms, Proposition 2 provides conditions under which the seller will profitably deviate from the optimal posted price mechanism. In other words, the outcome distribution induced by posted prices is not PBE-feasible:

Proposition 2 (Posted prices not PBE-feasible)**.** *Assume $J(\theta, F_1)$ is strictly increasing and $\underline{\theta} = 0$. Then, $\overline{\delta} \in (0,1)$ exists such that for all discount factors $\delta \in (\overline{\delta}, 1)$, the outcome distribution implemented by posted prices is not PBE-feasible.*

The proof of Proposition 2 is constructive. Starting from the optimal posted-price mechanism, we show the seller can deviate to an *obfuscated* non-uniform pricing mechanism. This mechanism is characterized by five parameters $(\tilde{\theta}_1, \tilde{\theta}_2, \gamma, p_\gamma, p_1)$ and works as follows. Pricing is non-uniform because in period 1 types above $\tilde{\theta}_2$ are served with probability 1 and pay p_1, while types in $(\tilde{\theta}_1, \tilde{\theta}_2)$ are *rationed*, that is, they are served with probability γ and pay p_γ if they receive the good. The remaining types are not served in period 1 and pay nothing. It is obfuscated because when $\gamma < 1$, the seller observes whether the good is sold in $t = 1$, but not whether the buyer's type is above or below $\tilde{\theta}_1$. Non-uniform pricing has been studied in the durable goods literature under commitment and limited capacity (see Loertscher and Muir (2021) and the references therein). Instead, Denicolo and Garella (1999) show that obfuscation may benefit a durable goods seller if he cannot commit to the sequence of posted

[25] Equation 2 defines an information design problem with a continuum state space where the designer's payoff depends on more than just the posterior mean, and hence is outside the scope of the existing tools in information design (c.f., Kolotilin, 2018; Dworczak and Martini, 2019).

prices.[26] As we explain next, it is the combination of non-uniform pricing and obfuscation that makes this mechanism more attractive to the seller than the optimal posted-price mechanism.

The proof of Proposition 2 shows the optimal posted-price mechanism is, in a sense, costly for the seller when the seller is patient: to avoid offering low prices in $t = 2$, the seller does not trade with buyer types with positive virtual values in $t = 1$. Instead, by carefully designing the interval $(\tilde{\theta}_1, \tilde{\theta}_2)$, the obfuscated non-uniform pricing mechanism allows the seller to serve these buyer types with positive probability in $t = 1$, without necessarily lowering the price in $t = 2$. The proof constructs a deviation to such a mechanism that guarantees that the unique best response for the buyer is to participate and truthfully report her type. It follows that in this setting the set of continuation outcomes described in Equation 1 is not large enough to deter the seller from this deviation.

We conclude Section 4.3 by discussing the reason that, unlike Skreta (2006), other mechanisms within our class can outperform posted prices in a two-period-setting:

Remark 2 (Comparison with Skreta, 2006)**.** *To understand why in our two-period example the seller can do better than by using the optimal posted price mechanism, it is instructive to compare the incentive constraints in OPT with those implied by mechanisms where the seller observes the buyer's choice of input message as in, for instance, Laffont and Tirole (1988); Bester and Strausz (2001); Skreta (2006). Although not expressed in the language of type reports or beliefs, the incentive constraints in Skreta (2006) require that for each F_2 in the support of $\beta(\cdot|\theta)$, the buyer prefers the tuple $(q(F_2), x(F_2), u^*(\theta, F_2))$ to any other tuple $(q(F_2'), x(F_2'), u^*(\theta, F_2'))$ in the mechanism. In particular, the buyer must be indifferent between any two tuples that she chooses with positive probability. Contrast this with the incentive constraints in OPT, where the buyer is not necessarily indifferent between the tuples $(q(F_2), x(F_2), u^*(\theta, F_2))$ in the support of $\beta(\cdot|\theta)$, although in expectation, the lottery she faces over such tuples under truthtelling must be better than the one she faces by lying. Indeed, in the obfuscated non-uniform pricing mechanism, the seller exploits the weaker incentive constraints in OPT: Buyer types in $(\tilde{\theta}_1, \tilde{\theta}_2)$ are not indifferent between receiving the good in period* 1 *at the rationing price and receiving the good in period* 2*.*

However, for longer horizons, the comparison of the seller's payoffs in the two models is not obvious because the larger set of canonical mechanisms also implies the seller has a larger set

[26] Denicolo and Garella (1999) consider a two-period model, where in each period the seller can choose both a price at which to sell the good, and a probability γ at which the buyer who wants to buy the good at the posted price, receives the good. Implicit in their analysis is that the seller only observes whether the good is sold but not whether the buyer is willing to buy the good at the posted price.

of deviations in our model than in the model in Skreta (2006).

The previous discussion highlights that under limited commitment, the principal may benefit from employing mechanisms where the output message (and hence, the allocation) does *not* reveal the input message that the agent submitted into the mechanism. Contrast this to the standard revelation principle for the case of commitment when the principal faces a privately informed agent (adverse selection): as we explained in the introduction, it follows from the result in Myerson (1982) that it is without loss of generality in that case to consider mechanisms whereby the principal learns the input message from observing the realization of the output message. Instead, Myerson (1982) shows that adding "noise" to the communication may be essential when the principal also faces an agent whose actions are not contractible (moral hazard). Indeed, pooling in the same output message different types of the privately informed agent to incentivize the agent whose action is not contractible to follow the recommendation may be beneficial. Mechanism design with limited commitment is closer to the hybrid model of adverse selection and moral hazard in Myerson (1982) than it is to the model of pure adverse selection. Indeed, note that in a given period, the principal faces, in a sense, *two* agents whose incentives he needs to manage: the privately informed agent (adverse selection) and his future self, whose choice of mechanism is not contractible (moral hazard). That is, today's principal needs to elicit the agent's information while simultaneously ensuring his future behavior is sequentially rational. In the same way that output messages are key in the presence of moral hazard in Myerson (1982), they feature prominently in our framework.

Following Myerson (1982), it would have been natural to consider mechanisms where output messages encode a recommended sequence of mechanisms from tomorrow onward. However, the language of recommendations is self-referential because the set of output messages would refer to continuation mechanisms, which are themselves defined by a set of output messages. Instead, the language of posterior beliefs avoids this potential infinite-regress problem, allowing us to identify a canonical set of output messages for mechanism design with limited commitment.

5 Conclusions and Further Directions

This paper provides a revelation principle for dynamic mechanism-selection games in which the designer can only commit to short-term mechanisms. In doing so, it opens the door to the study of the implications of limited commitment in fundamental problems in economics and political economy, such as optimal taxation, redistribution, and the design

of social insurance (Sleet and Yeltekin, 2008; Farhi et al., 2012; Golosov and Iovino, 2021), or environmental regulation (Hiriart et al., 2011), which, due to the difficulties with the revelation principle, have only been studied within simple informational environments, such as fully non-persistent private information. Since our model allows for non-separable payoffs and non-transferable utility, our results can be used in a broad range of applications.

A cornerstone of the analysis is the idea that a mechanism should encode not only the rules that determine the allocation, but also the information the designer obtains from the interaction with the agent. We expect that the idea that the mechanism's output messages should encode at the very least the principal's information about the agent is more far-reaching and carries to other forms of limited commitment, such as renegotiation with long-term mechanisms, where versions of the revelation principle have proved equally elusive. However, other aspects of our analysis may not carry immediately to the analysis of renegotiation: indeed, Proposition B.1 uncovered that PBE outcomes have a recursive structure. In games with time-separable payoffs, Proposition B.1 implies PBE payoffs can be characterized by relying on self-generating techniques as in Abreu et al. (1990) and Athey and Bagwell (2008), reducing the analysis of a complex dynamic game essentially to a series of static problems (see Doval and Skreta, 2020a for an application).

At the same time, by highlighting the canonical role of beliefs as the signals employed by the mechanism, our work opens up a new avenue for research in information design. Indeed, as the analysis in Section 4.3 and our companion work, Doval and Skreta (2021), highlights, developing information design tools for continuum type spaces when the designer's payoff does not depend only on the posterior mean would contribute to our understanding of classical problems in mechanism design.

References

Abreu, D., D. Pearce, and E. Stacchetti (1990): "Toward a theory of discounted repeated games with imperfect monitoring," *Econometrica*, 58, 1041–1063.

Acharya, A. and J. Ortner (2017): "Progressive learning," *Econometrica*, 85, 1965–1990.

Akbarpour, M. and S. Li (2020): "Credible auctions: A trilemma," *Econometrica*, 88, 425–467.

