Testing Rational Hypotheses in Signaling Games*

Adam Dominiak† and Dongwoo Lee‡

May 6, 2023

Abstract

We introduce a solution concept for signaling games, called Rational Hypothesis Testing Equilibrium (RHTE). Beliefs are updated via Ortoleva's (2012) Hypothesis Testing model, allowing for conditioning on information sets off the path. Hypotheses are conjectures by the uninformed player about opponent's strategies that rationalize sending an unexpected message. Each RHTE is a Perfect Bayesian Equilibrium, but not vice versa. RHTE features a number of desirable properties: First, beliefs are structurally consistent in the spirit of Kreps and Wilson (1982). Second, beliefs are consistent with mutual knowledge of rationality. Third, RHTE can be related to the prominent refinement concepts, including the Intuitive Criterion, strategic stability, and undefeated equilibrium. In the Spence game, RHTE restricts the admissible wages, significantly reducing the number of equilibria. Finally, we show that our equilibrium notion offers an alternative explanation for the experimental results in Brandts and Holt (1992, 1993).

Keywords: Signaling games, perfect Bayesian equilibrium, updating, off-path beliefs, hypothesis testing, rationality, refinements, Intuitive Criterion, strategic stability, undefeated equilibrium.
JEL Classification: C72, C73, D81, D83

*The authors are grateful to George Mailath, Hans Haller, Per Overgaard, Matthew Kovach, Gerelt Tserenjigmid, Kevin He, the audiences of the 28th International Conference on Game Theory, the 88th Southern Economic Association Meetings, and the Australian National University for their valuable comments and fruitful discussions. Parts of this paper are based on Chapter 3 in Lee's doctoral thesis written in the Department of Economics at Virginia Tech.

†Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus, Denmark. E-mail: adamdominiak@econ.au.dk; Department of Economics, Virginia Tech, Blacksburg, VA 24061, USA.

‡Corresponding Author: China Center for Behavioral Economics and Finance, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China. E-mail: dwlee05@gmail.com.

1 Introduction

Signaling games are an important class of dynamic games with incomplete information. Signaling refers to interactive situations in which an uninformed agent uses observable actions of an opponent to make inferences about hidden information. Signaling games have been widely applied to explain a variety of economic phenomena including job search (Spence, 1973), advertising (Nelson, 1974; Milgrom and Roberts, 1986), dividends (Bhattacharya, 1979; John and Williams, 1985), product quality (Miller and Plott, 1985), warranties (Gal-Or, 1989), limit pricing (Milgrom and Roberts, 1982), elections (Banks, 1990), social norms (Bernheim, 1994), or lobbying (Lohmann, 1995).

A Perfect Bayesian Equilibrium (PBE), or equivalently, sequential equilibrium (Kreps and Wilson, 1982) is a standard solution concept for signaling games.[1] A Sender, knowing his type, sends an optimal message. An uninformed Receiver only knows a prior over types of the Sender. Conditional on each message, the Receiver forms beliefs about types via Bayesian updating and responds with an optimal action. Since Bayes' rule does not specify how beliefs are determined at information sets off the equilibrium path, PBE admits arbitrary beliefs, leading to multiplicity of equilibria.

To discipline beliefs, we introduce a solution concept called *Rational Hypothesis Testing Equilibrium* (in short, *rational equilibrium*). The Receiver forms beliefs about types via the Hypothesis Testing model, axiomatized by Ortoleva (2012), allowing for conditioning on *all* information sets.

For each message, the Receiver selects the most-likely hypothesis according to her second-order prior over a set of hypotheses, and updates it by Bayes' rule. We require hypotheses to be about rational behavior. A rational hypothesis is a conjecture about a strategy for the Sender, which is calibrated by the prior over types, that best responds to some of the Receiver's rational strategies.

On the equilibrium path, the most-likely (initial) hypothesis conjectures that the informed opponent follows his equilibrium strategy which best responds to the Receiver's equilibrium strategy. When an unexpected message arrives, the Receiver rejects the initial hypothesis. She updates her second-order prior by Bayes' rule and selects a new hypothesis in the maximum likelihood-fashion. Off the path, each message is viewed as the outcome of some disequilibrium strategy of the Sender who mistakenly best responds to other rational strategy than the Receiver's equilibrium strategy.

Rational equilibrium is a PBE that adheres to *rational consistency*. It is a new consistency requirement on beliefs that strengthens the notion of *structural consistency*, suggested by Kreps and Wilson (1982) as a criterion to justify sequential equilibrium in general extensive-form games. Structural consistency requires each belief to be the Bayesian update of a single behavioral strategy that governed the previous moves. The authors even provided a hypothesis-testing interpretation of structural consistency which Kreps and Ramey (1987, p.1332) subsumed in the following way:

> "*... the player who is moving should posit some single strategy combination which, in*

[1] In signaling games, the two solution concepts coincide (Fudenberg and Tirole, 1991b, Proposition 3.2).

> *his view, has determined moves prior to his information set, and that his beliefs should be Bayes-consistent with this hypothesis. If the information set is reached with positive probability in equilibrium, then beliefs are formed using the equilibrium strategy. If, however, the information set lies off the equilibrium path, then the player must form some single "alternative hypothesis" as to the strategy governing prior play, such that under the hypothesis the information set is reached with positive probability."*

Rational equilibrium formalizes this interpretation under two additional requirements: (i) each "alternative hypothesis" is about a (second-order) rational strategy that governed the unexpected signaling; (ii) each "alternative hypothesis" is consistent with the prior over types and so are beliefs at information sets off the equilibrium path. These two requirements define rational consistency.

A PBE may fail rational consistency for two reasons. First, beliefs, or strategies used to justify the beliefs, are inconsistent with implications from the chief assumption in game theory that the two players mutually know their rationality (Aumann and Brandenburger, 1995). Second, beliefs are inconsistent with the prior information about types which is a primitive of signaling games.

Rational equilibrium substantially refines PBE in the education signaling game by Spence (1973). Rational consistency implies that a rational employer offers the highest wage for each education level off the path that only the rational worker with high productivity can choose. Whether a pooling rational equilibrium exists depends on the prior information about workers' productivity, eliminating all pooling PBEs if the fraction of high-productivity workers is sufficiently small. A separating rational equilibrium always exists and supports the efficient Riley (1979) outcome.

Rational equilibrium relates to prominent refinement concepts such as the Intuitive Criterion (Cho and Kreps, 1987), strategic stability (Kohlberg and Mertens, 1986), $D1$ Criterion (Banks and Sobel, 1987), and undefeated equilibrium (Mailath, Okuno-Fujiwara, and Postlewaite, 1993). Foremost, we prove the existence of a rational equilibrium under the same condition that the above criteria require: there is at least one Sender type that could benefit from defecting to an unsent message. This condition also implies that rational equilibrium entails strategic stability. However, rational equilibrium is neither nested with intuitive equilibrium nor with undefeated equilibrium. Under a mild condition, we establish the relationship with the Intuitive Criterion. Furthermore, we show that one can invoke the "is-not-defeated-by-another-PBE-test" as a condition testing whether the undefeated PBE is a rational equilibrium, facilitating applications of our solution concept.

Finally, we demonstrate that our solution concept can explain the main experimental findings in Brandts and Holt (1992, 1993). Both studies found an intriguing pattern of behavior where intuitive equilibrium predominates in some games, while unintuitive equilibrium prevails in other games. The authors attributed this phenomenon to "type-dependence". It is the salient separating strategy in a given game that triggers adjustments in behavior leading either to intuitive or to unintuitive equilibrium. We argue that "type-dependence" induces the most-likely rational hypothesis off

paths, showing that rational equilibrium offers an alternative account for the intriguing behavior.

This paper is organized as follows. Section 2 recalls the Hypothesis Testing model and defines rational consistency. Section 3 introduces rational equilibrium. Section 4 solves the Spence game. Section 5 compares rational equilibrium with intuitive equilibrium and undefeated equilibrium, respectively. Section 6 discusses the experimental findings in Brandts and Holt (1992, 1993). Section 7 concludes this work. Appendix A proves the existence of a rational equilibrium that is strategically stable. Appendix B deals with the $D1$ Criterion. Appendix C collects all proofs.

2 Preliminaries

2.1 Signaling Games

A signaling game consists of two players, called the *Sender* (he) and the *Receiver* (she). Nature draws a type for the Sender from a finite set of types Θ according to a prior probability distribution p on Θ. We assume that p has full support (i.e., $supp(p) = \Theta$), and is known by the players. The Sender observes his type and chooses a message m from a finite set $\mathcal{M}$. The Receiver observes the message, but not the type, chooses an action a from a finite set $\mathcal{A}$, and the game ends. Payoffs are given by $u_S, u_R : \Theta \times \mathcal{M} \times \mathcal{A} \to \mathbb{R}$. The class of finite signaling games is denoted by $\mathcal{G}$.

A behavioral strategy for the Receiver is a collection of message-contingent mixtures over actions, denoted by $b_R := (b_R(\cdot|m))_{m\in\mathcal{M}}$ (i.e., $\sum_{a\in\mathcal{A}} b_R(a|m) = 1$ for each $m \in \mathcal{M}$, where $b_R(a|m)$ is the probability that a is played in response to m). A pure strategy refers to a degenerate b_R (i.e., for each $m \in \mathcal{M}$, $b_R(a|m) = 1$ for some $a \in \mathcal{A}$). For each $m \in \mathcal{M}$, $\mu(\cdot|m) \in \Delta(\Theta)$ denotes a conditional belief (posterior) over types given m. We denote by $\mu := \{\mu(\cdot|m)\}_{m\in\mathcal{M}}$ a system of posteriors. A strategy b_R is (first-order) rational if for each $m \in \mathcal{M}$, any $a \in \mathcal{A}$ with $b_R(a|m) > 0$ is a best response with respect to some belief $\mu(\cdot|m)$ over Θ; i.e.,

$$b_R(a|m) > 0 \text{ implies } a \in BR(\Theta, m) := \bigcup_{\mu(\cdot|m)\in\Delta(\Theta)} BR(\mu, m), \tag{1}$$

where

$$BR(\mu, m) := \arg\max_{a\in\mathcal{A}} \sum_{\theta\in\Theta} \mu(\theta|m) u_R(\theta, m, a) \tag{2}$$

is the set of (pure) best responses with respect to μ. $BR(J, m)$ is the set of best responses for beliefs concentrated over a subset J of Θ. Furthermore, $MBR(\mu, m)$ and $MBR(J, m)$ denote the sets of mixed strategy best responses.[2] $\mathcal{B}_R^\bullet$ stands for the set of rational strategies for the Receiver.

[2] $MBR(\mu, m) := \arg\max_{b_R(\cdot|m)\in\Delta(\mathcal{A})} \sum_{a\in\mathcal{A}} \sum_{\theta\in\Theta} \mu(\theta|m) u_R(\theta, m, a) b_R(a|m)$.

A behavioral strategy for the Sender is a collection of type-contingent mixtures over messages, denoted by $b_S := (b_S(\cdot|\theta))_{\theta\in\Theta}$ (i.e., $\sum_{m\in\mathcal{M}} b_S(m|\theta) = 1$ for each $\theta \in \Theta$, where $b_S(m|\theta)$ denotes the probability that θ sends m). $\mathcal{B}_S = [\Delta(\mathcal{M})]^\Theta$ is the set of all such strategies. A degenerate b_S (i.e., for each $\theta \in \Theta$, $b_S(m|\theta) = 1$ for some $m \in \mathcal{M}$) is a pure strategy. A strategy b_S is (second-order) rational if it is a best response to some $b_R \in \mathcal{B}_R^\bullet$; i.e., for each $\theta \in \Theta$ and $m \in \mathcal{M}$,

$$b_S(m|\theta) > 0 \text{ implies } m \in \arg\max_{m'\in\mathcal{M}} \sum_{a\in\mathcal{A}} u_S(\theta, m', a)\, b_R(a|m'). \tag{3}$$

$\mathcal{B}_S^\bullet$ denotes the set of rational strategies for the Sender. A message $m^d \in \mathcal{M}^d$ is (strictly) dominated if for each $\theta \in \Theta$, playing m^d is a never-best response.

A *Perfect Bayesian Equilibrium* (PBE) consists of a strategy profile (b_S^*, b_R^*) and beliefs $\mu^* = \{\mu^*(\cdot|m)\}_{m\in\mathcal{M}}$ where b_S^* best responds to b_R^* and b_R^* best responds to b_S^* with respect to the conditional belief μ^* over Θ, which is derived via Bayesian updating whenever it is possible; i.e.,

$$\mu^*(\theta|m) = \frac{b_S^*(m|\theta)p(\theta)}{\sum_{\theta'\in\Theta} b_S^*(m|\theta')p(\theta')} \text{ for each } \theta \in \Theta \text{ if } \sum_{\theta'\in\Theta} b_S^*(m|\theta')p(\theta') > 0, \text{and} \tag{4}$$

$$\mu^*(\cdot|m) \text{ is an arbitrary probability distribution over } \Theta \text{ if } \sum_{\theta'\in\Theta} b_S^*(m|\theta')p(\theta') = 0. \tag{5}$$

For a message m° such that $\sum_{\theta\in\Theta} b_S^*(m^\circ|\theta)p(\theta) = 0$, we say that m° is off the path (unsent message).

2.2 Hypothesis Testing Model

We recapitulate Ortoleva's (2012) Hypothesis Testing model in the framework of signaling games.

The main component of Ortoleva's model is a set of hypotheses. Denote by $\beta := (\beta(\cdot|\theta))_{\theta\in\Theta}$ a system of type-contingent probability distribution on $\mathcal{M}$, where $\theta \in \Theta$, $\beta(\cdot|\theta)$ represents the Receiver's conjecture about message choices by type θ. Note that $\beta \in \mathcal{B}_S$. A system of beliefs (in short, *belief*) β combined with the prior information about types, p, defines a *hypothesis*.

Definition 1 (Hypothesis) *A* hypothesis π *is the probability distribution on* $\mathcal{M} \times \Theta$ *induced by a belief* $\beta \in \mathcal{B}_S$ *and the prior probability distribution* $p \in \Delta(\Theta)$*; i.e., for each* $(m, \theta) \in \mathcal{M} \times \Theta$*:*

$$\pi(m, \theta) = \beta(m|\theta)p(\theta). \tag{6}$$

A hypothesis π ascribes probability $\pi(m, \theta)$ to the state: "type θ signals m." We say that m is feasible under π if $\pi(m, \theta) = \beta(m|\theta)p(\theta) > 0$ for some θ; i.e., the Receiver believes that her opponent plays a strategy according to which type θ signals m with a strictly positive probability.[3]

[3]Recall, we assume $supp(p) = \Theta$.