Aliprantis, C. and K. Border (2006): *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer-Verlag Berlin and Heidelberg GmbH & Company KG.

Athey, S. and K. Bagwell (2008): "Collusion with persistent cost shocks," *Econometrica*, 76, 493–540.

Attar, A., E. Campioni, T. Mariotti, and A. Pavan (2021a): "Keeping the Agents in the Dark: Private Disclosures in Competing Mechanisms," Tech. rep., Toulouse School of Economics.

ATTAR, A., E. CAMPIONI, T. MARIOTTI, AND G. PIASER (2021b): "Competing mechanisms and folk theorems: Two examples," *Games and Economic Behavior*, 125, 79–93.

AUMANN, R. J. (1961): "Borel structures for function spaces," *Illinois Journal of Mathematics*, 5, 614–630.

——— (1964): "Mixed and behavior strategies in infinite extensive games," Tech. rep., Princeton University.

AUSUBEL, L. M. AND R. J. DENECKERE (1989): "Reputation in bargaining and durable goods monopoly," *Econometrica*, 511–531.

BECCUTI, J. AND M. MÖLLER (2018): "Dynamic adverse selection with a patient seller," *Journal of Economic Theory*, 173, 95–117.

BEST, J. AND D. QUIGLEY (2017): "Persuasion for the long run," *Working Paper*.

BESTER, H. AND R. STRAUSZ (2001): "Contracting with imperfect commitment and the revelation principle: the single agent case," *Econometrica*, 69, 1077–1098.

——— (2007): "Contracting with imperfect commitment and noisy communication," *Journal of Economic Theory*, 136, 236–259.

BISIN, A. AND A. A. RAMPINI (2006): "Markets as beneficial constraints on the government," *Journal of public Economics*, 90, 601–629.

BULOW, J. I. (1982): "Durable-goods monopolists," *Journal of Political Economy*, 90, 314–332.

DEB, R. AND M. SAID (2015): "Dynamic screening with limited commitment," *Journal of Economic Theory*, 159, 891–928.

DENICOLO, V. AND P. G. GARELLA (1999): "Rationing in a durable goods monopoly," *The RAND Journal of Economics*, 44–55.

DOVAL, L. AND V. SKRETA (2020a): "Optimal mechanism for the sale of a durable good," *arXiv preprint arXiv:1904.07456*.

——— (2020b): "Supplement to "Mechanism Design with Limited Commitment"," Click here.

——— (2021): "Purchase history and product personalization," *arXiv preprint arXiv:2103.11504*.

DWORCZAK, P. AND G. MARTINI (2019): "The simple economics of optimal persuasion," *Journal of Political Economy*, 127, 1993–2048.

FARHI, E., C. SLEET, I. WERNING, AND S. YELTEKIN (2012): "Non-linear capital taxation without commitment," *Review of Economic Studies*, 79, 1469–1493.

FIOCCO, R. AND R. STRAUSZ (2015): "Consumer standards as a strategic device to mitigate ratchet effects in dynamic regulation," *Journal of Economics & Management Strategy*, 24, 550–569.

FREIXAS, X., R. GUESNERIE, AND J. TIROLE (1985): "Planning under incomplete information and the ratchet effect," *The Review of Economic Studies*, 52, 173–191.

FUDENBERG, D. AND J. TIROLE (1991): "Perfect Bayesian equilibrium and sequential equilibrium," *journal of Economic Theory*, 53, 236–260.

GERARDI, D. AND L. MAESTRI (2020): "Dynamic contracting with limited commitment and the

ratchet effect," *Theoretical Economics*, 15, 583–623.

GOLOSOV, M. AND L. IOVINO (2021): "Social insurance, information revelation, and lack of commitment," *Journal of Political Economy*, 129, 2629–2665.

GUL, F., H. SONNENSCHEIN, AND R. WILSON (1986): "Foundations of dynamic monopoly and the coase conjecture," *Journal of Economic Theory*, 39, 155 – 190.

HIRIART, Y., D. MARTIMORT, AND J. POUYET (2011): "Weak enforcement of environmental policies: a tale of limited commitment and limited fines," *Annals of Economics and Statistics*, 25–42.

KAMENICA, E. AND M. GENTZKOW (2011): "Bayesian persuasion," *American Economic Review*, 101, 2590–2615.

KOLOTILIN, A. (2018): "Optimal information disclosure: A linear programming approach," *Theoretical Economics*, 13, 607–635.

LAFFONT, J.-J. AND J. TIROLE (1988): "The dynamics of incentive contracts," *Econometrica*, 1153–1175.

LIPNOWSKI, E. AND D. RAVID (2020): "Cheap talk with transparent motives," *Econometrica*, 88, 1631–1660.

LIU, Q., K. MIERENDORFF, X. SHI, AND W. ZHONG (2019): "Auctions with limited commitment," *American Economic Review*, 109, 876–910.

LOERTSCHER, S. AND E. V. MUIR (2021): "Monopoly pricing, optimal rationing, and resale," Tech. rep., University of Melbourne.

MCADAMS, D. AND M. SCHWARZ (2007): "Credible sales mechanisms and intermediaries," *American Economic Review*, 97, 260–276.

MCAFEE, R. P. AND D. VINCENT (1997): "Sequentially optimal auctions," *Games and Economic Behavior*, 18, 246–276.

MILGROM, P. AND I. SEGAL (2002): "Envelope theorems for arbitrary choice sets," *Econometrica*, 70, 583–601.

MYERSON, R. B. (1982): "Optimal coordination mechanisms in generalized principal–agent problems," *Journal of Mathematical Economics*, 10, 67–81.

POLLARD, D. (2002): *A user's guide to measure theoretic probability*, 8, Cambridge University Press.

ROCKAFELLAR, R. T. (1970): *Convex analysis*, Princeton University Press.

RUBIN, H. AND O. WESLER (1958): "A note on convexity in Euclidean n-space," *Proceedings of the American Mathematical Society*, 9, 522–523.

SKRETA, V. (2006): "Sequentially optimal mechanisms," *The Review of Economic Studies*, 73, 1085–1111.

——— (2015): "Optimal auction design under non-commitment," *Journal of Economic Theory*, 159, 854–890.

SLEET, C. AND Ş. YELTEKIN (2008): "Politically credible social insurance," *Journal of monetary Economics*, 55, 129–151.

SRIVASTAVA, S. M. (2008): *A course on Borel sets*, vol. 180, Springer Science & Business Media.

STOKEY, N. L. (1981): "Rational expectations and durable goods pricing," *The Bell Journal of Economics*, 112–128.

VARTIAINEN, H. (2013): "Auction design without commitment," *Journal of the European Economic Association*, 11, 316–342.

A COLLECTED DEFINITIONS AND NOTATION

Appendix A introduces the necessary notation to define the payoffs from an assessment and, hence, the definition of Perfect Bayesian equilibrium. It also collects notation that is used in the proofs. Throughout this section and also for most of Appendix B, we assume that Θ is finite and the mechanisms used by the principal have finite support and defer the proof of Theorem 1 for the general case to the Online Appendix.

Shorthand notation: To simplify notation, we do not explicitly include the agent's decision to participate in the mechanism in the histories of the game. Instead, we follow the convention that if the agent does not participate, the input message is $\emptyset$, the output message is $\emptyset$, and the allocation is a^*. With this convention in mind, we let $M_{i\emptyset} = M_i \cup \{\emptyset\}$, $M_i S_j A_\emptyset \equiv \big(M_i \times S_j \times A\big) \cup \{(\emptyset, \emptyset, a^*)\}$ and $S_j A_\emptyset \equiv \big(S_j \times A\big) \cup \{(\emptyset, a^*)\}$, denote the private and public outcomes in a given period when the principal uses a mechanism with labels (M_i, S_j).

Also, given a mechanism $\mathbf{M}_t$, let $z_{(s_t,a_t)}(\mathbf{M}_t)$ denote the tuple $\mathbf{M}_t, s_t, a_t$, which summarizes the period-t outcomes from offering $\mathbf{M}_t$, where $(s_t, a_t) \in S^{\mathbf{M}_t} A_\emptyset$. Note that any public history at the beginning of period t can be written as $h^t = (h^{t-1}, z_{(s_{t-1},a_{t-1})}(\mathbf{M}_{t-1}), \omega_t)$, with the convention that when $t = 1$, $h^1 = \{\omega_1\}$ for some $\omega_1 \in [0,1]$. Finally, given an assessment, $(\sigma_P, \sigma_A, \mu)$, it is useful to collapse the distribution on $M^{\mathbf{M}_t} S^{\mathbf{M}_t} A_\emptyset$, defined by

$$(1 - \pi_t(\theta, h_A^t, \mathbf{M}_t))\mathbb{1}[(m_t, s_t, a_t) = (\emptyset, \emptyset, a^*)] + \pi_t(\theta, h_A^t, \mathbf{M}_t) r_t(\theta, h_A^t, \mathbf{M}_t)(m_t) \varphi^{\mathbf{M}_t}(s_t, a_t | m_t) \tag{A.1}$$

and we denote it by $\kappa_t^{\sigma_A}(m_t, s_t, a_t | \theta, h_A^t, \mathbf{M}_t)$.