Note that, by construction, each hypothesis is consistent with the prior information p. That is,

$$\pi(\mathcal{M}, \theta) = \sum_{m \in \mathcal{M}} \pi(m, \theta) = \sum_{m \in \mathcal{M}} \beta(m|\theta) p(\theta) = p(\theta). \tag{7}$$

A hypothesis π is called *simple* if β is a system of degenerate beliefs (i.e., for each $\theta \in \Theta$, $\beta(m|\theta) = 1$ for $m \in \mathcal{M}$). In this case, the Receiver believes that the Sender plays a pure strategy.

We assume that the Receiver selects and updates hypotheses in the following way. For a signaling game in $\mathcal{G}$, we denote by $\Delta(\mathcal{M} \times \Theta)$ the set of all probability measures on $\mathcal{M} \times \Theta$. Let $\Pi \subset \Delta(\mathcal{M} \times \Theta)$ be the set of all hypotheses associated with the game. The Receiver holds a *second-order prior* over Π, denoted by ρ. The support of ρ is finite (i.e., $|supp(\rho)| \in \mathbb{N}$). We assume that ρ induces a strict partial order over $supp(\rho)$. Before any information is revealed, the Receiver selects an initial hypothesis π^* which is the most likely hypothesis with respect to ρ, i.e.,

$$\{\pi^*\} := \arg\max_{\pi \in supp(\rho)} \rho(\pi). \tag{8}$$

Upon arrival of a message m, the Receiver conducts a test. If m is feasible under π^*, she accepts π^* and updates it via Bayes' rule. However, if m is not feasible under π^* (i.e., $\pi^*(m, \Theta) = 0$), the Receiver rejects π^*. Then, she updates her second-order prior ρ via Bayes' rule. We assume that ρ_m, the Bayesian update of ρ given m is a strict partial order over $supp(\rho_m)$ for each $m \in \mathcal{M}$. The Receiver selects a new hypothesis π_m^{**} which is the most likely hypothesis according to ρ_m; i.e.,

$$\{\pi_m^{**}\} := \arg\max_{\pi \in supp(\rho)} \rho_m(\pi) \quad \text{where} \quad \rho_m(\pi) = \frac{\pi(m, \Theta)\rho(\pi)}{\sum_{\pi' \in supp(\rho)} \pi'(m, \Theta)\rho(\pi')}, \tag{9}$$

and updates it via Bayes' rule to determine her posterior over Θ. Posteriors are well-defined if for each $m \in \mathcal{M}$, there exists a hypothesis $\pi \in supp(\rho)$ under which m is feasible (i.e., $\pi(m, \Theta) > 0$).

A second-order prior ρ is called *focused* if its support contains only hypotheses that are used.[4] That is,

$$supp(\rho) := \{\pi^*\} \cup \bigcup_{\substack{m \in \mathcal{M} \ s.t. \\ \pi^*(m,\Theta)=0}} \{\pi_m^{**}\}, \tag{10}$$

where π_m^{**} is the alternative hypothesis conditional on m with zero-probability according to π^*. This is the essence of the Hypothesis Testing model.[5] In Section 3, we suggest a solution concept

[4] A focused ρ (i.e., a strict partial order) guarantees that (conditional) beliefs are unique in the following sense. There exists a strict partial order $\triangleright$ on Π, such that for any other (focused) second-order prior ρ' on Π with $supp(\rho) = supp(\rho')$, $\pi \triangleright \pi'$ implies $\rho(\pi) < \rho(\pi')$ and $\rho'(\pi) < \rho'(\pi')$. Thus, the updated beliefs under ρ' are the same as the ones associated with ρ, avoiding multiplicity of equilibria due to multiplicity of strict partial orders (see Section 3).

[5] Strictly speaking, we consider a special case of the Hypothesis Testing model. In the general version, an initial

that incorporates this model of updating.[6] However, before doing it, we need to elaborate more on the meaning of hypotheses which are used to justify beliefs on and off the equilibrium paths.

2.3 Structural versus Rational Consistency

Now, we introduce two notions of structural consistency in the spirit of Kreps and Wilson (1982).

Structural consistency requires that for each message (i.e., information set), each conditional belief should be derived from a single strategy for the Sender, according to which the message is sent with a (strictly) positive probability, via Bayes' rule. In addition, we require conditional beliefs to be consistent with the prior probability distribution over the Sender types (see Definition 1). A system of beliefs $\mu := \{\mu(\cdot|m)\}_{m\in\mathcal{M}}$ is said to be *structurally consistent* if, for each $m \in \mathcal{M}$, $\mu(\cdot|m)$ is the Bayesian update of some hypothesis π conditional on m.

Definition 2 (Structural Consistency) *For a given signaling game, a system of beliefs μ adheres to structural consistency if for each message $m \in \mathcal{M}$, there exists a hypothesis π with $\pi(m, \theta) = \beta(m|\theta)p(\theta) > 0$ for some $\theta \in \Theta$, and $\mu(\cdot|m)$ is the Bayesian update of π conditional on m.*

In signaling games, PBE beliefs are structurally consistent.[7]

Lemma 1 *Let (b_S^*, b_R^*, μ^*) be a PBE. Then, μ^* is structurally consistent.*

Structural consistency justifies all PBE beliefs. Using structural consistency as a criterion to refine PBE is too weak. In particular, it admits beliefs that can only be derived from "irrational" strategies (i.e., never-best responses). Such beliefs are inconsistent with the common assumption that players are rational and thus never play never-best responses (Bernheim, 1994). Moreover, under the assumption of players' mutual knowledge of rationality (i.e., each player is rational and believes that the opponent player is rational), each player best-responds with respect to beliefs that concentrate on the set of rational strategies of the opponent, other strategies are deemed impossible.

To maintain consistency with this assumption, one needs to require that the Receiver derives beliefs from her opponent's strategies that best respond to some of her own rational strategies. Thus, we require hypotheses to be conjectures about rational behavior, called *rational hypotheses*.

belief is rejected if the surprise event has a probability equal or smaller than $\epsilon \geq 0$. Such a model is said to be *minimal* if any $\epsilon' < \epsilon$ leads to different decisions than under ϵ (Ortoleva, 2012, Definition 3). In our setup, the initial hypothesis is rejected off the path (i.e., $\epsilon = 0$). Thus, we assume the Minimal Focused Hypothesis Testing model. The advantage of this model specification is its unique representation (Ortoleva, 2012, Proposition 2). It is important to remark that our results carry over to a model assuming $\varepsilon \geq 0$ since pooling equilibrium will be exactly the same as under $\epsilon = 0$.

[6] Galperti (2019) applies the Hypothesis Testing model to study optimal persuasion under strategic information design. The Sender can confirm or disconfirm the Receiver's understanding of a prior. The author explores different ways how the Receiver forms a new prior in light of unexpected signaling by the Sender.

[7] In general, however, this is not true. Kreps and Ramey (1987) provided examples for extensive-form games for which sequential-equilibrium beliefs violate structural consistency.

Definition 3 (Rational Hypothesis) *A rational hypothesis π is the joint probability distribution on $\mathcal{M} \times \Theta$ induced by a belief $\beta \in \mathcal{B}_S^\bullet$ and the prior $p \in \Delta(\Theta)$; i.e., for every $(m, \theta) \in \mathcal{M} \times \Theta$,*

$$\pi(m, \theta) = \beta(m|\theta)p(\theta). \tag{11}$$

Beliefs satisfy *rational consistency* if they are Bayesian updates of rational hypotheses.

Definition 4 (Rational Consistency) *A system of beliefs μ adheres to rational consistency with respect to a signaling game if for each $m \in \mathcal{M}$, there exists a rational hypothesis π such that $\pi(m, \theta) = \beta(m|\theta)p(\theta) > 0$ for some θ and $\mu(\cdot|m)$ is the Bayesian update of π conditional on m.*

A PBE is said to adhere to rational consistency if its beliefs are rationality consistent. An equilibrium may fail rational consistency for two reasons. First, some belief is inconsistent with mutual belief in players' rationality. Second, rational hypotheses exist, but none of them justifies a given PBE belief. In this case, the belief is inconsistent with the prior information about types. Below, we illustrate both instances.

2.4 Failures of Rational Consistency

To illustrate violations of rational consistency, we consider two games. In Example 1, we show that the education signaling game by Spence (1973) has pooling equilibria with beliefs that are inconsistent with mutual belief in rationality. In Example 2, we derive equilibria in the first game studied by Brandts and Holt (1992) of which one has beliefs that are inconsistent with the prior information.[8]

Example 1 (The Spence Game)

We consider a finite version of the signaling game studied by Spence (1973), called Spence game. There is a worker (he) and an employer (she). The worker has either low (L) or high (H) productivity; i.e., $\Theta = \{\theta_L, \theta_H\}$, where $\theta_L < \theta_H$. Let $p(\theta_L) = 1 - \alpha$ and $p(\theta_H) = \alpha \in (0, 1)$ be the prior on Θ. Knowing his type θ, the worker chooses an education level e from a finite set $\mathcal{M} = \{e_0, e_1, \ldots, e_N\}$, where $e_0 = 0 < e_1 < \ldots < e_N$.[9] The worker's payoff is given by

$$u_S(\theta, e, w) = w - \frac{e}{\theta} \quad \text{for } \theta \in \Theta, \tag{12}$$

where w denotes the wage and $\frac{e}{\theta}$ is the cost of choosing e by type θ. Education is more costly to the low-productivity type. It is assumed that the worker can always find a job at wage $w = \theta_L$. Figure 1 depicts type-dependent indifference curves (the red one for θ_L and the blue one for θ_H).

[8] In Section 6, we solve their second game. The two games are depicted in Figures 2 and 5.

[9] We assume a finite and rich set $\mathcal{M}$ in the sense that two adjacent education levels are very close to each other. In

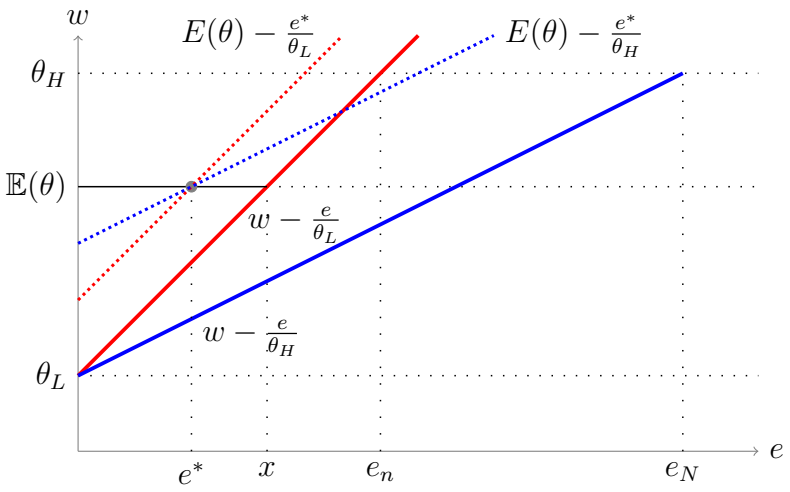

Figure 1: Pooling PBE with education level e^* and wage $w^* = \mathbb{E}(\theta)$.

The employer observes the education level e, but not the worker's type, and offers a wage w. Her payoff is given by

$$u_R(\theta, e, w) = -(\theta - w)^2 \quad \text{for } \theta \in \Theta. \tag{13}$$

A rational employer offers a wage that amounts to expected productivity. That is, the employer's best response to each e is given by $w(e) := \mathbb{E}(\theta|e) = \mu(\theta_H|e)\theta_H + \big(1 - \mu(\theta_H|e)\big)\theta_L$ where $\mu(\cdot|e)$ denotes a posterior belief over Θ. Although $\mathcal{A} = \mathbb{R}_+$, rationality implies that $w(e) \in [\theta_L, \theta_H]$. We denote by $\mathbb{E}(\theta) := \alpha\theta_H + (1-\alpha)\theta_L$ the average productivity when $\mu(\cdot|e)$ coincides with the prior p. To simplify notation, we write $w(e)$ to denote the employer's pure strategy $b_R(w(e)|e) = 1$.

In this example, we focus on pooling behavior; i.e., both worker types choose the same education level, denoted by e^*. In pooling equilibrium, the payoff of the low-productivity type from choosing e^* at the wage w^* that equals to the average productivity (i.e., $w^* = \mathbb{E}(\theta)$) must be greater than his payoff from choosing $e = 0$ at the minimum wage $w = \theta_L$. Thus, e^* must satisfy

$$u_S(\theta_L, e^*, w^*) = \mathbb{E}(\theta) - \frac{e^*}{\theta_L} \geq \theta_L, \quad \text{or equivalently,} \quad e^* \leq \underbrace{\alpha(\theta_H - \theta_L)\theta_L}_{:=x}, \tag{14}$$

specifying an upper bound x for an education level that can be supported by a Pooling PBE.

For $e^* \leq x$, neither type has an incentive to deviate from e^* at $w^* = \mathbb{E}(\theta)$ as long as the wages paid off the equilibrium paths satisfy the following conditions: For each $e < e^*$, $w(e)$ satisfies

$$u_S(\theta_L, e^*, w^*) = \mathbb{E}(\theta) - \frac{e^*}{\theta_L} \geq w(e) - \frac{e}{\theta_L} = u_S(\theta_L, e, w(e)), \tag{15}$$

particular, we assume that $\mathcal{M}$ contains education levels $e_n := \theta_L(\theta_H - \theta_L)$ and $e_N := \theta_H(\theta_H - \theta_L)$. For an analysis of the Spence model with $\mathcal{M} = \mathbb{R}_+$, see Fudenberg and Tirole (1991a, Chapter 8).

whereas for each $e' > e^*$, $w(e')$ satisfies

$$u_S(\theta_H, e^*, w^*) = \mathbb{E}(\theta) - \frac{e^*}{\theta_H} \geq w(e') - \frac{e'}{\theta_H} = u_S(\theta_H, e', w(e')). \tag{16}$$

Hence, there is a continuum of pooling equilibria. Besides multiple levels of education that can be supported by a pooling equilibrium, each pooling equilibrium admits various wage schemes off the path. For the sake of our argument, we consider the following Pooling PBE: Both workers choose e^* (i.e., $b_S^* = (b_S^*(e^*|\theta_L) = b_S^*(e^*|\theta_H) = 1)$) such that $0 \leq e^* \leq x := \alpha(\theta_H - \theta_L)\theta_L$ and the employer with beliefs $\mu^* := (\mu^*(\cdot|e))_{e\in\mathcal{M}}$ offers the wages $w_R^* := (w^*(e))_{e\in\mathcal{M}}$ given by

$$w^*(e) = \begin{cases} \theta_L & \text{if } 0 \leq e < e^*, \\ \mathbb{E}(\theta) & \text{if } e^* \leq e < e_N, \end{cases} \quad \text{where} \quad \mu^*(\theta_L|e) = \begin{cases} 1 & \text{if } 0 \leq e < e^*, \\ (1-\alpha) & \text{if } e^* \leq e \leq e_N. \end{cases} \tag{17}$$

The employer pays the minimum wage for any education level below e^* believing it is chosen by the low-productivity type. For any education equal to or above e^*, the employer pays the average productivity as she believes it is chosen by both worker types with the respective prior probabilities.