Perfect Bayesian equilibrium: To define Perfect Bayesian equilibrium, we need to define the principal and the agent's payoff from a given assessment. To do so, fix an assessment, $(\sigma_P, \sigma_A, \mu)$. The prior μ_1 and the strategy profile $\sigma = (\sigma_P, \sigma_A)$ induce a probability distribution over the terminal histories $\Theta \times H_A^{T+1}$, P^σ, via the Ionescu-Tulcea theorem (Pollard, 2002).[27] Moreover, fixing t and (θ, h_A^t), the measure $P^{\sigma|\theta,h_A^t}$ corresponds to the measure induced by drawing with probability 1 (θ, h_A^t) and then using the continuation strategy profile to determine the distribution over the continuation histories. Fix a public history h^t. The

[27] Online Appendix D provides a formal definition of this distribution.

principal's payoff at h^t is given by

$$W(\sigma,\mu|h^t)=\sum_{\theta,h_A^t\in H_A^t(h^t)}\mu_t(\theta,h_A^t|h^t)\mathbb{E}_{\sigma_{Pt}(h^t)}\left[\mathbb{E}^{P^{\sigma|(\theta,h_A^t,\mathbf{M}_t)}}\left[W(a^{t-1},\cdot,\theta)\right]\right]\equiv\mathbb{E}_{\sigma_{Pt}(h^t)}\left[W(\sigma,\mu|h^t,\mathbf{M}_t)\right],$$

where $a^{t-1}=(a_1,\ldots,a_{t-1})$ is the allocation through the beginning of period t that is consistent with h^t (with the convention that $a^0=\{\emptyset\}$). Note that the principal's payoff from offering $\mathbf{M}_t$ at h^t, $W(\sigma,\mu|h^t,\mathbf{M}_t)$, depends on the belief system μ only through $\mu_t(\cdot|h^t)$.

For any mechanism $\mathbf{M}_t$, $W(\sigma,\mu|h^t,\mathbf{M}_t)$ equals

$$\sum_{\theta,h_A^t\in H_A^t(h^t)}\mu_t(\theta,h_A^t|h^t)\sum_{(m_t,s_t,a_t)\in M^{\mathbf{M}_t}S^{\mathbf{M}_t}A_\emptyset}\kappa_t^{\sigma_A}(m_t,s_t,a_t|\theta,h_A^t,\mathbf{M}_t)\mathbb{E}^{P^{\sigma|(\theta,h_A^t,\mathbf{M}_t,m_t,s_t,a_t)}}[W(a^{t-1},a_t,\cdot,\theta)].$$

The principal's beliefs at h^t, together with mechanism $\mathbf{M}_t$ and the agent's strategy, define a distribution over the continuation public histories as follows:

$$\nu_{t+1}^{\mu,\sigma_A}(h^t,z_{(s_t,a_t)}(\mathbf{M}_t)|h^t,\mathbf{M}_t)=\sum_{\theta,h_A^t\in H_A^t(h^t),m_t\in M_\emptyset^{\mathbf{M}_t}}\mu_t(\theta,h_A^t|h^t)\kappa_t^{\sigma_A}(m_t,s_t,a_t|\theta,h_A^t,\mathbf{M}_t).\quad\text{(A.2)}$$

With this notation in hand, we can express the principal's payoff $W(\sigma,\mu|h^t,\mathbf{M}_t)$ as:

$$W(\sigma,\mu|h^t,\mathbf{M}_t)=\sum_{(s_t,a_t)\in S^{\mathbf{M}_t}A_\emptyset}\nu_{t+1}^{\mu,\sigma_A}(h^t,z_{(s_t,a_t)}(\mathbf{M}_t)|h^t,\mathbf{M}_t)W(\sigma,\mu|h^t,z_{(s_t,a_t)}(\mathbf{M}_t)).\quad\text{(A.3)}$$

Similarly, the agent's payoff at private history (θ,h_A^t), when the principal offers mechanism $\mathbf{M}_t$, is given by:

$$U(\sigma|\theta,h_A^t,\mathbf{M}_t)=\sum_{(m_t,s_t,a_t)\in M^{\mathbf{M}_t}S^{\mathbf{M}_t}A_\emptyset}\kappa_t^{\sigma_A}(m_t,s_t,a_t|\theta,h_A^t,\mathbf{M}_t)\mathbb{E}^{P^{\sigma|(\theta,h_A^t\mathbf{M}_t,m_t,s_t,a_t)}}\left[U(a^{t-1},a_t,\cdot,\theta)\right].\quad\text{(A.4)}$$

With this, we can formally define Perfect Bayesian equilibrium.

Definition A.1. *An assessment* (σ_P,σ_A,μ) *is* sequentially rational *if for all* $t\geq 1$ *and public histories* h^t*, the following hold:*

1. *If* $\mathbf{M}_t$ *is in the support of* $\sigma_{Pt}(h^t)$*, then* $W(\sigma,\mu|h^t,\mathbf{M}_t)\geq W(\sigma,\mu|h^t,\mathbf{M}_t')$ *for all* $\mathbf{M}_t'\in\mathcal{M}_\mathcal{I}$*,*
2. *For all* $(\theta,h_A^t)\in\Theta\times H_A^t(h^t)$*, and* $\mathbf{M}_t$ *in* $\mathcal{M}_\mathcal{I}$*,* $U(\sigma|\theta,h_A^t,\mathbf{M}_t)\geq U(\sigma_P,\sigma'_A|\theta,h_A^t,\mathbf{M}_t)$ *for all* σ'_A*.*

Definition A.2. *An assessment* (σ_P,σ_A,μ) satisfies Bayes' rule where possible *if for all public*

histories h^t and mechanisms $\mathbf{M}_t$ the following holds:

$$\mu_{t+1}(\theta, h_A^t, m_t, z_{(s_t,a_t)}(\mathbf{M}_t), \omega_{t+1}|h^t, z_{(s_t,a_t)}(\mathbf{M}_t), \omega_{t+1})\nu_{t+1}^{\mu,\sigma_A}(h^t, z_{(s_t,a_t)}(\mathbf{M}_t)|h^t, \mathbf{M}_t) \tag{A.5}$$
$$= \mu_t(\theta, h_A^t|h^t)\kappa_t^{\sigma_A}(m_t, s_t, a_t|\theta, h_A^t, \mathbf{M}_t).$$

Definition A.3. *An assessment $(\sigma_P, \sigma_A, \mu)$ is a* Perfect Bayesian equilibrium *if it is sequentially rational and satisfies Bayes' rule where possible.*

Prunning: Given a mechanism $\mathbf{M}_t$, let $(S^{\mathbf{M}_t} \times A)_+ = \{(s_t, a_t) : (\exists m \in M^{\mathbf{M}_t})\varphi^{\mathbf{M}_t}(s_t, a_t|m) > 0\}$. The set $S^{\mathbf{M}_t} \times A \setminus (S^{\mathbf{M}_t} \times A)_+$ has zero probability *regardless* of the agent's strategy. Hence, if we remove from the tree those paths that are consistent with mechanism $\mathbf{M}_t$ and $(s, a) \notin (S^{\mathbf{M}_t} \times A)_+$, this does not change the set of equilibrium outcomes. Hereafter, these histories are removed from the tree.

Principal pure strategies: Lemma D.1 in Online Appendix D.2 shows that without loss of generality, we can focus on PBE assessments of $G_{\mathcal{I}}$ in which the principal plays a pure strategy, so in what follows, we focus on PBE assessments that satisfy Lemma D.1.

B Proof of Theorem 1

The proof of Theorem 1 follows from the proof of Propositions B.1-B.2 below.

Proposition B.1. *For every PBE assessment $(\sigma_P, \sigma_A, \mu)$ of $G_{\mathcal{I}}$, an outcome-equivalent PBE assessment $(\sigma'_P, \sigma'_A, \mu')$ exists such that the agent's strategy only depends on her type and the public history.*

We relegate the proof of Proposition B.1 to Online Appendix D.1.1. In what follows, we focus on PBE of the mechanism-selection game that satisfy the properties of Proposition B.1 and abuse notation in the following two ways: First, we write the agent's strategy as a function of her private type and the public history alone, with the understanding that $\sigma_{At}(\theta, h_A^t) = \sigma_{At}(\theta, h^t)$ whenever $h_A^t \in H_A^t(h^t)$. Similarly, we write the belief system at history h^t as inducing distributions over Θ and not over $\Theta \times H_A^t(h^t)$.

Proposition B.2. *Fix $\mathcal{I}$ and let $(\sigma_P, \sigma_A, \mu)$ be a PBE assessment of $G_{\mathcal{I}}$ that satisfies Proposition B.1. Then, an outcome-equivalent canonical PBE assessment $(\sigma'_P, \sigma'_A, \mu')$ of $G_{\mathcal{I}}$ exists.*

Proof of Proposition B.2. Let $(\sigma_P, \sigma_A, \mu)$ be as in the statement of Proposition B.2. Let h^t be a public history and let $\mathbf{M}_t$ denote the mechanism that the principal offers at h^t under σ_{Pt}. Let Θ^+ denote the support of the principal's beliefs at h^t, $\mu_t(h^t)$.

For types in Θ^+, use Equation A.1 to define an auxiliary mapping $\varphi' : \Theta^+ \mapsto \Delta\left(S^{\mathbf{M}_t} A_\emptyset\right)$, as

follows:

$$\varphi'(s_t, a_t|\theta) = \sum_{m \in M_\emptyset^{\mathbf{M}_t}} \kappa_t^{\sigma_A}(m_t, s_t, a_t|\theta, h^t, \mathbf{M}_t). \tag{B.1}$$

That is, φ' corresponds to the *direct* version of $\varphi^{\mathbf{M}_t}$ for $\theta \in \Theta^+$; we use it in what follows to construct an alternative mechanism for the principal, $\mathbf{M}_t'$, that uses message sets $(\Theta, \Delta(\Theta))$.

Omitting the dependence on (σ_A, μ), recall that $\nu_{t+1}(h^t, z_{(s_t,a_t)}(\mathbf{M}_t)|h^t, \mathbf{M}_t)$ denotes the probability of history $(h^t, z_{(s_t,a_t)}(\mathbf{M}_t))$ under the equilibrium strategy when the principal offers $\mathbf{M}_t$ at h^t (Equation A.2). Equation A.2 implies we can write ν_{t+1} using φ' as follows:

$$\nu_{t+1}(h^t, z_{(s_t,a_t)}(\mathbf{M}_t)|h^t, \mathbf{M}_t) = \sum_{\theta \in \Theta^+} \mu_t(\theta|h^t)\varphi'(s_t, a_t|\theta).$$

In what follows, to simplify notation we omit the dependence of $\nu_{t+1}(\cdot|h^t, \mathbf{M}_t)$ on $\mathbf{M}_t$.

The first step is to show that the distribution over continuation histories ν_{t+1} can be seen as inducing a distribution over posterior beliefs, allocations, and realizations of a public randomization device. To see this, for $\mu \in \Delta(\Theta)$, let $B(\mu)$ denote the set

$$B(\mu) = \left\{(s_t, a_t) \in S^{\mathbf{M}_t} A_\emptyset : \mu_{t+1}(\cdot|h^t, z_{(s_t,a_t)}(\mathbf{M}_t)) = \mu\right\},$$

and let $B_{a_t}(\mu)$ denote the projection of $B(\mu)$ onto $\{a_t\}$. In what follows, for any subset $B \subseteq S^{\mathbf{M}_t} A_\emptyset$, we abuse notation and write $\nu_{t+1}(B|h^t)$ instead of $\sum_{(s_t,a_t)\in B} \nu_{t+1}(h^t, z_{(s_t,a_t)}(\mathbf{M}_t)|h^t)$.

Let $\Delta(\Theta)^+$ denote the smallest subset of $\Delta(\Theta)$ such that $\nu_{t+1}(B(\Delta(\Theta)^+)|h^t) = 1$. That is, $\Delta(\Theta)^+$ is the set of principal posterior beliefs that are pinned down via Bayes' rule. We can write the principal's payoff at history h^t when he offers $\mathbf{M}_t$ as follows:

$$\begin{aligned}
&\sum_{(s_t,a_t)} \nu_{t+1}(h^t, z_{(s_t,a_t)}(\mathbf{M}_t)|h^t) W(\sigma, \mu_{t+1}(\cdot|h^t, z_{(s_t,a_t)}(\mathbf{M}_t))|h^t, z_{(s_t,a_t)}(\mathbf{M}_t)) \qquad \text{(B.2)}\\
&= \sum_{\mu \in \Delta(\Theta)^+} \nu_{t+1}(B(\mu)|h^t) \sum_{a_t \in A} \underbrace{\frac{\nu_{t+1}(B_{a_t}(\mu)|h^t)}{\nu_{t+1}(B(\mu)|h^t)}}_{\text{allocation rule}} \sum_{s_t \in B_{a_t}(\mu)} \underbrace{\frac{\nu_{t+1}(h^t, z_{(s_t,a_t)}(\mathbf{M}_t)|h^t)}{\nu_{t+1}(B_{a_t}(\mu)|h^t)}}_{\text{public randomization}} W(\sigma, \mu|h^t, z_{(s_t,a_t)}(\mathbf{M}_t)).
\end{aligned}$$

The above equation shows two ways in which we can think of the distribution over continuation histories starting from h^t. The first is standard: we draw history $(h^t, z_{(s_t,a_t)}(\mathbf{M}_t))$ using the distribution induced by the equilibrium strategy, $\nu_{t+1}(\cdot|h^t)$. The second is the one that delivers the canonical mechanisms: we first draw a belief μ using the distribution over continuation equilibrium beliefs induced by $\nu_{t+1}(\cdot|h^t)$ and then we draw an allocation a_t, con-

ditional on the continuation equilibrium belief coinciding with μ. The principal's posterior belief μ and the allocation a_t may still not be enough to pin down the continuation history, so we draw the output message s_t conditional on s_t being consistent with a_t and μ.

The second step is to show that, conditional on the induced posterior belief μ, (i) the probability that the allocation is a_t is independent of θ, and (ii) the probability that the output message is $s_t \in B_{a_t}(\mu)$ is independent of θ. To see this, note that for any belief $\mu \in \Delta(\Theta)^+$, for any $(s_t, a_t) \in B(\mu)$ and for any θ such that $\mu(\theta) > 0$, we have[28]

$$\mu(\theta) = \frac{\mu_t(\theta|h^t)\varphi'(s_t, a_t|\theta)}{\nu_{t+1}(h^t, z_{(s_t,a_t)}(\mathbf{M}_t)|h^t)} = \frac{\mu_t(\theta|h^t)\sum_{s'_t \in B_{a_t}(\mu)}\varphi'(s'_t, a_t|\theta)}{\nu_{t+1}(B_{a_t}(\mu)|h^t)} = \frac{\mu_t(\theta|h^t)\sum_{(s'_t,a'_t)\in B(\mu)}\varphi'(s'_t, a'_t|\theta)}{\nu_{t+1}(B(\mu)|h^t)}. \tag{B.3}$$

That is, the principal updates to μ either when (i) he observes $(s_t, a_t) \in B(\mu)$, (ii) he learns that a_t is the realized allocation, that is, he learns that $s_t \in B_{a_t}(\mu)$, or (iii) he learns that the output message and the allocation belong to $B(\mu)$. Thus, for all θ in the support of μ, we have

$$\begin{aligned} \frac{\nu_{t+1}(B_{a_t}(\mu)|h^t)}{\nu_{t+1}(B(\mu)|h^t)} &= \frac{\sum_{s'_t \in B_{a_t}(\mu)}\varphi'(s'_t, a_t|\theta)}{\sum_{(s'_t,a'_t)\in B(\mu)}\varphi'(s'_t, a'_t|\theta)} \\ \frac{\nu_{t+1}(h^t, z_{(s_t,a_t)}(\mathbf{M}_t)|h^t)}{\nu_{t+1}(B_{a_t}(\mu)|h^t)} &= \frac{\varphi'(s_t, a_t|\theta)}{\sum_{s'_t \in B_{a_t}(\mu)}\varphi'(s'_t, a_t|\theta)}, \end{aligned} \tag{B.4}$$

where each of the equalities follows from applying Equation B.3. Equation B.4 shows (i) the probability that the allocation is a_t conditional on the induced belief being μ is independent of θ, and (ii) the probability that the output message is $s_t \in B_{a_t}(\mu)$ conditional on the allocation being a_t and the induced belief μ is independent of θ. It follows that for all $\mu \in \Delta(\Theta)^+$, $(s_t, a_t) \in B(\mu)$ and for all θ in the support of μ, we can *split* the auxiliary mapping as follows:

$$\begin{aligned} \varphi'(s_t, a_t|\theta) &= \left(\sum_{(s'_t,a'_t)\in B(\mu)}\varphi'(s'_t, a'_t|\theta)\right)\frac{\nu_{t+1}(B_{a_t}(\mu)|h^t)}{\nu_{t+1}(B(\mu)|h^t)}\frac{\nu_{t+1}(h^t, z_{(s_t,a_t)}(\mathbf{M}_t)|h^t)}{\nu_{t+1}(B_{a_t}(\mu)|h^t)} \\ &= \frac{\mu(\theta)}{\mu_t(\theta|h^t)}\nu_{t+1}(B(\mu)|h^t)\frac{\nu_{t+1}(B_{a_t}(\mu)|h^t)}{\nu_{t+1}(B(\mu)|h^t)}\frac{\nu_{t+1}(h^t, z_{(s_t,a_t)}(\mathbf{M}_t)|h^t)}{\nu_{t+1}(B_{a_t}(\mu)|h^t)}, \end{aligned}$$

where the last equality follows from the last equality in Equation B.3.