The above equilibrium fails rational consistency. For any education level $e' \in \{e_{n+1}, \ldots, e_N\}$, the employer's belief that both worker types chose e' with the respective prior probabilities (i.e., $\mu^*(\theta_L|e') = (1-\alpha)$ and $\mu^*(\theta_H|e') = \alpha$) cannot be justified by a rational hypothesis. Note that the low-productivity worker has no incentive to choose such an e' even if the highest wage, $w(e') = \theta_H$, were paid. The lowest effort level at the minimum wage makes him strictly better off than e', i.e.,

$$\text{for each } e' > e_n,\ u_S(\theta_L, 0, \theta_L) = \theta_L > \theta_H - \frac{e'}{\theta_L} = u_S(\theta_L, e', \theta_H) \text{ since } e' > \theta_L(\theta_H - \theta_L),$$

To make e' attractive for the low-productivity worker, the employer needs to offer a wage above the highest wage that she could pay, θ_H. However, since such wages cannot be rational, the employer cannot conjecture that workers pool on e' as a best response to some rational wage. For this reason, there is no rational hypothesis according to which θ_L chooses an effort level above e_n. Consequently, for each $e' \in \{e_{n+1}, \ldots, e_N\}$, each $\mu^*(\theta_L|e') > 0$ violates rational consistency.[10]

Paying average productivity for any such e' is inconsistent with mutual belief in rationality. Knowing that the employer is rational, only the (rational) worker with high productivity can choose e' as a best response to a rational wage. Knowing this, the rational employer will offer the highest wage θ_H for each $e' \in \{e_{n+1}, \ldots, e_N\}$, and the rational workers know this fact. Stated differently, the rational employer must believe that any education $e' \in \{e_{n+1}, \ldots, e_N\}$ perfectly correlates with

[10]Note that e_n makes the θ_L-worker indifferent between choosing $e = 0$ for $w(0) = \theta_L$ and e_n for $w(e_n) = \theta_H$ (see Figure 1).

high productivity. Therefore, $\mu(\theta_H|e') = 1$ is the only belief that adheres to rational consistency. In Section 4, we elaborate on the implications of this restriction on pooling behavior in the Spence game, showing that rational consistency substantially reduces the number of pooling equilibria.

In the next example, we illustrate another reason why rational consistency may fail.

Example 2

Consider Game 1 depicted in Figure 2 which represents a simpler version of the Spence game. Again, θ_L and θ_H refer to workers' productivity. Each worker type decides whether to invest in additional education (E) or not (N). An employer either assigns the worker to an executive job (e) or to a manual job (m). The prior is $p(\theta_L) = 1/3$. Each worker type prefers the executive job regardless of his education status. For the employer, education is not productive since her payoff is unaffected by the signal. Therefore, the employer prefers to match the high-productivity worker with the executive job and the low-productivity worker with the manual job.

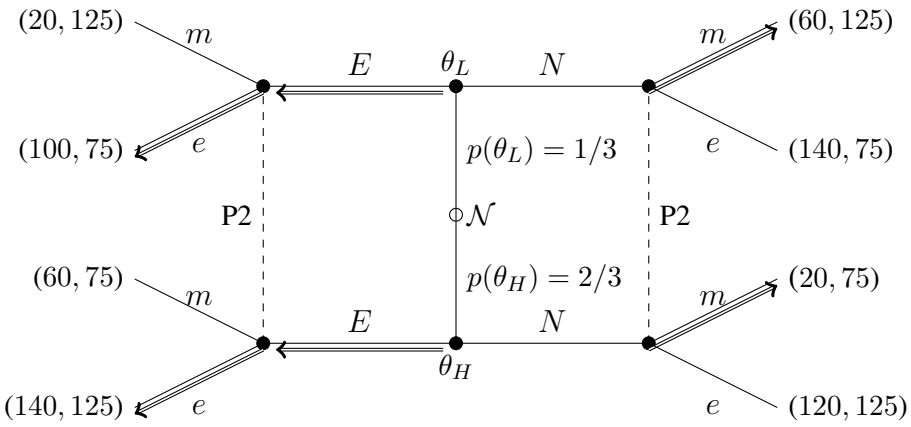

Figure 2: Game 1 in Brandts and Holt (1992) and RHTE-1 ("⟶")

This game has two pooling equilibria. In the first PBE, both worker types pool on E; i.e.,

$$b_S^*(E|\theta_L) = b_S^*(E|\theta_H) = 1,\ b_R^*(e|E) = b_R^*(m|N) = 1, \qquad \text{(PBE-1)}$$
$$\mu^*(\theta_L|E) = 1/3 \text{ and } \mu^*(\theta_L|N) \geq 1/2.$$

In the second PBE, both types pool on N; i.e.,

$$b_S^*(N|\theta_L) = b_S^*(N|\theta_H) = 1,\ b_R^*(m|E) = b_R^*(e|N) = 1, \qquad \text{(PBE-2)}$$
$$\mu^*(\theta_L|E) \geq 1/2 \text{ and } \mu^*(\theta_L|N) = 1/3.$$

PBE-1 adheres to rational consistency, but not PBE-2. To illustrate, consider the former equilibrium. Note that beliefs on the equilibrium path satisfy rational consistency by definition.

Let β_λ be the employer's belief that the high-productivity worker chooses N with a positive probability $\lambda \geq 0$ and the low-productivity worker signals N with certainty; i.e.,

$$\beta_\lambda = (\beta(E|\theta_H) = (1-\lambda), \beta(N|\theta_H) = \lambda, \text{ and } \beta(N|\theta_L) = 1)\,.$$

Since β_λ best responds against $b_R(m|E) = 1/4$, $b_R(e|E) = 3/4$, and $b_R(e|N) = 1$, it is rational. Each β_λ together with p induces a rational hypothesis:

$$\pi_\lambda = \{\pi(N,\theta_L) = 1/3,\ \pi(E,\theta_L) = 0,\ \pi(N,\theta_H) = \lambda 2/3,\ \pi(E,\theta_H) = (1-\lambda)2/3\}.$$

By updating π_λ conditional on N, Bayes' rule yields

$$\mu^*(\theta_L|N) = \frac{1}{1+2\lambda} \geq \frac{1}{2} \text{ for each } \lambda \in [0, 1/2],$$

showing that PBE-1 adheres to rational consistency.

Now, consider PBE-2. None of its beliefs off the path are rationality consistent. Although there are strategies that rationalize signaling E by a low-productivity worker, allowing us to construct rational hypotheses under which E is feasible, none of the hypotheses justify $\mu(\theta_L|E) \geq \frac{1}{2}$.

To illustrate, let β_γ be the employer's belief that the low-productivity worker chooses E with a positive probability $\gamma \geq 0$ while the high-productivity worker does it with certainty; i.e.,

$$\beta_\gamma = (\beta(E|\theta_L) = \gamma,\ \beta(N|\theta_L) = 1-\gamma,\ \text{and } \beta(E|\theta_H) = 1)\,.$$

For each $\gamma \in [0,1]$, the above belief defines a rational hypothesis π_γ given by

$$\pi_\gamma := \big\{\pi(N,\theta_L) = (1-\gamma)1/3,\ \pi(E,\theta_L) = \gamma 1/3,\ \pi(N,\theta_H) = 0,\ \pi(E,\theta_H) = 2/3\big\},$$

as β_γ best responds to $b_R = (b_R(e|E) = 1,\ b_R(m|N) = 1/2,\ b_R(e|N) = 1/2)$. However,

$$\mu(\theta_L|E) = \frac{\pi_\gamma(E,\theta_L)}{\pi_\gamma(E,\Theta)} = \frac{\gamma}{\gamma+2} \leq \frac{1}{3} \text{ for each } \gamma \in [0,1],^{11}$$

showing that PBE-2 violates rational consistency.

Our next goal is to introduce an equilibrium concept that adheres to rational consistency.

[11] This is true for all rational hypotheses according to which E is feasible.

3 Rational Hypothesis Testing Equilibrium

In this section, we introduce a solution concept, called Rational Hypothesis Testing Equilibrium.

PBE beliefs on the path satisfy rational consistency by definition. The Sender's equilibrium strategy combined with the prior information about types induces the rational hypothesis that justifies the PBE beliefs on the path (see Equation (4)). However, PBE beliefs off the path may violate rational consistency (see Examples 1 and 2). Yet, if both players are rational and each one believes that the opponent is rational, they best respond with respect to beliefs whose support contains rational strategies of the opponent. To preserve consistency with mutual belief in rationality, we require PBE beliefs to be derived from rational hypotheses via Ortoleva's Hypothesis Testing model.

Formally, a *Rational Hypothesis Testing Equilibrium* (RHTE) consists of a strategy profile (b_S^*, b_R^*), a second-order prior ρ, and beliefs $\mu_\rho^* = \left\{\mu_\rho^*(\cdot|m)\right\}_{m\in\mathcal{M}}$ that satisfy rational consistency.

Definition 5 (RHTE) $(b_S^*, b_R^*, \rho, \mu_\rho^*)$ *constitutes a Rational Hypothesis Testing Equilibrium if:*

$$(i)\ \ b_S^*(m|\theta) > 0 \textit{ implies } m \in \arg\max_{m'\in\mathcal{M}} \sum_{a\in\mathcal{A}} u_S(\theta, m', a) b_R^*(a|m') \textit{ for each } \theta \in \Theta,$$

$$(ii)\ \ b_R^*(a|m) > 0 \textit{ implies } a \in \arg\max_{a'\in\mathcal{A}} \sum_{\theta\in\Theta} u_R(\theta, m, a') \mu_\rho^*(\theta|m) \textit{ for each } m \in \mathcal{M},$$

$$(iii)\ \ \mu_\rho^*(\theta|m) = \frac{\pi^*(m,\theta)}{\pi^*(m,\Theta)} \textit{ if } \pi^*(m,\Theta) > 0, \textit{ where } \{\pi^*\} := \arg\max_{\pi\in supp(\rho)} \rho(\pi), \textit{ and}$$

$$\pi^*(m,\theta) = \begin{cases} \beta^*(m|\theta)p(\theta), & \textit{if } \beta^*(m|\theta) > 0 \textit{ where } \beta^* = b_S^*, \\ 0, & \textit{otherwise}, \end{cases}$$

$$(iv)\ \ \mu_\rho^*(\theta|m) = \frac{\pi_m^{**}(m,\theta)}{\pi_m^{**}(m,\Theta)} \textit{ if } \pi^*(m,\Theta) = 0, \textit{ where } \{\pi_m^{**}\} := \arg\max_{\pi\in supp(\rho)} \rho_m(\pi),$$

$$\pi^{**}(m,\theta) = \begin{cases} \beta(m|\theta)p(\theta), & \textit{if } \beta(m|\theta) > 0 \textit{ where } \beta \in \mathcal{B}_S^\bullet, \\ 0, & \textit{otherwise}, \end{cases}$$

$$(v)\ \ \mu_\rho^*(\cdot|m) \textit{ is an arbitrary probability distribution on } \Theta \textit{ if } m \textit{ is strictly dominated}.$$

Conditions (i) and (ii) ensure sequential rationality. Conditions (iii) and (iv) ensure that conditional beliefs satisfy rational consistency. By conditions $(i), (ii)$, and (iii), each RHTE is a PBE. To avoid abusing acronyms, we will sometimes refer to RHTE simply as *rational equilibrium*.

In some games, there might be a very costly message that none of the Sender's types will ever play. In this case, there is no rational strategy that generates the message. Since (strictly) dominated messages do not affect equilibrium behavior, condition (v) admits arbitrary beliefs for

any $m^d \in \mathcal{M}^d$, allowing us to apply RHTE to a broader family of signaling games. From now on, we let $\mathcal{M}^\circ$ denote the set of messages off the equilibrium path which are not strictly dominated.

In rational equilibrium, beliefs are consistent with mutual belief in rationality, provided $\mathcal{M}^d = \emptyset$. That is, the Receiver believes that messages are generated by (second-order) rational strategies.[12]

Example 2 (continued)

Consider again the game in Figure 2. We have shown that PBE-1 adheres to rational consistency. Therefore, the only rational equilibrium in this game is with pooling on education; e.g.,

$$b_S^*(E|\theta_L) = b_S^*(E|\theta_H) = 1,\ b_R^*(e|E) = b_R^*(m|N) = 1, \qquad \text{(RHTE-1)}$$
$$supp(\rho) = \{\pi_1, \pi_1'\} \text{ such that } \rho(\pi_1') < \rho(\pi_1),\ \ \mu_\rho^*(\theta_L|E) = 1/3 \ \text{ and } \ \mu_\rho^*(\theta_L|N) = 1,$$

where π_1 is the initial hypothesis and π_1' is the alternative hypothesis which are given by[13]

$$\pi_1 := \{\pi_1(E,\theta_L) = 1/3, \pi_1(E,\theta_H) = 2/3\} \text{ where } \beta_1 = (\beta_1(E|\theta_L) = 1,\ \beta_1(E|\theta_H) = 1))\,,$$
$$\pi_1' := \{\pi_1'(N,\theta_L) = 1/3, \pi_1'(E,\theta_H) = 2/3\} \text{ where } \beta_1' = (\beta_1'(N|\theta_L) = 1,\ \beta_1'(E|\theta_H) = 1))\,.$$

Ortoleva (2012) suggested an equilibrium notion, called *Hypothesis Testing Equilibrium* (HTE). It is important to remark on the conceptual differences between RHTE and Ortoleva's HTE.

Remark 3 Similar to us, Ortoleva assumed that the initial hypothesis is rejected for each unsent message (i.e., $\epsilon = 0$). Besides this, there are two substantial differences. First, Ortoleva's HTE builds on simple hypotheses (i.e., pure strategies). In contrast, we allow for behavioral strategies and non-simple hypotheses. Second, and more importantly, Ortoleva assumes that beliefs are derived from first-order rational strategies. That is, his hypothesis notion, defined by a (single) pure strategy that best responds to *some* (not necessarily rational) pure strategy of the Sender, is weaker. Thus, all our results carry over to Ortoleva's equilibrium. Since Ortoleva's HTE may violate rational consistency, we elaborate on the main consequences for the Spence game (see Section 4).

Remark 4 Sun (2019) extended Ortoleva's HTE by allowing for a strictly positive threshold, $\epsilon \geq 0$.