Thus, the agent's payoff at history h^t, when the principal offers mechanism $\mathbf{M}_t$ and her

[28] Note that if $\mu \in \Delta(\Theta)^+$ and θ is such that $\mu(\theta) > 0$, then $\theta \in \Theta^+$.

type is $\theta \in \Theta^+$, can be written as follows:

$$\sum_{(s_t,a_t)\in S^{\mathbf{M}_t}A_\emptyset} \varphi'(s_t,a_t|\theta)\mathbb{E}^{P^{\sigma|(\theta,h^t,z_{(s_t,a_t)}(\mathbf{M}_t))}}\left[U(a^{t-1},a_t,\cdot,\theta)\right] = \tag{B.5}$$

$$\sum_{\mu\in\Delta(\Theta)} \frac{\mu(\theta)}{\mu_t(\theta|h^t)}\nu_{t+1}(B(\mu)|h^t) \sum_{a_t\in A} \frac{\nu_{t+1}(B_{a_t}(\mu)|h^t)}{\nu_{t+1}(B(\mu)|h^t)} \sum_{s_t\in B_{a_t}(\mu)} \frac{\nu_{t+1}(h^t,z_{(s_t,a_t)}(\mathbf{M}_t)|h^t)}{\nu_{t+1}(B_{a_t}(\mu)|h^t)}\mathbb{E}^{P^{\sigma|(\theta,h^t,z_{(s_t,a_t)}(\mathbf{M}_t))}}[U(a^{t-1},a_t,\cdot,\theta)].$$

The difference between the principal and the agent's payoff in Equations B.2 and B.5 is that the agent cares only about the distribution over (μ,s_t,a_t) *conditional* on θ, whereas the principal's payoff is expressed in terms of the unconditional distribution. For this reason the agent's payoff features the term $\mu(\theta)/\mu_t(\theta|h^t)$.

We now define the canonical mechanism $\mathbf{M}_t^C = (\Theta,\Delta(\Theta),\varphi^{\mathbf{M}_t^C})$: First, for $\theta\in\Theta^+$,

$$\varphi^{\mathbf{M}_t^C}(\mu,a_t|\theta) = \underbrace{\frac{\mu(\theta)}{\mu_t(\theta|h^t)}\nu_{t+1}(B(\mu)|h^t)}_{\beta^{\mathbf{M}_t^C}(\mu|\theta)} \underbrace{\frac{\nu_{t+1}(B_{a_t}(\mu)|h^t)}{\nu_{t+1}(B(\mu)|h^t)}}_{\alpha^{\mathbf{M}_t^C}(a_t|\mu)},$$

where the decomposition in terms of $\beta^{\mathbf{M}_t^C},\alpha^{\mathbf{M}_t^C}$ is well-defined because of the independence properties highlighted after Equation B.4. Second, if $\theta\notin\Theta^+$, let $\theta^*(\theta)$ denote the solution to

$$\max_{\tilde\theta\in\Theta^+}\sum_{(\mu,a_t)\in\Delta(\Theta)\times A}\varphi^{\mathbf{M}_t^C}(\mu,a_t|\tilde\theta)\sum_{s_t\in B_{a_t}(\mu)}\frac{\nu_{t+1}(h^t,z_{(s_t,a_t)}(\mathbf{M}_t)|h^t)}{\nu_{t+1}(B_{a_t}(\mu)|h^t)}\mathbb{E}^{P^{\sigma|(\theta,h^t,z_{(s_t,a_t)}(\mathbf{M}_t))}}\left[U(a^{t-1},a_t,\cdot,\theta)\right] \tag{B.6}$$

and let $\varphi^{\mathbf{M}_t^C}(\cdot|\theta) = \varphi^{\mathbf{M}_t^C}(\cdot|\theta^*(\theta))$. Change the principal's strategy at h^t so that he offers $\mathbf{M}_t^C$ instead of $\mathbf{M}_t$. Change the agent's strategy so that, conditional on participating, the agent truthfully reports her type, $r_t'(\theta,h^t,\mathbf{M}_t^C)=\delta_\theta$.

For $\mu\in\Delta(\Theta)^+$ and allocation a_t, enumerate the output messages in $B_{a_t}(\mu)$ as $s_t^1,\dots,s_t^K$. (We omit the dependence of K on μ and a_t to simplify notation.) Define the sequence $\{\omega_k\}_{k=0}^K$ such that $\omega_0=0,\omega_K=1$ and for $k=1,\dots,K-1$,

$$\omega_k-\omega_{k-1} = \frac{\nu_{t+1}(h^t,z_{(s_t^k,a_t)}(\mathbf{M}_t)|h^t)}{\nu_{t+1}(B_{a_t}(\mu)|h^t)}.$$

Modify the continuation strategies as follows: for $k=1,\dots,K$ and $\omega\in[\omega_{k-1},\omega_k]$, let $\sigma|_{(h^t,z_{(\mu,a_t)}(\mathbf{M}_t^C),\omega)}$ coincide with $\sigma|_{(h^t,z_{(s_t^k,a_t)}(\mathbf{M}_t),\frac{\omega-\omega_{k-1}}{\omega_k-\omega_{k-1}})}$. Note these strategies imply the principal and the agent's payoffs remain the same as in the original equilibrium whenever $\theta\in\Theta^+$. Furthermore, modify the continuation strategies so that $\sigma|_{(h^t,z_{(\emptyset,a^*)}(\mathbf{M}_t^C))}=\sigma|_{(h^t,z_{(\emptyset,a^*)}(\mathbf{M}_t))}$.

For θ in Θ^+, set $\pi_t'(\theta,h^t,\mathbf{M}_t^C)=1$. For types not in Θ^+, use Equation B.6 to compute

$\pi'_t(\theta, h^t, \mathbf{M}_t^C)$ accordingly. Conditional on participating, the agent can guarantee at most the payoff from imitating the strategy followed by θ' for some $\theta' \in \Theta^+$. This strategy was already feasible in the original PBE, so the agent has no new deviations. It follows that the new assessment is a PBE of the auxiliary game. □

B.1 Proof of Proposition 1

Fix $\mathcal{I}$ and let $(\sigma_P, \sigma_A, \mu)$ denote a canonical PBE of $G_{\mathcal{I}}$, like the one constructed in Proposition B.2. Fix a history h^t and let $\mathbf{M}_t$ denote the mechanism the principal offers at h^t under σ_{Pt}. Let Θ^+ denote the support of $\mu_t(h^t)$. To simplify notation, we use the shorthand notation $(w^*, (u^*_\theta)_{\theta\in\Theta^+}, (u^\pi_\theta)_{\theta\notin\Theta^+})$ to denote the principal and the agent's payoff vector at h^t when the agent participates in $\mathbf{M}_t$, $(W(\sigma,\mu|h^t,\mathbf{M}_t), U(\sigma|\theta,h^t,\mathbf{M}_t,1))$. Theorem 1 implies that for the principal and for $\theta\in\Theta^+$, w^* and u^*_θ are their equilibrium payoffs at h^t.