[12] One could require higher-order rationality to justify beliefs, resulting in a stronger notion of rational consistency. However, it is well-known that mutual belief (knowledge) in rationality and in strategies (conjectures) are sufficient conditions for Nash equilibrium (Aumann and Brandenburger, 1995) as well as sequential equilibrium (Asheim and Perea, 2005) in two-player games. Building on rational consistency, we aim to explore implications of this assumption for PBE beliefs. Moreover, we are unaware of an interesting example where an additional level of rationality plays a role. We should also remark that there exists a literature that connects different notions of extensive-form rationalizability with the iterative Intuitive Criterion by Cho and Kreps (1987), who assumed common knowledge of rationality. E.g., Sobel, Stole, and Zapater (1990) characterize the iterative Intuitive Criterion as a fixed-equilibrium rationalizable outcome in games with imperfect information (i.e., types are interpreted as players). Battigalli (2006) solves signaling games by applying the Δ-rationalizability introduced by Battigalli and Siniscalchi (2002). The latter authors characterize the Intuitive Criterion outcome in a self-confirming equilibrium of Fudenberg and Levine (1993).

[13] Note that $\pi_1' = \pi_{\lambda=0}$ (see Example 2). Since β_1' best responds to $b_R' = (b_R(e|E) = b_R(e|N) = 1)$, it is rational.

That is, the Receiver may reject the initial hypothesis on the path if she is surprised by a message that is (ex-ante) strictly less likely than ϵ. Sun provided two examples, one showing that HTE ($\epsilon = 0$) can refine PBE. The second one illustrates that HTE with $\epsilon > 0$ can coarsen PBE. Since HTE builds on first-order rationality, RHTE is a refinement of Ortoleva's HTE with $\epsilon = 0$. Thus, the solution concepts are nested in the following way: HTE ($\epsilon > 0$) $\supseteq$ HTE ($\epsilon = 0$) $\supseteq$ RHTE.

4 Rational Equilibrium in the Spence Game

Now, we are ready to solve the Spence game described in Section 2.4. First, we derive the set of education levels that can constitute a pooling rational equilibrium. Later, we show that the efficient separating equilibrium, known as the Riley outcome (Riley, 1979), is a rational equilibrium too. Besides this, we highlight the key difference between rational equilibrium and Ortoleva's HTE.

4.1 Pooling RHTE

In Example 1, we derived a necessary condition for a Pooling PBE to adhere to rational consistency. The employer must believe that any education level $e' \in \{e_{n+1}, \dots, e_N\}$ perfectly correlates with high productivity. In other words, rational consistency requires the highest wage to be paid for each such e' since $\mu(\theta_H|e') = 1$ is the only belief that can be justified by a rational hypothesis. This condition restricts the levels of education that can be supported by a Pooling RHTE.

Note the payoff for type θ_H from choosing a (pooling) message e^* at $w(e^*) = \mathbb{E}(\theta)$ must be greater than his payoff from choosing e_{n+1} at the highest wage θ_H. That is, e^* must satisfy

$$\mathbb{E}(\theta) - \frac{e^*}{\theta_H} \geq \theta_H - \frac{e_{n+1}}{\theta_H}, \quad \text{or equivalently,} \quad e^* \leq \underbrace{(\theta_H - \theta_L)(\theta_L - (1-\alpha)\theta_H)}_{:=y}, \tag{18}$$

specifying an upper bound y for education that can be supported by a rational equilibrium. The "number" of rational equilibria is a function of α, the fraction of high-productivity workers. When a relatively small fraction of high-productivity workers occupies the market; i.e., $\alpha < \frac{\theta_H - \theta_L}{\theta_H}$, none of the Pooling PBEs, which always exist, adheres to rational consistency (see Figure 3). On the other hand, the larger α is, the more education levels can be supported by a Pooling RHTE. Overall, rational equilibrium refines Pooling PBE since $y < x$. The next result summarizes these observations.

Proposition 1 *A Pooling RHTE exists if and only if* $\alpha \geq \frac{\theta_H - \theta_L}{\theta_H}$*; equivalently* $y = e_n - (1-\alpha)e_N \geq 0$*. Then, each education level* e^* *such that* $0 \leq e^* \leq y$ *is supported by a Pooling RHTE where*

$b_S^* = (b_S^*(e^*|\theta_L) = b_S^*(e^*|\theta_H) = 1))$ *and* $w_R^* = (w^*(e))_{e\in\mathcal{M}}$ *with* $\mu^* = (\mu^*(\cdot|e))_{e\in\mathcal{M}}$ *are given by*[14]

$$w^*(e) = \begin{cases} \theta_L & \text{if } e < e^*, \\ \mathbb{E}(\theta) & \text{if } e^* \leq e \leq e_n, \\ \theta_H & \text{if } e_n < e \leq e_N, \end{cases} \quad \text{and} \quad \mu^*(\theta_H|e) = \begin{cases} 0 & \text{if } e < e^*, \\ \alpha & \text{if } e^* \leq e \leq e_n, \\ 1 & \text{if } e_n < e \leq e_N. \end{cases} \tag{19}$$

Rational equilibrium refines Pooling PBE in two dimensions, the level of education and wages off the path. Interestingly enough, there is experimental evidence in favor of such behavior. Especially, Kübler, Müller, and Normann (2008) found pooling behavior at lower education levels.

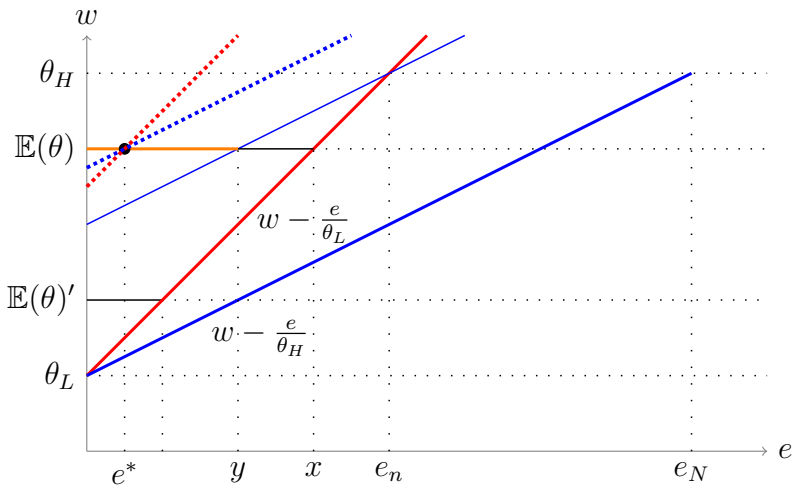

Figure 3: Pooling RHTE at $\mathbb{E}(\theta)$, but no Pooling RHTE at $\mathbb{E}(\theta)'$.

At this stage, let us briefly remark on the role of rationality for refining pooling behavior. Suppose for a moment that hypotheses build on first-order rationality as in Ortoleva (2012) (see Remark 3). Such hypotheses are about worker's strategies that best respond to some (not-necessarily rational) strategy of the employer, including wages paying more than θ_H; e.g., for each $e \in \mathcal{M}$, let

$$w(e) = \begin{cases} \theta_L + \frac{e}{\theta_L} + \varepsilon & \text{if } e \in \{e_{n+1}, \ldots, e_N\} \\ \theta_L & \text{elsewhere,} \end{cases} \tag{20}$$

where $\varepsilon > 0$.[15] Given $w(e)$, the only best response for both types is to choose e'. Hence, we can construct a hypothesis that justifies $\mu^*(\theta_H|e) = \alpha$ for each $e' \in \{e_{n+1}, \ldots, e_N\}$, allowing us to explain all Pooling PBEs, including the one with the average productivity, for each e' (see (19)).

[14] Recall that $e_n := \theta_L(\theta_H - \theta_L)$ and $e_N := \theta_H(\theta_H - \theta_L)$.

[15] It is clear that $w(e)$ is not rational. For example, when $e' > e_n$, $w(e') > \theta_H$.

Pooling equilibria that fail rational consistency are not "robust" under mutual belief in rationality. To be more specific, consider a PBE with pooling on e^* such that $y < e^* \leq x$ in which the average productivity, $\mathbb{E}(\theta)$, is paid for each $e' \in \{e_{n+1}, \ldots, e_N\}$ (see Equation (17)). Note that the payoff of the high-productivity worker from e_{n+1} at the highest wage, θ_H, is higher than his equilibrium payoff. Knowing that players are rational, the high-productivity worker must believe that the low-productivity worker has no incentive to choose e_{n+1}, and thus that the rational employer can only offer the highest wage for this education level. This reasoning provides the high-productivity worker an incentive to choose e_{n+1} instead of e^*. Rational equilibrium eliminates such incentives for defecting due to its consistency with mutual belief in rationality, making the equilibrium behavior "robust" .

4.2 Separating RHTE

In a separating PBE, both worker types choose different education levels revealing their types. When the employer observes $e(\theta_i)$, she infers that the worker-type is θ_i (i.e., $\mu(\theta_H \mid e(\theta_L)) = 0$ and $\mu(\theta_H \mid e(\theta_H)) = 1$), and offers the corresponding wages: $w(e(\theta_L)) = \theta_L$ and $w(e(\theta_H)) = \theta_H$.

At the minimum wage, the low-productivity type chooses $e(\theta_L) = 0$. Otherwise, he will get a strictly lower payoff. For the high-productivity type, the chosen education level $e(\theta_H)$ must satisfy

$$u_S(\theta_L, e(\theta_H), \theta_H) = \theta_H - \frac{e(\theta_H)}{\theta_L} \leq \theta_L - \frac{0}{\theta_L} = \theta_L = u_S(\theta_L, e(\theta_L), \theta_L), \tag{21}$$

$$u_S(\theta_H, e(\theta_H), \theta_H) = \theta_H - \frac{e(\theta_H)}{\theta_H} \geq \theta_L - \frac{0}{\theta_H} = \theta_L = u_S(\theta_H, e(\theta_L), \theta_L). \tag{22}$$

The first restriction ensures that θ_L prefers $e(\theta_L)$ for $w = \theta_L$ to $e(\theta_H)$ for $w = \theta_H$, while the second one ensures θ_H prefers $e(\theta_H)$ for $w = \theta_H$ to $e(\theta_L)$ for $w = \theta_L$. Thus, any education level $e(\theta_H)$ such that $e_n = \theta_L(\theta_H - \theta_L) \leq e(\theta_H) \leq \theta_H(\theta_H - \theta_L) = e_N$ can be supported by an equilibrium with respect to arbitrary wage schemes offered off the equilibrium path (see Figure 4). For instance, for each $e^*(\theta_H) \in \{e_n, \ldots, e_N\}$, the separating strategy $b_S^*(0|\theta_L) = 1$ and $b_S^*(e^*(\theta_H)|\theta_H) = 1$ together with a family of wages $w_R^* := (w^*(e))_{e \in \mathcal{M}}$ and beliefs $\mu^* := (\mu^*(\cdot|e))_{e \in \mathcal{M}}$ given by

$$w^*(e) = \begin{cases} \theta_L & \text{if } 0 \leq e < e^*(\theta_H), \\ \theta_H & \text{if } e = e^*(\theta_H), \\ \mathbb{E}(\theta|e) & \text{if } e^*(\theta_H) < e \leq e_N, \end{cases} \quad \text{and} \quad \mu^*(\theta_H|e) = \begin{cases} 0 & \text{if } 0 \leq e < e^*(\theta_H), \\ 1 & \text{if } e = e^*(\theta_H), \\ [0,1] & \text{if } e^*(\theta_H) < e \leq e_N, \end{cases}$$

where $\mathbb{E}(\theta|e) = \mu^*(\theta_H|e)\theta_H + (1 - \mu^*(\theta_H|e))\theta_L$, constitute a Separating PBE.

As previously argued, no rational hypothesis according to which θ_L chooses $e' \in \{e_{n+1}, \ldots, e_N\}$

with a strictly positive probability exists. Only the high-productivity type can choose such an e' as a best response and thus $\mu(\theta_H|e') = 1$ is the only belief that adheres to rational consistency. As a consequence, rational equilibrium eliminates most Separating PBEs except the Riley outcome; i.e., the most efficient equilibrium that Pareto dominates other separating equilibria (Riley, 1979).

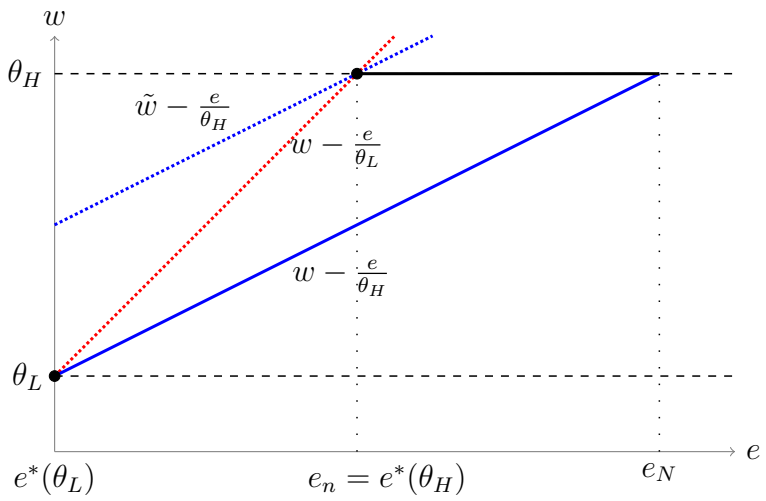

Figure 4: RHTE supporting the Riley outcome ($e^*(\theta_L) = 0$ and $e^*(\theta_H) = e_n$).

Proposition 2 *The Riley outcome is supported by a Separating RHTE with* $b_S^* = (b_S^*(0|\theta_L) = 1,\ b_S^*(e_n|\theta_H) = 1)$ *where wages* $w_R^* = (w(e))_{e\in\mathcal{M}}$ *and beliefs* $\mu^* = (\mu^*(\theta_H|e))_{e\in\mathcal{M}}$ *are given by*

$$w^*(e) = \begin{cases} \theta_L & \text{if } e_0 \le e < e_n, \\ \\ \theta_H & \text{if } e_n \le e \le e_N, \end{cases} \quad \text{where } \ \mu^*(\theta_H|e) = \begin{cases} 0 & \text{if } e_0 \le e < e_n, \\ \\ 1 & \text{if } e_n \le e \le e_N. \end{cases} \tag{23}$$

Rational equilibrium refines separating equilibria with respect to education levels and wages. As the above proposition shows, the (unique) Riley outcome is supported by the rational equilibrium with the unique wage scheme offered for any education level above the equilibrium level, e_n^*.[16]

We close this section with a final remark on the difference between Ortoleva's HTE and RHTE.

Remark 6 Ortoleva's HTE (with $\epsilon = 0$) supports *all* PBEs in the Spence game, including pooling and separating behavior. The reasons are twofold: First, first-order rationality allows the employer to form conjectures about workers' strategies that best respond to wage schemes that do not need to be rational, generating multiple hypotheses (see Section 4.1). Second, the single-crossing property, which is necessary to distinguish between types based on education they are willing to obtain,

[16]The Riley equilibrium is the only Intuitive PBE. Moreover, Eső and Schummer (2009) showed that the Riley equilibrium is not vulnerable to deviations; i.e., a unique set of types with an incentive to deviate does not exist.

allows us to construct hypotheses that justify all conceivable beliefs, $\mu(\theta_L|e') \in [0,1]$, for each information set $e' \in \{e_{n+1}, \ldots, e_N\}$. As a result, all PBEs are supported by Ortoleva's HTE.

5 Rational Equilibrium and Other Refinements

This section relates rational equilibrium with two refinement concepts, *intuitive equilibrium* (Cho and Kreps, 1987) and *undefeated equilibrium* (Mailath, Okuno-Fujiwara, and Postlewaite, 1993).

To cope with the multiplicity of equilibria, various refinement concepts have been suggested in the economic literature (e.g., Kohlberg and Mertens, 1986; Cho and Kreps, 1987; Banks and Sobel, 1987; Cho, 1987; Mailath, Okuno-Fujiwara, and Postlewaite, 1993; Farrell, 1993; Fudenberg and He, 2020). The existing refinements divide into two main categories. The first category is designed as a test for *strategic stability*, introduced by Kohlberg and Mertens (1986), including the *Intuitive Criterion* by Cho and Kreps (1987) and the $D1$ *Criterion* by Banks and Sobel (1987).[17] The second category of refinements contains criteria unrelated with strategic stability or the Intuitive Criterion such as the notion of undefeated equilibrium by Mailath, Okuno-Fujiwara, and Postlewaite (1993).

These theories pose different criteria for stating when a message off the equilibrium path can or cannot be "reasonably" expected to be sent by a Sender type. For the Intuitive Criterion, beliefs are "reasonable" if they assign positive probabilities to types that could benefit from deviating to an unsent message. For undefeated equilibrium, beliefs are "reasonable" if they assign positive probabilities to types that may use an unsent message as an attempt to achieve another equilibrium. Both concepts apply to games where at least one type could benefit from deviation, independent of whether it is rational or not. Our concept requires deviations to be governed by rational strategies.

Our comparison begins with the Intuitive Criterion which is broadly applied in economics.

5.1 Intuitive PBE

The Intuitive Criterion eliminates beliefs off the path that assign a positive probability to the Sender types that do not have an incentive to deviate from equilibrium. Let us briefly recall this criterion. Consider a PBE, (b_S^*, b_R^*, μ^*) and denote by $u_S^*(\theta)$ the expected equilibrium payoff for type θ; i.e.,

$$u_S^*(\theta) := \sum_{m \in \mathcal{M}} \left(\sum_{a \in \mathcal{A}} u_S(\theta, m, a) b_R^*(a|m) \right) b_S^*(m|\theta). \tag{24}$$

For an unsent message m°, let $T(m^\circ) \subseteq \Theta$ be the set of types that cannot improve upon the equilibrium payoff by deviating to m°, no matter how the Receiver responds to m°; i.e., for each

[17] Appendix A relates rational equilibrium to *strategic stability* and the $D1$ *Criterion*.

$\theta \in T(m^\circ)$ and $a \in BR(\Theta, m^\circ)$[18], $u_S^*(\theta) > u_S(\theta, m^\circ, a)$. Thus, $I(m^\circ) := \Theta \setminus T(m^\circ)$ is the set of types that could be (weakly) better off by choosing m° than following the equilibrium strategy b_S^*,

$$I(m^\circ) := \{\theta \in \Theta : u_S^*(\theta) \leq u_S(\theta, m^\circ, a) \text{ for some } a \in BR(\Theta, m^\circ)\}. \tag{25}$$

A PBE fails the Intuitive Criterion if a type that would be better off by choosing m° exists, provided that the Receiver best responds to m° with respect to beliefs that concentrate on $I(m^\circ)$; otherwise, the equilibrium survives the Intuitive Criterion.

Definition 6 (Intuitive Criterion) *A PBE, (b_S^*, b_R^*, μ^*), fails the Intuitive Criterion if for some out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$, there is a type $\theta \in I(m^\circ)$ such that for all $a \in BR(I(m^\circ), m^\circ)$,*

$$u_S^*(\theta) < u_S(\theta, m^\circ, a) \tag{26}$$

where $BR(I(m^\circ), m^\circ)$ is the set of best responses with respect to beliefs in $\Delta(I(m^\circ))$.

If a PBE passes the Intuitive Criterion, there is a best reply for the Receiver with respect to some belief in $\Delta(I(m^\circ))$, which makes all types better off by following the equilibrium strategy (i.e., there is $a \in BR(I(m^\circ), m^\circ)$ such that $u_S^*(\theta) \geq u_S(\theta, m^\circ, a)$ for each θ). An equilibrium that passes (resp., fails) the Intuitive Criterion is called *Intuitive* (resp., *Unintuitive) PBE*.

In the game of Figure 2, rational equilibrium and Intuitive PBE predict the same behavior.

Example 3 First, consider PBE-1. Given the equilibrium payoff, type θ_L could be better off by playing N if he expects to obtain the executive job. That is, $I(N) = \{\theta_L\}$ and $T(N) = \{\theta_H\}$. As long as the employer believes that N is signaled by θ_L, no type has an incentive to deviate to N, showing that PBE-1 with $\mu(\theta_L|N) = 1$, which is RHTE-1, passes the Intuitive Criterion.

Now, consider PBE-2, which was shown to violate rational consistency in Example 2. Only the high-productivity type could benefit from detecting to E. However, the employer believes that E is chosen by θ_H (i.e., $\mu(\theta_H|E) = 1$) responds with e. Consequently, the high-productivity type will indeed signal E instead of N, showing that this equilibrium fails the Intuitive Criterion.

In general, however, rational equilibrium and intuitive equilibrium are unrelated concepts.

Proposition 3 *RHTE and Intuitive PBE are not nested.*

There is a rational equilibrium that fails the Intuitive Criterion. Likewise, there is an intuitive equilibrium that violates rational consistency. The former case is illustrated below.[19]

[18] Recall, $BR(\Theta, m^\circ)$ is the set of (pure) best-responses with respect to beliefs in $\Delta(\Theta)$; see Equations (1) and (2).
[19] For the latter, see the proof of Proposition 3 in Appendix C.

Example 4 Game 2 in Figure 5 has two rational equilibria. In the first one, both types pool on E;

$$b_S^*(E|\theta_L) = b_S^*(E|\theta_H) = 1, \;\; b_R^*(e|E) = b_R^*(m|N) = 1, \qquad \text{(RHTE-2)}$$

$supp(\rho) = \{\pi_3, \pi_3'\}$ such that $\rho(\pi_3') < \rho(\pi_3)$, $\mu_\rho^*(\theta_L|E) = 1/3$ and $\mu_\rho^*(\theta_L|N) = 1$, where

$\pi_3 := \{\pi_3(E,\theta_L) = 1/3, \pi_3(E,\theta_H) = 2/3\}$ where $\beta_3 = (\beta_3(E|\theta_L) = 1, \; \beta_3(E|\theta_H) = 1)$,

$\pi_3' := \{\pi_3'(N,\theta_L) = 1/3, \pi_3'(E,\theta_H) = 2/3\}$ where $\beta_3' = (\beta_3'(N|\theta_L) = 1, \; \beta_3'(E|\theta_H) = 1)$.[20]

In the second rational equilibrium, both types pool on N; i.e.,

$$b_S^*(N|\theta_L) = b_S^*(N|\theta_H) = 1, \;\; b_R^*(m|E) = b_R^*(e|N) = 1, \qquad \text{(RHTE-3)}$$

$supp(\rho) = \{\pi_4, \pi_4'\}$ such that $\rho(\pi_4') < \rho(\pi_4)$, $\mu_\rho^*(\theta_L|E) = 1$ and $(\theta_L|N) = 1/3$, where

$\pi_4 := \{\pi_4(N,\theta_L) = 1/3, \; \pi_4(N,\theta_H) = 2/3\}$ given $\beta_4 := (\beta_4(N|\theta_L) = 1, \; \beta_4(N|\theta_H) = 1)$,

$\pi_4' := \{\pi_4'(E,\theta_L) = 1/3, \; \pi_4'(N,\theta_H) = 2/3\}$ given $\beta_4' := (\beta_4'(E|\theta_L) = 1, \; \beta_4'(N|\theta_H) = 1)$.[21]

In this game, RHTE-2 passes the Intuitive Criterion, but not RHTE-3.[22]

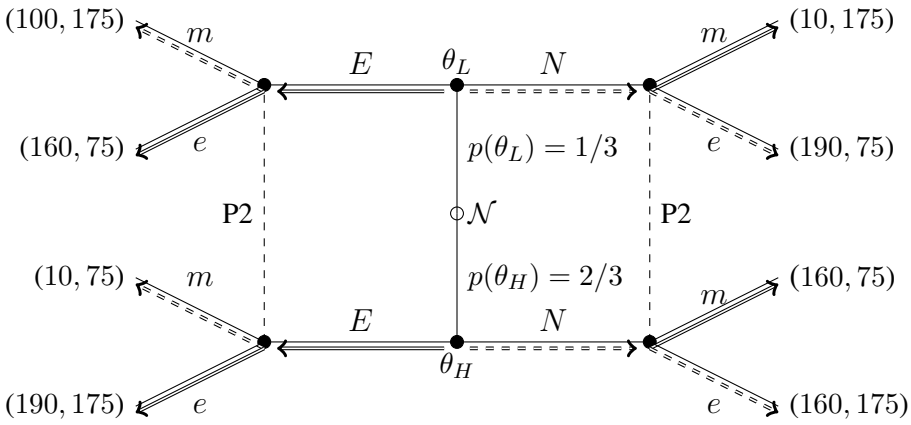

Figure 5: Game 2 in Brandts and Holt (1992) with RHTE-2 ("⟶") and RHTE-3 ("⇢")

Next, we derive a condition under which intuitive equilibrium satisfies rational consistency. We focus on equilibria for which at least one rational hypothesis that makes an unsent message m° feasible exists. As the next lemma shows, a rational hypothesis exists, presupposed that some type could benefit by deviating to m° no matter whether the given PBE is intuitive or not.[23]

[20] β_3 and β_3' best respond to $b_R(e|E) = b_R(m|N) = 1$ and $b_R'(e|E) = b_R'(e|N) = 1$, respectively.

[21] β_4 and β_4' best respond to $b_R(m|E) = b_R(e|N) = 1$ and $b_R'(m|E) = b_R'(m|N) = 1$, respectively.

[22] In RHTE-2, $I(N) = \{\theta_L\}$. So the Intuitive Criterion selects $\mu(\theta_L|N) = 1$ under which both worker types have no incentive to deviate. In RHTE-3, $I(E) = \{\theta_H\}$. So the Intuitive Criterion selects $\mu(\theta_H|E) = 1$. Under this belief, the Receiver responds with e to E which in turn causes that the worker of type θ_H will indeed deviate to E.

[23] Note that the Intuitive Criterion has no bite if $I(m^\circ) = \emptyset$. In this case, all beliefs in $\Delta(I(m^\circ))$ are admitted.

Lemma 2 *If $I(m^\circ) \neq \emptyset$, then a rational hypothesis according to which m° is feasible exists.*

The non-emptiness of $I(m^\circ)$ is, however, not sufficient to guarantee rational consistency (see Example 2) unless the intuitive equilibrium is strategically stable (Appendix A, Theorem 2). Note that the Intuitive Criterion admits arbitrary beliefs over the set of potentially deviating types (i.e., $\Delta(I(m^\circ))$), provided $I(m^\circ)$ is a non-singleton set. To identify the "intuitive" beliefs that adhere to rational consistency, an additional condition is needed. For $m^\circ \in \mathcal{M}^\circ$ and $a \in BR(\Theta, m^\circ)$,

$$I(m^\circ; a) \equiv \{\theta \mid u_S^*(\theta) \leq u_S(\theta, m^\circ, a)\} \subseteq I(m^\circ) \tag{27}$$

denotes the set of types that have an incentive to deviate to m° if a is the Receiver's response to m°. If the equilibrium is intuitive, at least one a for which $I(m^\circ; a) \neq \emptyset$ exists. The Receiver may reason that all the types in $I(m^\circ; a)$ pool on m° in response to a, and consequently, she will derive her posterior by updating the prior distribution p conditional on $I(m^\circ; a)$. If the PBE belief at m° is the Bayesian update of p given $I(m^\circ; a)$, then the Intuitive PBE constitutes a rational equilibrium.

Theorem 1 *Let (b_S^*, b_R^*, μ^*) be an Intuitive PBE. Suppose that for each unsent message $m^\circ \in \mathcal{M}^\circ$, there is a best response $a \in BR(\Theta, m^\circ)$ for which $I(m^\circ; a) \neq \emptyset$, and the belief at m° is given by*

$$\mu^*(\theta | m^\circ) = \begin{cases} \frac{p(\theta)}{\sum_{\theta' \in I(m^\circ; a)} p(\theta')} & \text{if } \theta \in I(m^\circ; a), \\ 0 & \text{if } \theta \notin I(m^\circ; a). \end{cases} \tag{28}$$

Then, an RHTE that supports the Intuitive PBE exists.[24]

Theorem 1 contains a special case that deserves a remark. When $I(m^\circ) = \{\theta_{m^\circ}\}$ is a singleton, the Receiver learns the only type that could benefit from playing m°. In this case, a rational equilibrium that supports the unique Intuitive-Criterion outcome (i.e., $\mu(\theta_{m^\circ} | m^\circ) = 1$) exists.

Corollary 1 *If (b_S^*, b_R^*, μ^*) is an Intuitive PBE with a singleton $I(m^\circ)$ for each unsent message m°, then an RHTE that supports the PBE with $\mu^*(\theta_{m^\circ} | m^\circ) = 1$ where $\{\theta_{m^\circ}\} = I(m^\circ)$ exists.*

5.2 Undefeated PBE

In the previous section, we have shown that rational equilibrium and intuitive equilibrium do not need to be nested. In this section, we compare rational equilibrium with another popular concept, the undefeated equilibrium suggested by Mailath, Okuno-Fujiwara, and Postlewaite (1993).

[24] As Cho (1987) pointed out, "intuitive" beliefs do not need to assign zero probabilities to the types in $T(m^\circ)$, violating the so-called *introspective consistency*. For this reason, the author suggested a stronger equilibrium concept, called forward induction equilibrium. It requires PBE beliefs to satisfy introspective consistency, i.e., $\mu(\theta | m^\circ) = 0$ for each $\theta \in T(m^\circ)$. The rational equilibrium derived in Theorem 1 is in fact a forward induction equilibrium.