Similar steps to those in the proof of Proposition B.2 in Online Appendix D.1 imply a distribution $\tau\in\Delta(\Delta(\Theta))$ exists such that for all measurable subsets $\widetilde{U}$ of $\Delta(\Theta)$ and for all $\theta\in\Theta$,

$$\beta^{\mathbf{M}_t}(\widetilde{U}|\theta)\mu_t(\theta|h^t) = \int_{\widetilde{U}} \mu'(\theta)\tau(d\mu').$$

Therefore, we can write the principal and the agent's payoffs as follows:

$$w^* = \int_{\Delta(\Theta)} \sum_{\theta\in\Theta} \mu'(\theta) \left[\int_A \mathbb{E}^{P^{\sigma|(\theta,h^t,\mathbf{M}_t,\mu',a_t)}} \left[W(a^{t-1}, a_t, \cdot, \theta) \right] \alpha^{\mathbf{M}_t}(da_t|\mu') \right] \tau(d\mu') \equiv \int_{\Delta(\Theta)} w(\mu')\tau(d\mu'),$$

$$u^*_\theta = \int_{\Delta(\Theta)} \int_A \mathbb{E}^{P^{\sigma|(\theta,h^t,\mathbf{M}_t,\mu',a_t)}} \left[U(a^{t-1}, a_t, \cdot, \theta) \right] \alpha^{\mathbf{M}_t}(da_t|\mu') \frac{\mu'(\theta)}{\mu_t(\theta|h^t)} \tau(d\mu') \equiv \int_{\Delta(\Theta)} u(\mu', \theta)\tau(d\mu'),$$

$$u^\pi_\theta = \int_{\Delta(\Theta)} \int_A \mathbb{E}^{P^{\sigma|(\theta,h^t,\mathbf{M}_t,\mu',a_t)}} \left[U(a^{t-1}, a_t, \cdot, \theta) \right] \alpha^{\mathbf{M}_t}(da_t|\mu') \frac{\mu'(\theta^*(\theta))}{\mu_t(\theta^*(\theta)|h^t)} \tau(d\mu') \equiv \int_{\Delta(\Theta)} u(\mu', \theta)\tau(d\mu'),$$

where $\theta^*(\theta)\in\Theta^+$ is as in the proof of Proposition B.2 (recall Equation B.6). Finally, note that the payoff of the agent of type θ from reporting $\theta'\in\Theta^+$ at history h^t when the mechanism is $\mathbf{M}_t$, and then following her equilibrium strategy is given by:

$$\int_{\Delta(\Theta)} \int_A \mathbb{E}^{P^{\sigma|(\theta,h^t,\mathbf{M}_t,\mu',a_t)}} \left[U(a^{t-1}, a_t, \cdot, \theta) \right] \alpha^{\mathbf{M}_t}(da_t|\mu') \beta^{\mathbf{M}_t}(d\mu'|\theta')$$
$$= \int_{\Delta(\Theta)} \int_A \mathbb{E}^{P^{\sigma|(\theta,h^t,\mathbf{M}_t,\mu',a_t)}} \left[U(a^{t-1}, a_t, \cdot, \theta) \right] \alpha^{\mathbf{M}_t}(da_t|\mu') \frac{\mu'(\theta')}{\mu_t(\theta'|h^t)} \tau(d\mu') \equiv \int_{\Delta(\Theta)} d_\theta(\theta', \mu')\tau(d\mu') = d^*_{\theta,\theta'}.$$

The notation d_θ denotes that these are the payoffs that θ obtains from deviating to θ'. The payoff $d^*_{\theta,\theta'}$ can be similarly defined for $\theta'\notin\Theta^+$ using $\theta^*(\theta')$ as in the definition of u^π_θ.

Let $S = \text{supp}\tau$. Define $C = \{(\mu', w(\mu'), (u(\theta,\mu'))_{\theta\in\Theta}, (d_\theta(\theta',\mu'))_{\theta,\theta'\in\Theta}) : \mu'\in S\}$. Because Θ

is finite, Rubin and Wesler (1958) implies that $(\mu(h^t), w^*, (u^*_\theta)_{\theta\in\Theta^+}, (u^\pi_\theta)_{\theta\notin\Theta^+}, (d^*_{\theta,\theta'})_{\theta,\theta'\in\Theta}) \in \text{co}C$. Letting $N = |\Theta|$, Caratheodory's theorem implies $M \leq N(N+1)+1$ exists such that $(\mu(h^t), w^*, (u^*_\theta)_{\theta\in\Theta^+}, (u^\pi_\theta)_{\theta\notin\Theta^+}, (d^*_{\theta,\theta'})_{\theta,\theta'\in\Theta})$ can be written as the convex combination of M elements of C. That is, $(\lambda_i, \mu'_i)_{i=1}^M$ exist such that $\lambda_i \geq 0, \sum_{i=1}^M \lambda_i = 1$, and

$$\mu_t(h^t) = \sum_{i=1}^M \lambda_i \mu'_i \quad w^* = \sum_{i=1}^M \lambda_i w(\mu'_i) \quad u^*_\theta = \sum_{i=1}^M \lambda_i u(\theta,\mu'_i) \quad u^\pi_\theta = \sum_{i=1}^M \lambda_i u(\theta,\mu'_i) \quad d^*_{\theta,\theta'} = \sum_{i=1}^M \lambda_i d_\theta(\theta',\mu'_i).$$

Consider the following canonical mechanism $\mathbf{M}'_t$: For types in Θ^+, let $\beta^{\mathbf{M}'_t}(\{\mu'_i\}|\theta) = \frac{\mu'_i(\theta)}{\mu_t(\theta|h^t)}\lambda_i$, and otherwise, $\beta^{\mathbf{M}'_t}(\cdot|\theta) = 0$. For types not in Θ^+, let $\beta^{\mathbf{M}'_t}(\cdot|\theta) = \beta^{\mathbf{M}'_t}(\cdot|\theta^*(\theta))$. Furthermore, $\alpha^{\mathbf{M}'_t}(\cdot|\mu'_i) = \alpha^{\mathbf{M}_t}(\cdot|\mu'_i)$. Modify the PBE assessment so that the principal offers $\mathbf{M}'_t$ at h^t. Furthermore, modify the continuation strategies so that $(\sigma_P,\sigma_A,\mu)|_{(h^t,\mathbf{M}'_t,z_{(\mu'_i,a_t)}(\mathbf{M}'_t))} = (\sigma_P,\sigma_A,\mu)|_{(h^t,\mathbf{M}_t,z_{(\mu'_i,a_t)}(\mathbf{M}_t))}$ and $(\sigma_P,\sigma_A,\mu)|_{(h^t,\mathbf{M}'_t,z_{(\emptyset,a^*)}(\mathbf{M}'_t))} = (\sigma_P,\sigma_A,\mu)|_{(h^t,\mathbf{M}_t,z_{(\emptyset,a^*)}(\mathbf{M}_t))}$. Clearly, mechanism $\mathbf{M}'_t$ delivers the same payoff to the principal and the agent conditional on the agent participating and truthfully reporting her type. We now show it is optimal for the agent to truthfully report her type. Suppose the agent of type θ reports instead that her type is $\theta' \in \Theta^+$ (the case $\theta' \notin \Theta^+$ is similar). In this case, she obtains:

$$\sum_{i=1}^M \lambda_i \frac{\mu'_i(\theta')}{\mu_t(\theta'|h^t)} \int_A \mathbb{E}^{P^{\sigma|\theta,h^t,z_{(\mu'_i,a_t)}(\mathbf{M}'_t)}} \left[U(a^{t-1},a_t,\cdot,\theta)\right] \alpha^{\mathbf{M}'_t}(da_t|\mu'_i) = \sum_{i=1}^M \lambda_i d_\theta(\theta',\mu'_i) = d^*_{\theta,\theta'}$$

Since $u^*_\theta \geq d^*_{\theta,\theta'}$, then truthtelling is still optimal. Thus, because the payoffs from participating and not participating are the same as in the original assessment for all types, $\pi(\theta,h^t,\mathbf{M}'_t) = \pi_t(\theta,h^t,\mathbf{M}_t)$ is a best response.