To recapitulate this refinement concept, we follow Mailath, Okuno-Fujiwara, and Postlewaite (1993) and consider equilibria in pure strategies. The key idea is to interpret each message off the path as an attempt by some types of the Sender to achieve another equilibrium, which the types strictly prefer to the given one. Hence, the refinement criterion compares two equilibria: a given PBE, (b_S^*, b_R^*, μ^*), and an alternative PBE', $(b_S^{**}, b_R^{**}, \mu^{**})$. The main step in the refinement criterion is to verify if the alternative equilibrium defeats the given one. Suppose that for some unsent message m° in the given PBE, there is a non-empty set of types signaling m° in the alternative PBE' which all the types, and at least one of them does even strictly, prefer to the given PBE; i.e.,

$$K(m^\circ) := \{\theta \in \Theta : b_S^{**}(m^\circ|\theta) = 1 \text{ and } u_S^{**}(\theta) \geq u_S^*(\theta)\} \neq \emptyset, \tag{29}$$

and $u_S^{**}(\theta) > u_S^*(\theta)$ for at least one $\theta \in K(m^\circ)$. If the belief at m° of the given PBE is consistent with the Bayesian update of the prior distribution p conditional on the types in $K(m^\circ)$ which strictly prefer PBE', then the given PBE is *undefeated* by the alternative PBE'. Otherwise, it is *defeated*.[25]

Definition 7 (Defeated Equilibrium) *A given PBE, (b_S^*, b_R^*, μ^*), is defeated by an alternative PBE', $(b_S^{**}, b_R^{**}, \mu^{**})$, if there exists an unsent message $m^\circ \in \mathcal{M}^\circ$ (i.e., $b_S^*(m^\circ|\theta) = 0$ for all $\theta \in \Theta$) and a non-empty set $K(m^\circ) := \{\theta \in \Theta : b_S^{**}(m^\circ|\theta) = 1$ and $u_S^{**}(\theta) \geq u_S^*(\theta)\}$, containing at least one θ that strictly prefers the alternative equilibrium, i.e., $u_S^{**}(\theta) > u_S^*(\theta)$, and for some $\theta \in K(m^\circ)$,*

$$\mu^*(\theta|m^\circ) \neq \frac{p(\theta)\xi(\theta)}{\sum\limits_{\theta' \in K(m^\circ)} p(\theta')\xi(\theta')}, \tag{30}$$

for any mapping $\xi : \Theta \to [0,1]$ given by

$$\xi(\theta) = \begin{cases} 1 & \text{if } \theta \in K(m^\circ) \text{ and } u_S^{**}(\theta) > u_S^*(\theta); \\ 0 & \text{if } \theta \notin K(m^\circ), \end{cases} \tag{31}$$

allowing that any type in $K(m^\circ)$, which is indifferent (i.e., $u_S^{**}(\theta) = u_S^*(\theta)$), may have randomized.

An equilibrium is *undefeated* if no other equilibrium that defeats it exists. The refinement concept selects an equilibrium that is undefeated. As in the case of intuitive equilibrium, rational equilibrium and undefeated equilibrium are not nested concepts.[26] However, we can use the conditions testing for "undefeated equilibrium" as a condition that identifies a rational equilibrium, relating the two refinement concepts in the following proposition.

[25] If the belief concentrates on the set of such types, then the given equilibrium defeats the other equilibrium.

[26] Example 5 proves one direction. An example proving the reverse direction can be provided upon request.

Proposition 4 *Let (b_S^*, b_R^*, μ^*) be a PBE. If for each unsent message m° there is an alternative PBE $(b_S^{**}, b_R^{**}, \mu^{**})$ that does not defeat the given PBE and $u^*(\theta) < u^{**}(\theta)$ for all $\theta \in K(m^\circ) \neq \emptyset$, then the given PBE constitutes an RHTE.*

In the next example below, we illustrate these results to derive an RHTE in the game in Figure 5.

Example 5 Recall, Game 2 has two equilibria (Example 4). Let RHTE-3 be the given equilibrium and RHTE-2 be the alternative one. In this case, the high-productivity worker prefers the equilibrium with pooling on E; i.e., $K(E) = \{\theta_H\}$. Yet, the employer believes that education is signaled by the low-productivity type (i.e., $\mu_\rho^*(\theta_L|E) = 1$), showing that RHTE-2 defeats RHTE-3.

When RHTE-2 is the given equilibrium, the low-productivity worker strictly prefers the alternative one with pooling on N; i.e., $K(N) = \{\theta_L\}$. Thus, he may signal N to achieve RHTE-3. In RHTE-2, however, the employer's belief at N is consistent with this reasoning (i.e., $\mu_\rho^*(\theta_L|N) = 1$), showing that RHTE-2 is undefeated by RHTE-3. It is the only undefeated equilibrium.[27]

The above argument proves that the undefeated equilibrium adheres to rational consistency. The fact that the low-productivity worker best responds with N, and the high-productivity worker with E whenever the employer matches both signals with the executive job, induces a rational hypothesis, π_4', that justifies the Bayesian update of the prior distribution on $K(N)$, showing that the undefeated equilibrium is a rational equilibrium.

6 Experimental Findings and "Type-Dependence"

In this section, we show that rational equilibrium can account for the main experimental findings in Brandts and Holt (1992, 1993). By testing intuitive versus unintuitive equilibrium, both studies found an intriguing pattern of behavior. Intuitive equilibrium predominates in some games, while unintuitive equilibrium prevails in other games. The authors attribute this phenomenon to "type-dependence" which is a salient separating behavior in a given game. We show that rational equilibrium incorporates "type-dependence" as the most likely rational hypothesis off the path, offering an alternative explanation.

From now on, we refer to the studies of Brandts and Holt (1992, 1993) as BH-92 and BH-93. Table 1 summarizes the predictions for the two games implemented by BH-92 (Figures 2 and 5).[28] Either of the games has two pooling equilibria of which one is intuitive and one is unintuitive. The intuitive equilibrium is undefeated, and the unintuitive equilibrium is defeated (see Example 5).[29]

[27] For the same reasons, RHTE-1 is the undefeated equilibrium in Game 2.

[28] Game 2 corresponds to Game 3R in Brandts and Holt (1992).

[29] Remarks on undefeated equilibrium are our responsibility, as BH-92 and BH-93 did not study this refinement.

In Game 1, the only rational equilibrium is the intuitive equilibrium (Example 2). This is different in Game 2 where both the intuitive equilibrium and the unintuitive equilibrium are rational equilibria (Example 4). Hence, the three refinement concepts predict the PBE with pooling on E in both games. Besides that, rational equilibrium predicts the PBE with pooling on N in Game 2.

Signaling Game	Equilibrium Concept			
Game 1 & Game 2	PBE with pooling on E			PBE with pooling on N
Game 1	Intuitive	Undefeated	RHTE-1	
Game 2	Intuitive	Undefeated	RHTE-2	RHTE-3

Table 1: Equilibrium predictions

In Game 1, BH-92 found that a vast majority of subjects' behaviors (79.7%) matched with the intuitive equilibrium. In Game 2, however, a larger fraction of subjects (58.3%) behaved consistently with the unintuitive equilibrium as opposed to the intuitive equilibrium (16.0%). This finding has challenged both intuitive equilibrium and undefeated equilibrium as a positive theory.

BH-93 found the same pattern of behavior in a series of games, each of which has one intuitive and one unintuitive equilibrium.[30] In particular, Game 3 and Game 5 in BH-93 have exactly the same equilibrium structure as Game 1 and Game 2 (Table 1). In Game 3, more than half of the subjects (57.0%) behaved consistently with the intuitive equilibrium which is also undefeated. In Game 5, slightly less than half of the subjects' behaviors matched with the unintuitive equilibrium which is defeated. The preponderance of the unintuitive equilibrium in Game 4, which is undefeated in this game, was even more pronounced (61.5%) as compared to the intuitive (and defeated) equilibrium. Table 2 summarizes the frequencies reported by BH-92 and BH-93, showing that rational equilibrium can account for the authors' main experimental findings.[31]

	BH-92		BH-93		
Equilibrium	Game 1	Game 2	Game 3	Game 4	Game 5
Intuitive PBE	79.7% (RE)	16.0% (RE)	57.0% (RE)	8.0% (RE)	10.5% (RE)
Unintuitive PBE	5.5%	58.3% (RE)	12.5%	61.5% (RE)	48.0% (RE)

Table 2: Equilibrium frequencies (RE:= RHTE)

[30] Game 3 in BH-93, which is similar to Game 1, has the same equilibrium predictions presented in Table 1. Also Games 4 and 5 in BH-93, which resemble Game 2, have the same predictions as in Table 1 except that for Game 4, the Intuitive PBE is defeated, and the Unintuitive PBE is undefeated. The respective proofs can be provided upon request.

[31] There are other studies testing the Intuitive-Criterion predictions: Banks, Camerer, and Porter (1994) found evidence in favor of intuitive equilibrium. On the contrary, Kübler, Müller, and Normann (2008) found support for both separating and pooling behavior in the Spence game, where the latter fails the Intuitive Criterion (see Section 4).

BH-92 and BH-93 attribute this phenomenon to adjustment processes governed by a property called *type-dependence*. It refers to a salient separating behavior in a given game. Adjustments of behavior triggered by a type-dependence "select" an equilibrium that is intuitive or unintuitive.

As BH-92 argue, Game 1 features *normal* type-dependence: "*[...] only high-ability workers obtain an education*; *knowing this, the employer assigns a worker with no education to the manual job. In order to obtain the preferred job, the low-ability workers would want to invest in the education signal*" (Brandts and Holt, 1992, p.1359), leading to the intuitive (education) equilibrium.

Game 2 features *reverse* type-dependence: "*[...] only a low-ability worker obtains an education and is, therefore, assigned to the manual job. This gives a low-ability worker the incentive to signal differently in order to be pooled with the uneducated, high-ability workers in executive jobs. In this way, unintuitive (no-education) equilibrium might be reached*" (Brandts and Holt, 1992, p.1359).

To be more precise, BH-92 identify type-dependence with a salient or "reasonable" separating behavior. Let (m_L, m_H) denote a type-contingent choice, associated with a pure strategy $b_S = (b_s(m_L|\theta_L) = 1, b_S(m_H|\theta_H) = 1)$, where m_L reads "a low-productivity worker signals m_L."[32] For a tuple of probabilities $(b_R(\cdot|E), b_R(\cdot|N))$, a type-contingent choice (m_L, m_H) is "optimal" if

$$\sum_{a\in\{e,m\}} u(\theta_L, m_L, a) b_R(a|m_L) \geq \sum_{a\in\{e,m\}} u(\theta_L, m, a) b_R(a|m) \text{ for } m \neq m_L. \tag{32}$$

$$\sum_{a\in\{e,m\}} u(\theta_H, m_H, a) b_R(a|m_H) \geq \sum_{a\in\{e,m\}} u(\theta_H, m, a) b_R(a|m) \text{ for } m \neq m_H, \tag{33}$$

A pair (m_L, m_H) with $m_L \neq m_H$ (i.e., separating behavior) defines type-dependence if the set of probabilities that rationalize (m_L, m_E) is "larger" than for other type-contingent choices.[33]

Figure 6 depicts the sets of probabilities that rationalize type-contingent choices in Game 1 and Game 2, respectively. In the left diagram, the shaded area represents the set of probabilities that rationalize (N_L, E_H). Notably this set is "larger" than the set of probabilities that rationalize (E_L, N_H) which is empty. This is why Game 1 features normal type-dependence, (N_L, E_H). In Game 2, on the other hand, reverse type-dependence (N_L, E_H) is the salient separating behavior, as the right diagram below proves.

In the approaches by BH-92 and BH-93, an equilibrium emerges gradually through adjustments of behavior that best respond to type-dependence. In the end, a single equilibrium is determined. Since type-dependence describes "optimal" behavior, it is not surprising that the final equilibrium can be explained by a rational equilibrium whose alternative hypotheses capture type-dependence.

[32] We use type-contingent choices to disentangle this approach from other concepts in the paper.

[33] BH-93 use a weaker definition. A separating behavior (m_L, m_H) defines type-dependence if (m_L, m_H) is a best-response with respect to a uniform probability distribution over the set of responses for the Receiver.

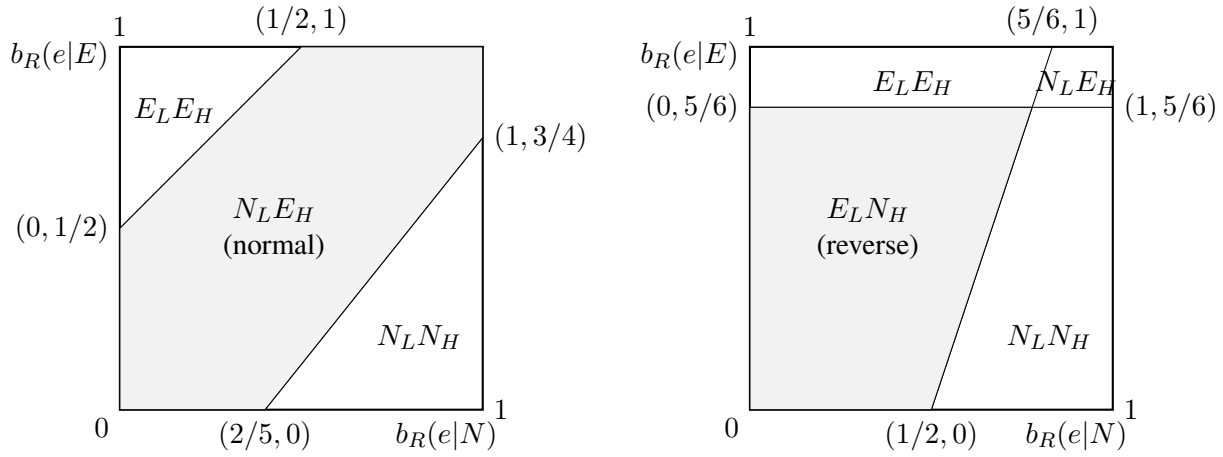

Figure 6: Type-dependencies: "normal" in Game 1 (left) and "reverse" in Game 2 (right)

In particular, π_1' in RHTE-1 captures normal type-dependence (N_L, E_H), while π_4' in RHTE-3 displays reverse type-dependence (E_L, N_H) (see Examples 2 and 4). In BH-92, type-dependence is an initial conjecture about "optimal" behavior that leads to a final equilibrium. In our approach, type-dependence is a posterior conjecture that justifies the off-path belief in the final equilibrium.

In rational equilibrium, beliefs are justified by rational strategies. Our approach is silent about the origin of rational behavior. It may emerge as adjustment processes triggered by type-dependence or stem from iterative reasoning such as mutual knowledge of rationality. For this reason, our approach is general and flexible enough to facilitate behavior adaptation in various contexts where type-dependence, learning, experience, or heuristics affect equilibrium behavior.