C Proofs of Section 4.3

Envelope theorem: We establish the envelope theorem and the virtual surplus representation of the seller's payoff. The buyer's payoff in the mechanism when her type is θ is given by

$$U(\theta) = \int_{\Delta(\Theta)} \left(\theta q(F_2) - x(F_2) + (1-q(F_2))\delta u^*(\theta,F_2)\right)\beta(dF_2|\theta), \tag{C.1}$$

where $u^*(\theta,F_2) = \max\{0,\theta - p_2(F_2)\}$. Equation C.2 defines the buyer's payoff when her type is θ and she deviates by first reporting θ' and then following θ''s strategy in $t=2$:

$$V(\theta',\theta) = \int_{\Delta(\Theta)} \left(\theta q(F_2) - x(F_2) + (1-q(F_2))\delta\left[u^*(\theta',F_2) + (\theta-\theta')q^*_2(\theta',F_2)\right]\right)\beta(dF_2|\theta'), \tag{C.2}$$

where $q^*_2(\theta',F_2) = \mathbb{1}[\theta' \geq p_2(F_2)]$. The optimality of truthtelling implies $U(\theta) = \max\left\{V(\theta',\theta) : \theta' \in \Theta\right\}$. We now establish that the family $\{V(\theta',\cdot) : \theta' \in \Theta\}$ is equi-Lipschitz

continuous. To see this, let $\theta, \tilde{\theta}$ such that $\tilde{\theta} \neq \theta$ and consider

$$|V(\theta',\theta) - V(\theta',\tilde{\theta})| = \left| \int_{\Delta(\Theta)} \left[(\theta - \tilde{\theta}) q(F_2) + (1 - q(F_2)) \delta (\theta - \tilde{\theta}) q_2^*(\theta', F_2) \right] \beta(dF_2|\theta') \right|$$
$$\leq |\theta - \tilde{\theta}| \left| \int_{\Delta(\Theta)} \left[q(F_2) + (1 - q(F_2)) \delta q_2^*(\theta', F_2) \right] \beta(dF_2|\theta') \right| \leq |\theta - \tilde{\theta}|(1 + \delta).$$

Moreover, for θ in the interior of Θ we have that

$$\frac{\partial}{\partial \theta} V(\theta',\theta) = \lim_{h \to 0} \frac{V(\theta',\theta + h) - V(\theta',\theta)}{h} = \int_{\Delta(\Theta)} \left[q(F_2) + (1 - q(F_2)) \delta q_2^*(\theta', F_2) \right] \beta(dF_2|\theta').$$

$U(\theta)$ is then Lipschitz continuous because it is the maximum over a family of equi-Lipschitz continuous functions. Thus, it is differentiable almost everywhere. Theorem 1 in Milgrom and Segal (2002) implies that at any point of differentiability of $U(\cdot)$,

$$U'(\theta) = \frac{\partial}{\partial \tilde{\theta}} V(\theta, \tilde{\theta})|_{\tilde{\theta} = \theta} = \int_{\Delta(\Theta)} \left[q(F_2) + (1 - q(F_2)) \delta q_2^*(\theta, F_2) \right] \beta(dF_2|\theta).$$

It follows that

$$\int_\Theta \left[\int_{\Delta(\Theta)} x(F_2) \beta(dF_2|\theta) \right] F_1(d\theta) = \int_\Theta \int_{\Delta(\Theta)} \left[\theta q(F_2) + (1 - q(F_2)) \delta u^*(\theta, F_2) \right] \beta(dF_2|\theta) F_1(d\theta)$$
$$- \int_\Theta \int_{\underline{\theta}}^{\theta} \left[\int_{\Delta(\Theta)} \left(q(F_2) + (1 - q(F_2)) \delta q_2^*(u, F_2) \right) \beta(dF_2|u) \right] du F_1(d\theta), \quad \text{(C.3)}$$

where we are already replacing that at the optimum, transfers will be chosen so that $U(\underline{\theta}) = 0$. Recall from OPT that the seller's revenue is given by $\mathbb{E}_{F_1} \left[\int_{\Delta(\Theta)} \left(x(F_2) + (1 - q(F_2)) \delta p_2(F_2) q_2^*(\theta, F_2) \right) \beta(dF_2|\theta) \right]$. Replacing the transfers (Equation C.3) and integrating by parts, we obtain:

$$\int_\Theta \int_{\Delta(\Theta)} \left[q(F_2) J(\theta, F_1) + (1 - q(F_2)) \delta q_2^*(\theta, F_2) J(\theta, F_1) \right] \beta(dF_2|\theta) F_1(d\theta). \quad \text{(C.4)}$$

Denote by P the distribution on $\Theta \times \Delta(\Theta)$ defined as $P(\widetilde{\Theta} \times \widetilde{U}) = \int_{\widetilde{\Theta}} \beta(\widetilde{U}|\theta) F_1(d\theta)$, for all measurable subsets $\widetilde{\Theta}, \widetilde{U}$ of Θ and $\Delta(\Theta)$. Let $P_{\Delta(\Theta)}$ denote its marginal on $\Delta(\Theta)$, Pollard (2002, Appendix F, Theorem 6) implies Equation C.4 equals

$$\int_{\Delta(\Theta)} \left[q(F_2) \int_\Theta J(\theta, F_1) F_2(d\theta) + (1 - q(F_2)) \delta \int_{p_2(F_2)}^{\overline{\theta}} J(\theta, F_1) F_2(d\theta) \right] P_{\Delta(\Theta)}(dF_2),$$

which is the expression in Equation 2.

Monotonicity: Together with the envelope representation of the buyer's payoffs, the mechanism must also satisfy the following monotonicity constraint for all $\theta \geq \theta'$:

$$\int_{\Delta(\Theta)} \left[(\theta - \theta') q(F_2) + \delta(1 - q(F_2)) \int_{\theta'}^{\theta} q_2^*(u, F_2) \, du \right] \left(\beta(dF_2|\theta) - \beta(dF_2|\theta') \right) \geq 0. \tag{C.5}$$

Proof of Proposition 2. We first establish properties that the optimal posted-price mechanism satisfies. Let $F_2^{\hat{\theta}}$ denote the distribution of θ conditional on $\theta \leq \hat{\theta}$. Since $J(\theta, F_1)$ is strictly increasing, it follows that $J(\theta, F_2^{\hat{\theta}})$ is also strictly increasing. Furthermore, since $\underline{\theta} = 0$, $J(\underline{\theta}, F_2^{\hat{\theta}}) < 0 < J(\hat{\theta}, F_2^{\hat{\theta}})$ and there exists a unique $p_2(\hat{\theta})$ such that $J(p_2(\hat{\theta}), F_2^{\hat{\theta}}) = 0$. It follows that $p_2(\hat{\theta}) = \arg\max_{p_2} p_2(1 - F_2^{\hat{\theta}}(p_2))$. The theorem of the maximum implies that $p_2(\hat{\theta})$ is continuous in $\hat{\theta}$. Thus, the optimal posted-price mechanism solves

$$R(\delta) = \max_{\hat{\theta}} \int_{\hat{\theta}}^{\overline{\theta}} J(\theta, F_1) F_1(d\theta) + \delta \int_{p_2(\hat{\theta})}^{\hat{\theta}} J(\theta, F_1) F_1(d\theta). \tag{C.6}$$

The theorem of the maximum implies that a solution exists. Note that $R(0) = R(1) = \int_{p_M}^{\overline{\theta}} J(\theta, F_1) F_1(d\theta)$, where p_M is the monopoly price, that is, $J(p_M, F_1) = 0$. For each $\delta \in (0,1)$, let $\hat{\theta}_\delta^*$ denote a solution to the problem in Equation C.6 and define

$$V(\delta', \delta) = \int_{\hat{\theta}_{\delta'}^*}^{\overline{\theta}} J(\theta, F_1) F_1(d\theta) + \delta \int_{p_2(\hat{\theta}_{\delta'}^*)}^{\hat{\theta}_{\delta'}^*} J(\theta, F_1) F_1(d\theta) \equiv \int_{\hat{\theta}_{\delta'}^*}^{\overline{\theta}} J(\theta, F_1) F_1(d\theta) + \delta C(\delta').$$

Note that $R(\delta) = \sup_{\delta' \in [0,1]} V(\delta', \delta)$. Furthermore, $R(\delta) \geq V(\delta', \delta)$ and $R(\delta') \geq V(\delta, \delta')$ imply that if $\delta > \delta'$, then $C(\delta') \leq C(\delta)$. Note that $C(0) < 0 < C(1) = R(1)$ since the unique solution at $\delta = 0$ is to set $\hat{\theta}_0^* = p_M$. Steps similar to those leading to the envelope representation of the buyer's payoffs imply that (i) $\{V(\delta', \cdot) : \delta' \in [0,1]\}$ is equi-Lipschitz continuous, (ii) for all $\delta' \in [0,1]$, $V(\delta', \cdot)$ is differentiable on $(0,1)$ with derivative equal to $C(\cdot)$, and (iii) R is Lipschitz continuous and at any point of differentiability, $R'(\delta) = C(\delta)$.