Our equilibrium concept leaves it to a player to formulate rational hypotheses at information sets on and off paths, and to evaluate multiple hypotheses. For instance, consider Game 2. Suppose that the employer, initially believing that neither worker signals education, observes the unexpected signal. Conditional on E, the employer could potentially face a choice between π_3' and π_4'; each one is a rational hypothesis about workers' separating behavior. How could the employer decide?

It is conceivable that the employer forms her second-order prior based on type-dependence. Given an information set m, let ρ_m denote a conditional ranking of the "sizes" of probability sets that rationalize type-contingent choices. Let $\pi_{|(E_L,N_H)}$ denote the rational hypothesis induced by a type-contingent choice (m_L, m_H) which is calibrated by a given prior information p about types.

In Game 2 of Figure 5, when conditioning on E, we have the following ranking:[34]

$$\rho_E(\pi_4' := \pi_{|(E_L,N_H)}) > \rho_E(\pi_3 := \pi_{|(E_L,E_H)}) > \rho_E(\pi_3' := \pi_{|(N_L,E_H)}). \tag{34}$$

When the employer initially believes that both worker types pool on N and yet, the unexpected

[34]Recall, π_3, π_3' and π_4' are rational hypotheses constructed in Example 4.

signal E arrives, reverse type-dependence implies that π_4' is the most-likely rational hypothesis. This captures RHTE-3. On the other hand, in Game 1 of Figure 2, conditional on N, we have[35]

$$\rho_N(\pi_1' := \pi_{|(N_L,E_H)}) > \rho_N(\pi_{\lambda=1} := \pi_{|(N_L,N_H)}). \tag{35}$$

When pooling on E is initially assumed and the employer receives the unexpected signal N, normal type-dependence implies that π_1' is the most-likely rational hypothesis, as in RHTE-1. This shows that the "sizes" of probability sets that rationalize type-contingent choices induce a (strict) likelihood relation on the set of rational hypotheses associated with the respective pure strategies. Moreover, a rational hypothesis associated with the salient separating behavior is the most-likely hypothesis conditional on an unexpected message in either of the two games implemented by BH-92. Whether this relationship is generally true is an open question which is left for future research.

7 Conclusion

We have suggested a solution concept for signaling games that allows for belief updating at information sets off the path via the Hypothesis Testing model of Ortoleva (2012). In Rational Hypothesis Testing Equilibrium, the uninformed player forms beliefs by selecting and updating hypotheses about rational behavior of the informed player. Such beliefs adhere to rational consistency. That is, beliefs are not only structurally consistent in the spirit of Kreps and Wilson (1982), but consistent with implications of mutual knowledge of rationality and the prior information about types.

The Rational Hypothesis Testing Equilibrium provides an alternative tool for equilibrium selection in signaling games in which at least one type has an incentive to defect to an unsent message. On the one hand, our solution concept refines sequential equilibria by eliminating those that fail rational consistency. On the other hand, our solution concept is general enough to be related with the Intuitive Criterion as well as with the undefeated equilibrium, and to account for the experimental finding in Brandts and Holt (1992, 1993), contrary to these two prominent refinement concepts.

A possible direction for future research is to extend the Rational Hypothesis Testing Equilibrium to general extensive-form games with incomplete information and observable actions. To this end, however, one needs to solve the intriguing result by Kreps and Ramey (1987) and characterize the class of extensive-form games for which sequential equilibria maintain structural consistency. We leave this for future research.

[35] Recall, π_1' and π_λ are rational hypotheses constructed in Example 2.

A Strategic Stability and Existence of RHTE

In this appendix, we show that rational equilibrium entails *strategic stability* (Kohlberg and Mertens, 1986). Furthermore, we can show that a rational equilibrium exists for each (finite) signaling game.

The Intuitive Criterion has the same limitation as PBE when $I(m^\circ)$ is a non-singleton set. In this case, all probability distributions in $\Delta(I(m^\circ))$ are admissible. To cope with this limitation, more stringent refinement concepts have been suggested, including strategic stability by Kohlberg and Mertens (1986). Roughly speaking, an equilibrium is strategically stable if each "nearby" game, obtained by perturbing both players' strategies by a tremble, has a "nearby" equilibrium.

Banks and Sobel (1987) and Cho and Sobel (1990) provide an elegant characterization of strategic stability, which we use as a definition. Fix a PBE (b_S^*, b_R^*, μ^*). For each $\theta \in \Theta$, let $u_S^*(\theta)$ be the equilibrium expected payoff, defined in (24).

For $b_R(\cdot|m^\circ) \in MBR(\Theta, m^\circ)$, let $u_S(\theta, m^\circ, b_R(\cdot|m^\circ)) := \sum_{a\in\mathcal{A}} u_S(\theta, m^\circ, a) b_R(a|m^\circ)$ denote the expected payoff from playing m° by θ. For each unsent message m°, denote by $\Psi(m^\circ)$ the set of all pairs $(\mu, H(m^\circ))$ where μ is a probability distribution on Θ and $H(m^\circ)$ is a subset of Θ such that, for the given PBE, there exists a best response $b_R \in MBR(\Theta, m^\circ)$ to m° such that

$$u^*(\theta) = u_S(\theta', m, b_R(\cdot|m)) \text{ for all } \theta \in H(m^\circ), \tag{36}$$

$$u^*(\theta') > u_S(\theta', m, b_R(\cdot|m)) \text{ for all } \theta' \notin H(m^\circ). \tag{37}$$

That is, b_R makes any type in $H(m^\circ)$ indifferent between defection and the equilibrium behavior, while other types strictly prefer to follow b_S^*. Note that it is possible to support the given equilibrium outcome (i.e., the expected equilibrium payoffs associated with b_S^* and b_R^* on the path) with belief μ conditional on m° if and only if $(\mu, H(m^\circ)) \in \Psi(m^\circ)$ for some $H(m^\circ) \subseteq \Theta$.

Cho and Sobel (1990) proved that a PBE is *strategically stable* if and only if for all $m^\circ \in \mathcal{M}^\circ$ and $q \in \Delta(\Theta)$, there exists a tuple $(\mu, H(m^\circ)) \in \Psi(m^\circ)$ such that $\mu \in$ *co-hull*$(q, \Delta(H(m^\circ)))$.

Example 6 Consider RHTE-1 (see Example 2). Suppose that the employer mixes between e and m with equal probabilities. This makes θ_L indifferent between choosing the equilibrium message, E, and deviating to the unsent message, N. Hence, $H(N) = \{\theta_L\}$ since θ_H strictly prefers to signal education no matter how the employer best responds to N. Knowing that only θ_L is vulnerable to defect to N, the employer best responds with m, showing that the RHTE-1 is strategically stable.[36]

In signaling games for which there is at least one Sender's type that could be vulnerable to defection to m° (i.e, $H(m^\circ) \neq \emptyset$), a PBE that is strategically stable constitutes a rational equilibrium.

Theorem 2 *A strategically stable PBE with $H(m^\circ) \neq \emptyset$ for each $m^\circ \in \mathcal{M}^\circ$ constitutes an RHTE.*

[36] RHTE-2 in Example 4 is strategically stable too.

If $H(m^\circ) = \emptyset$, no type has a (weak) incentive to deviate. In this case, stability has no bite. The notion of strategic stability is designed to "stabilize " the types that are vulnerable to deviations.

In Section 5, we have shown that the non-emptiness of $I(m^\circ)$ is not sufficient for the existence of an Intuitive RHTE. We are ready to prove that this condition implies the existence of an RHTE.

Proposition 5 *An RHTE exists, provided $I(m^\circ) \neq \emptyset$ for each unsent message m°.*

B RHTE in Monotone Signaling Games

In this appendix, we relate RHTE and *D1* Criterion, introduced by Banks and Sobel (1987). To this end, we focus on monotone signaling games (see Cho and Sobel, 1990). Many signaling games studied in the literature are monotone; e.g., the education game (Spence, 1973), the limit-pricing model (Milgrom and Roberts, 1982) or the product quality model (Miller and Plott, 1985).

$D1$ Criterion is a test for strategic stability which is stronger than the Intuitive Criterion.[37] In analogy to the stronger criterion, $D1$ Criterion reasons about types that have an incentive to deviate from a given equilibrium. However, $D1$ makes sharper inferences about beliefs by asking if some types in $I(m^\circ)$ are more likely to deviate than other ones. If such types exist, $D1$ requires to assign a zero probability to those Sender types that are less likely to send a given message off the path.[38]

Fix a PBE (b_S^*, b_R^*, μ^*). For each $m^\circ \in \mathcal{M}^\circ$, define

$$D(\theta \mid m^\circ) = \{b_R(\cdot|m^\circ) \in MBR(\Theta, m^\circ) : u^*(\theta) < u_S(\theta, m^\circ, b_R(\cdot|m^\circ))\}, \tag{38}$$

to be the set of best responses that make the type θ Sender strictly prefer defection, and

$$D^0(\theta \mid m^\circ) = \{b_R(\cdot|m^\circ) \in MBR(\Theta, m^\circ) : u^*(\theta) = u_S(\theta, m^\circ, b_R(\cdot|m^\circ))\}. \tag{39}$$

to be the set of best responses that make the type θ Sender indifferent towards defection.

$D1$ Criterion requires that each belief at m° has zero probability that θ defected to m° if there is another type θ' that strictly benefits from the deviation whenever θ weakly benefits from it; i.e.,

$$D(\theta \mid m^\circ) \cup D^0(\theta \mid m^\circ) \subseteq D(\theta' \mid m^\circ) \quad \text{for some } \theta' \in \Theta. \tag{40}$$

[37] Other known concepts that nest intuitive equilibrium and strategic stability include the forward induction equilibrium by Cho and Kreps (1987) as well as the divine and universal divine equilibrium by Banks and Sobel (1987).

[38] Fudenberg and He (2020) suggest a solution concept, called (uniformly) rationally-compatible equilibrium, that builds on a similar "more-likely-to-deviate-than" relation, called rationally-compatible order. For each Receiver's strategy, the authors compare the expected payoff from deviating to m° by a type with the maximum expected payoff that the type could achieve by playing another message than m°. If some type θ weakly prefers a given message m° to the alternative, best-paying message, and another type θ' strictly prefers the former message to the latter, the ratio of the posterior of θ' to the posterior of θ should be not smaller than the same prior ratio. Upon request, we can prove that for each (uniformly) rationally-compatible equilibrium, a path-equivalent rational equilibrium exists.

In general, $D1$ is weaker than strategic stability. Yet, Cho and Sobel (1990) derived a remarkable result showing that $D1$ and strategic stability are equivalent in monotone signaling games.

Definition 8 (Monotonicity) *For all $m \in \mathcal{M}$, and all $b_R, b'_R \in BR(\Theta, m)$ if $u_S(\theta, m, b_R(\cdot|m)) > u_S(\theta, m, b'_R(\cdot|m))$ for some $\theta \in \Theta$, then $u_S(\theta', m, b_R(\cdot|m)) > u_S(\theta', m, b'_R(\cdot|m))$ for all $\theta' \in \Theta$.*

In monotone games, all types have identical (strict) preferences over the Receiver's actions.[39] In such games, a PBE that passes $D1$ Criterion is a rational equilibrium that is strategically stable.

Proposition 6 *In monotone signaling games, each $D1$ equilibrium is a strategically stable RHTE.*

C Proofs

Proof of Lemma 1. Let (b_S^*, b_R^*, μ^*) be a PBE. Consider the equilibrium strategy b_S^*, and the Receiver's belief $\beta = b_S^*$. This belief together with the prior p defines a hypothesis such that $\sum_{\theta\in\Theta} b_S^*(m^*|\theta)p(\theta) > 0$. Hence, m^* is an information set on the path. For each $(m, \theta) \in \mathcal{M} \times \Theta$,

$$\pi^*(m, \theta) = b_S^*(m|\theta)p(\theta). \tag{41}$$

By Equation (4), if $\sum_{\theta\in\Theta} b_S^*(m^*|\theta)p(\theta) > 0$, $\mu^*(\theta|m)$ should be

$$\mu^*(\theta|m) = \frac{b_S^*(m|\theta)p(\theta)}{\sum_{\theta'\in\Theta} b_S^*(m|\theta')p(\theta')} \quad \text{for any } \theta \in \Theta.$$

Then, for any $m \in \mathcal{M}$ such that $\pi^*(m, \Theta) > 0$, we have

$$\mu(\theta|m) = \frac{\pi^*(m, \theta)}{\pi^*(m, \Theta)} = \frac{b_S^*(m|\theta)p(\theta)}{\sum_{\theta'\in\Theta} b_S^*(m|\theta')p(\theta')} = \mu^*(\theta|m) \quad \text{for each } \theta \in \Theta, \tag{42}$$

showing that π^* induces the PBE beliefs on the path.

Now, fix $m' \in \mathcal{M} \setminus \{m^*\}$. Consider the following strategy b_S for the Sender:

$$b_S(m|\theta) = \begin{cases} \left(\frac{\mu^*(\theta|m')}{p(\theta)}\right)\frac{1}{X}, & \text{for } m = m', \\ 1 - \left(\frac{\mu^*(\theta|m')}{p(\theta)}\right)\frac{1}{X}, & \text{for some } m \in (\mathcal{M} \setminus \{m'\}), \\ 0, & \text{otherwise}, \end{cases} \tag{43}$$

[39] In monotone signaling games, Liu and Pei (2020) show that there may exist a non-monotone equilibrium in which a "higher" type signals a "lower" message. The authors provide conditions, including monotone-supermodularity, under which all non-monotone strategies are non-rationalizable, implying that only monotone equilibria exist.

where $X := \sum_{\theta\in\Theta} \frac{\mu^*(\theta|m')}{p(\theta)}$. Since $X \geq \frac{\mu^*(\theta|m')}{p(\theta)}$ and $\sum_{m\in\mathcal{M}} b_S(m|\theta) = 1$ for any $\theta \in \Theta$, b_S is well-defined. According to this strategy, only the types in the support of $\mu^*(\cdot|m')$ play m' with a strictly positive probability (i.e., $b_S(m'|\theta) > 0$ for each $\theta \in \Theta$, such that $\mu^*(\theta|m') > 0$).

Thus, $\beta = b_S$ and the prior p define the hypothesis $\pi_{m'}$ as follows: For each $(m, \theta) \in \mathcal{M} \times \Theta$,

$$\pi_{m'}(m, \theta) = b_S(m|\theta)p(\theta). \tag{44}$$

By updating $\pi_{m'}$ conditional on m', by (43) and (44), we have

$$\mu(\theta|m') = \frac{\pi_{m'}(m', \theta)}{\pi_{m'}(m', \Theta)} = \frac{\frac{\mu^*(\theta|m')}{\sum_{\theta\in\Theta} \frac{\mu^*(\theta|m')}{p(\theta)}}}{\frac{1}{\sum_{\theta\in\Theta} \frac{\mu^*(\theta|m')}{p(\theta)}}} = \mu^*(\theta|m') \text{ for every } \theta \in \Theta, \tag{45}$$

yielding the PBE belief off the path.