We now show that $\overline{\delta} \in (0,1)$ exists so that $C(\delta) > 0$ for $\delta \in (\overline{\delta}, 1]$. Toward a contradiction, suppose that $C(\delta) \leq 0$ for all $\delta \in (0,1)$. Then, R is decreasing on $(0,1)$ and because $R(0) = R(1)$ it follows that $C(\cdot) \equiv 0$ almost everywhere. Otherwise, a set $D \subseteq (0,1)$ of non-zero Lebesgue measure exists such that $C(u) < 0$ for $u \in D$. Then,

$$R(1) = R(0) + \int_0^1 C(u) \, du = R(0) + \int_{[0,1] \setminus D} C(u) \, du + \int_D C(u) \, du \leq R(0) + \int_D C(u) \, du < R(0),$$

where the first inequality follows from $C(u) \leq 0$ on $(0,1)$ and the second by the definition of

D. It follows that $C \equiv 0$ on $(0,1)$ and hence $R(\cdot)$ is constant and equal to the commitment payoff for all $\delta \in (0,1)$, a contradiction. Thus, a subset of $(0,1)$ exists on which C is strictly positive and because C is increasing, $\overline{\delta} \in (0,1)$ exists so that $C(\delta) > 0$ for $\delta \in (\overline{\delta}, 1]$.

Second, we use the above properties to construct a deviation for the seller. Define $\overline{\delta} \in (0,1)$ as above and let $\delta > \overline{\delta}$. We show $\epsilon, \eta, \gamma > 0$ exist such that the seller can deviate to the following mechanism. Let $p^*_{2\delta} \equiv p_2(\hat{\theta}^*_\delta)$ and define three posterior beliefs:

$$F^1_{2S}(\theta) = \frac{F_1(\theta) - F_1(\hat{\theta}^*_\delta)}{1 - F_1(\hat{\theta}^*_\delta)} \mathbb{1}[\theta \geq \hat{\theta}^*_\delta], \quad F^\epsilon_{2S}(\theta) = \frac{F_1(\theta) - F_1(p^*_{2\delta} - \eta)}{F_1(\hat{\theta}^*_\delta) - F_1(p^*_{2\delta} - \eta)} \mathbb{1}\left[\theta \in [p^*_{2\delta} - \eta, \hat{\theta}^*_\delta]\right]$$

$$F_{2D}(\theta) = \begin{cases} \frac{F_1(\theta)}{F_1(p^*_{2\delta}-\eta)+(1-\epsilon)(F_1(\hat{\theta}^*_\delta)-F_1(p^*_{2\delta}-\eta))} & \text{if } \theta < p^*_{2\delta} - \eta \\ \frac{F_1(p^*_{2\delta}-\eta)+(1-\epsilon)(F_1(\theta)-F_1(p^*_{2\delta}-\eta))}{F_1(p^*_{2\delta}-\eta)+(1-\epsilon)(F_1(\hat{\theta}^*_\delta)-F_1(p^*_{2\delta}-\eta))} & \text{if } \theta \in [p^*_{2\delta} - \eta, \hat{\theta}^*_\delta] \end{cases},$$

and let $q(F^1_{2S}) = q(F^\epsilon_{2S}) = 1$ and $q(F_{2D}) = 0$. Furthermore, let $\beta(\{F^1_{2S}\}|\theta) = \mathbb{1}[\theta > \hat{\theta}^*_\delta]$, and let

$$\beta(\{F_{2D}\}|\theta) = \begin{cases} 1 & \text{if } \theta < p^*_{2\delta} - \eta \\ 1-\epsilon & \text{if } \theta \in [p^*_{2\delta} - \eta, \hat{\theta}^*_\delta] \\ 0 & \text{otherwise} \end{cases}, \beta(\{F^\epsilon_{2S}\}|\theta) = \begin{cases} 0 & \text{if } \theta < p^*_{2\delta} - \eta \\ \epsilon & \text{if } \theta \in [p^*_{2\delta} - \eta, \hat{\theta}^*_\delta] \\ 0 & \text{otherwise} \end{cases}.$$

Finally, for $\gamma > 0$, define payments as follows:

$$x(F_{2D}) = -\gamma\epsilon, x(F^\epsilon_{2S}) = -\gamma\epsilon + p^*_{2\delta} - \eta, x(F^1_{2S}) = -\gamma\epsilon + \epsilon(p^*_{2\delta} - \eta) + (1-\epsilon)((1-\delta)\hat{\theta}^*_\delta + \delta p^*_{2\delta}).$$

That is, the seller now serves with probability ϵ all types in $[p^*_{2\delta} - \eta, \hat{\theta}^*_\delta]$, and leaves the probability of trade for the remaining types unchanged; furthermore, he pays the buyer $\gamma\epsilon$, regardless of her type. The value of η is chosen so that $p^*_{2\delta}$ remains optimal in period 2.

In what follows, we first argue that for any agent-extensive form that is feasible given the above mechanism in period 1 and in any agent-assessment of that extensive form, the buyer must accept the above mechanism with probability 1. That, conditional on participating, truthtelling is optimal is immediate. We then show that given ϵ we can always find η so that $p^*_{2\delta}$ remains optimal in period 2. Finally, we show that for ϵ, γ small, the seller prefers the mechanism defined above to the optimal posted-price mechanism.

The buyer participates with probability 1**:** Note first that buyer types in $[\underline{\theta}, \underline{\theta} + \gamma\epsilon)$ must accept the mechanism: the best price they can obtain in period 2 is $\underline{\theta}$, whereas they obtain at least $\gamma\epsilon$ by participating in the mechanism offered by the seller. Since $\gamma\epsilon > \theta - \underline{\theta}$ for all such types, they must accept the mechanism. It then follows that, conditional on rejecting

the mechanism, the price is at least $\gamma\epsilon + \underline{\theta}$ in period 2, so that it must be the case that types in $[\underline{\theta} + \gamma\epsilon, \underline{\theta} + 2\gamma\epsilon)$ must also accept the mechanism. Proceeding inductively, it follows that the buyer accepts the mechanism with probability 1. Thus, Bayes' rule does not pin down beliefs conditional on non-participation, so that we can specify them arbitrarily: In particular, we can specify them so that the seller assigns probability 1 to $\overline{\theta}$.

$p^*_{2\delta}$ remains optimal in period 2: The virtual values under F_{2D} are given by

$$J(\theta, F_{2D}) = \begin{cases} \theta - \frac{(1-\epsilon)F_1(\hat{\theta}^*_\delta) + \epsilon F_1(p^*_{2\delta} - \eta) - F_1(\theta)}{f_1(\theta)} & \text{if } \theta \leq p^*_{2\delta} - \eta \\ J(\theta, F_2^{\hat{\theta}^*_\delta}) & \text{if } \theta > p^*_{2\delta} - \eta \end{cases},$$

so that the virtual values under F_{2D} (i) are piecewise increasing in θ on $[\underline{\theta}, \hat{\theta}^*_\delta]$, (ii) are (weakly) negative on $(p^*_{2\delta} - \eta, p^*_{2\delta}]$, and (iii) they jump down at $\theta = p^*_{2\delta} - \eta$, whenever $\epsilon > 0$. Let $H(\eta, \epsilon)$ denote the limit from below of $J(\theta, F_{2D})$ as $\theta \to p^*_{2\delta} - \eta$. That is,

$$H(\eta, \epsilon) \equiv p^*_{2\delta} - \eta - \frac{(1-\epsilon)(F_1(\hat{\theta}^*_\delta) - F_1(p^*_{2\delta} - \eta))}{f_1(p^*_{2\delta} - \eta)}.$$

Using the strict monotonicity of the virtual values and that $\underline{\theta} = 0$, it is immediate to show that for each $\epsilon > 0$, $\eta(\epsilon)$ exists such that $H(\eta(\epsilon), \epsilon) = 0$. Thus, for $\eta = \eta(\epsilon)$, $J(\theta, F_{2D})$ is (weakly) negative for types θ below $p^*_{2\delta} - \eta$, and hence $p^*_{2\delta}$ remains optimal. In what follows, it is important to note $\eta(\cdot)$ is increasing in ϵ and the definition of $p^*_{2\delta}$ implies $\eta(0) = 0$.

The seller has a deviation: We now verify that for small enough ϵ, γ the seller prefers the above mechanism to the optimal posted price mechanism. Indeed, the difference in payoffs between the new mechanism and the optimal posted-price mechanism is as follows:

$$\epsilon \left[(1-\delta) \int_{p^*_{2\delta}}^{\hat{\theta}^*_\delta} J(\theta, F_1) F_1(d\theta) + \int_{p^*_{2\delta} - \eta(\epsilon)}^{p^*_{2\delta}} J(\theta, F_1) F_1(d\theta) - \gamma \right].$$

By assumption, the first term in the square brackets is positive and independent of ϵ, η, γ, whereas for small ϵ, the second term, although negative, is vanishing, and γ is also small. Thus, for $\delta > \overline{\delta}$, small enough ϵ, γ exist such that the seller has a deviation. □