Since m' was chosen arbitrarily, for each out-of-equilibrium message $m' \in \mathcal{M}\backslash\{m^*\}$, there is a hypothesis $\pi_{m'}$ that induces $\mu^*(\cdot|m')$. Let $\{\pi_{m'}\}_{m'\in\mathcal{M}\backslash\{m^*\}}$ be the family of such hypotheses. Thus, $\pi^* \cup \{\pi_{m'}\}_{m'\in\mathcal{M}\backslash\{m^*\}}$ induces $\mu^* = \{\mu^*(\cdot|m)\}_{m\in\mathcal{M}}$, showing that μ^* is structurally consistent. ■

Proof of Proposition 1. We show that the following family of PBEs is supported by an RHTE:

(i) $b_S^*(e^*|\theta_L) = b_S^*(e^*|\theta_H) = 1$ such that $e_0 \leq e^* \leq y := (\theta_H - \theta_L)\big(\theta_L - (1-\alpha)\theta_H\big)$.

$$(ii)\ w^*(e) = \begin{cases} \theta_L & \text{if } e < e^*, \\ \mathbb{E}(\theta) & \text{if } e^* \leq e \leq e_n, \\ \theta_H & \text{if } e_n < e \leq e_N. \end{cases} \qquad (iii)\ \mu^*(\theta_L|e) = \begin{cases} 1 & \text{if } e < e^*, \\ 1-\alpha & \text{if } e^* \leq e \leq e_n, \\ 0 & \text{if } e_n < e \leq e_N. \end{cases}$$

Fix a pooling message e_i^* such that $e_0 \leq e_i^* \leq y$. We first construct a rational hypothesis π_0^* that justifies the posterior on the path $\mu^*(\theta|e_i^*)$. The equilibrium strategy b_S^* is rational as it best responds to $w^*(e)$. Hence, $\beta = b_S^*$, together with p, induces the following rational hypothesis:

$$\pi_0^*(e, \theta) = b_S^*(e|\theta)p(\theta) \text{ for each } (e, \theta) \in \mathcal{M} \times \Theta. \tag{46}$$

By updating π_0^* conditional on e_i^*, we obtain the PBE belief $\mu^*(\theta_L|e_i^*) = p(\theta_L) = 1 - \alpha$.

Now, for each $e \in \mathcal{M}^\circ = \mathcal{M} \setminus \{e_i^*\}$, we construct a rational hypothesis under which e is feasible. W.l.o.g., we limit our attention to the following partition of $\mathcal{M}^\circ$:

$$\mathcal{P}(\mathcal{M}^\circ) = \big\{ \underbrace{\{e_0, \ldots, e_{i-1}\}}_{\text{Case 1}}, \underbrace{\{e_{i+1}, \ldots, e_n\}}_{\text{Case 2}}, \underbrace{\{e_{n+1}, \ldots, e_N\}}_{\text{Case 3}} \big\}, \tag{47}$$

where $e_n := \theta_L(\theta_H - \theta_L)$ and $e_N := \theta_H(\theta_H - \theta_L)$.

We consider the three cases.

Case 1. Fix $e' \in \{e_0, \ldots, e_{i-1}\}$. Consider the following strategy $w_1(e)$ for the employer together with the posterior that rationalizes it:

$$w_1(e) = \begin{cases} w' & \text{if } e = e', \\ \theta_H, & \text{if } e = e_n, \\ \theta_L, & \text{elsewhere,} \end{cases} \quad \text{and} \quad \mu(\theta_L|e) = \begin{cases} \frac{\theta_H - w'}{\theta_H - \theta_L}, & \text{if } e = e', \\ 0, & \text{if } e = e_n \\ 1, & \text{elsewhere.} \end{cases} \tag{48}$$

Note that w' must satisfy the following conditions. First, w' for e' has to make the low-productivity type better off than his payoff for offering education level 0 at the lowest wage θ_L; i.e,

$$u(\theta_L, e', w') = w' - \frac{e'}{\theta_L} > \theta_L - \frac{0}{\theta_L} = u(\theta_L, e_0, \theta_L), \text{ or equivalently, } w' > \theta_L + \frac{e'}{\theta_L}. \tag{49}$$

Second, w' for e' has to make the high-productivity type worse off than his payoff for offering e_n at the highest wage θ_H; i.e.,

$$u(\theta_H, e', w') = w' - \frac{e'}{\theta_H} < \theta_H - \frac{e_n}{\theta_H} = u(\theta_H, e_n, \theta_H), \text{ or equivalently, } w' < \theta_H + \frac{e'}{\theta_H} - \frac{e_n}{\theta_H}. \tag{50}$$

By (49) and (50), $\mu(\theta_L|e')$ is defined by

$$\frac{e_n - e'}{\theta_H(\theta_H - \theta_L)} < \mu(\theta_L|e') := \frac{\theta_H - w'}{\theta_H - \theta_L} < 1 - \frac{e'.}{\theta_L(\theta_H - \theta_L)}. \tag{51}$$

The worker's strategy $b_S := (b_S(e'|\theta_L) = 1, b_S(e_n|\theta_H) = 1)$ best responds to $w_1(e)$. Hence, it is rational. Therefore, $\beta = b_S$, together with the prior p, induces the rational hypothesis $\pi_1(e')$, yielding the PBE belief $\mu^*(\theta_L|e') = 1$ for each $e' \in \{e_0, \ldots, e_{i-1}\}$.

Case 2. Fix $e' \in \{e_{i+1}, \ldots, e_n\}$. Consider the following strategy $w_2(e)$ for the employer together with the posterior that rationalizes it:

$$w_2(e) = \begin{cases} \theta_H, & \text{if } e = e', \\ \theta_L, & \text{elsewhere,} \end{cases} \quad \text{and} \quad \mu(\theta_L|e) = \begin{cases} 0, & \text{if } e = e', \\ 1, & \text{elsewhere.} \end{cases} \tag{52}$$

The strategy $b'_S := (b'_S(e'|\theta_L) = 1, b'_S(e'|\theta_H) = 1)$ best responds to $w_2(e)$. Hence, it is rational. Therefore, $\beta' = b'_S$, together with the prior p, induces the rational hypothesis $\pi_2(e')$, yielding the PBE belief $\mu^*(\theta_L|e') = p(\theta_L) = 1 - \alpha$ for each $e' \in \{e_{i+1}, \ldots, e_n\}$.

Case 3. Fix $e' \in \{e_{n+1}, \ldots, e_N\}$. Consider the following strategy $w_3(e)$ for the employer together

with the posterior that rationalizes it:

$$w_3(e) = \begin{cases} \theta_H, & \text{if } e = e', \\ \theta_L, & \text{elsewhere,} \end{cases} \quad \text{and} \quad \mu(\theta_L|e) = \begin{cases} 0, & \text{if } e = e', \\ 1, & \text{elsewhere.} \end{cases} \tag{53}$$

The worker's strategy $b''_S := (b''_S(e_0|\theta_L) = 1, b''_S(e'|\theta_H) = 1)$ best responds to $w_3(e)$. Hence, it is rational. Therefore, $\beta'' = b''_S$, together with the prior p, induces the rational hypothesis $\pi_3(e')$, yielding the PBE belief $\mu^*(\theta_L|e') = 0$ for each $e' \in \{e_{n+1}, \ldots, e_N\}$.

Finally, we can suitably choose a second-order prior ρ such that

$$supp(\rho) = \{\pi_0, \pi_1(e)_{e\in\{e_0,\ldots,e_{i-1}\}}, \pi_2(e)_{e\in\{e_{i+1},\ldots,e_n\}}, \pi_3(e)_{e\in\{e_{n+1},\ldots,e_N\}}\},$$

$$\{\pi_0\} := \arg\max_{\pi\in supp(\rho)} \rho(\pi) \text{ and } \{\pi^{**}(e)\} := \arg\max_{\pi\in supp(\rho)} \rho_e(\pi) \text{ for each } e \in \mathcal{M}^\circ,$$

The constructed RHTE, $(b^*_S, w^*, \rho, \mu^*_\rho)$, is the PBE with pooling on e^*_i where $e_0 \le e^*_i \le y$. ■

Proof of Proposition 2. Consider the following separating PBE which is the Riley outcome:

(i) $b^*_S(e_0|\theta_L) = 1$ and $b^*_S(e_n|\theta_H) = 1$,

$$(ii)\ w^*(e) = \begin{cases} \theta_L & \text{if } e_0 \le e < e_n, \\ \theta_H & \text{if } e_n \le e \le e_N, \end{cases} \quad \text{and} \quad \mu^*(\theta_H|e) = \begin{cases} 0 & \text{if } e_0 \le e < e_n, \\ 1 & \text{if } e_n \le e \le e_N. \end{cases}$$

We first construct a rational hypothesis π^*_0 that justifies the posterior beliefs on the path, $\mu^*(\theta_H|e_L) = 0$ and $\mu^*(\theta_H|e_H) = 1$. Since b^*_S best responds to $w^*(e)$, the equilibrium strategy b^*_S is rational. Hence, $\beta = b^*_S$, together with p, induces the following rational hypothesis:

$$\pi^*_0(e, \theta) = b^*_S(e|\theta)p(\theta) \text{ for each } (e, \theta) \in \mathcal{M} \times \Theta. \tag{54}$$

By updating π^*_0 on $e^*_L = e_0$ and $e^*_H = e_n$, we obtain $\mu(\theta_H|e_0) = 0$ and $\mu(\theta_H|e_n) = 1$, respectively.

For each $e \in \mathcal{M}^\circ = \mathcal{M} \setminus \{e_0, e_n\}$, we construct a rational hypothesis under which e is feasible. In Case 1, we consider $e \in \{e_1, \ldots, e_{n-1}\}$ and in Case 2, $e \in \{e_{n+1}, \ldots, e_N\}$.

Case 1: Fix $e' \in \{e_1, \ldots, e_{n-1}\}$. Consider the following strategy $w_1(e)$ for the employer and the PBE belief, $\mu(\theta_L|e)$, that rationalizes it:

$$w_1(e) = \begin{cases} \frac{(\theta_H - \theta_L)e}{e_n} + \theta_L & \text{if } e \le e_n, \\ \theta_H, & \text{if } e > e_n, \end{cases} \quad \text{and} \quad \mu(\theta_L|e) = \begin{cases} 1 - \frac{e}{e_n}, & \text{if } e \le e_n, \\ 0, & \text{if } e > e_n. \end{cases} \tag{55}$$

Note that $w_1(e)$ is rational since $\mu(\theta_L|e)\theta_L + (1 - \mu(\theta_L|e))\theta_H = w_1(e)$ for any e.

Given $w_1(e)$, the low-productivity type cannot be better off than getting the lowest wage θ_L by choosing e'; i.e.,

$$u_S(\theta_L, e', w_1(e')) = \frac{(\theta_H - \theta_L)e'}{e_n} + \theta_L - \frac{e'}{\theta_L} = \theta_L = u_S(e_L, e, w_1(e)) \qquad \text{for any } e \leq e_n,$$
$$u_S(\theta_L, e', w_1(e')) = \frac{(\theta_H - \theta_L)e'}{e_n} + \theta_L - \frac{e'}{\theta_L} > \theta_H - \frac{e}{\theta_L} = u_S(\theta_L, e, w_1(e)) \quad \text{for any } e > e_n.$$

Hence, e' is optimal for θ_L-type. Choosing e_n is optimal for the high-productivity type; i.e.,

$$u_S(\theta_H, e_n, w_1(e_n)) = \theta_H - \frac{e_n}{\theta_H} \geq \frac{(\theta_H - \theta_L)e}{e_n} + \theta_L - \frac{e}{\theta_H} = u_S(\theta_H, e, w_1(e)) \qquad \text{for any } e \leq e_n,$$
$$u_S(\theta_H, e_n, w_1(e_n)) = \theta_H - \frac{e_n}{\theta_H} > \theta_H - \frac{e}{\theta_H} = u_S(\theta_H, e, w_1(e)) \qquad \text{for any } e > e_n.$$

In this way, the worker's strategy $b_S := (b_S(e'|\theta_L) = 1,\ b_S(e_n|\theta_H) = 1)$ best responds to $w_1(e)$, showing it is rational. Thus, $\beta = b_S$ and the prior p define the rational hypothesis $\pi_1(e')$ which after updating conditional on e' yields the PBE belief $\mu^*(\theta_L|e') = 1$ for each $e' \in \{e_1, \ldots, e_{n-1}\}$.

Case 2: Fix $e' \in \{e_{n+1}, \ldots, e_N\}$. Consider the following strategy $w_2(e)$ for the employer and the PBE belief, $\mu(\theta_L|e)$, that rationalizes it:

$$w_2(e) = \begin{cases} \theta_L & \text{if } e \leq e_n, \\ \frac{(\theta_H - \theta_L)e}{e_N} + \theta_L, & \text{if } e > e_n, \end{cases} \quad \text{and} \quad \mu(\theta_L|e) = \begin{cases} 1, & \text{if } e \leq e_n, \\ 1 - \frac{e}{e_N}, & \text{if } e > e_n. \end{cases} \tag{56}$$

Again, $w_2(e)$ is rational since $\mu(\theta_L|e)\theta_L + (1 - \mu(\theta_L|e))\theta_H = w_2(e)$ for any e.

Given $w_2(e)$, choosing $e_0 = 0$ is optimal for the low-productivity type; i.e.,

$$u_S(\theta_L, e_0, w_2(e_0)) = \theta_L \geq \theta_L - \frac{e}{\theta_L} = u_S(\theta_L, e, w_2(e)) \qquad \text{for any } e \leq e_n,$$
$$u_S(\theta_L, e_0, w_2(e_0)) = \theta_L > \frac{(\theta_H - \theta_L)e}{e_N} + \theta_L - \frac{e}{\theta_L} = u_S(\theta_L, e, w_2(e)) \qquad \text{for any } e > e_n.$$

The high-productivity type cannot be paid better than the lowest wage θ_L by choosing e'; i.e.,

$$u_S(\theta_H, e', w_2(e')) = \frac{(\theta_H - \theta_L)e'}{e_N} + \theta_L - \frac{e'}{\theta_H} > \theta_L - \frac{e}{\theta_H} = u_S(\theta_H, e, w_2(e)) \qquad \text{for any } e \leq e_n,$$
$$u_S(\theta_H, e', w_2(e')) = \theta_L = \frac{(\theta_H - \theta_L)e}{e_N} + \theta_L - \frac{e}{\theta_H} = u_S(\theta_H, e, w_2(e)) \qquad \text{for any } e > e_n.$$

Hence, e' is optimal for θ_H-type. In this way, the worker's strategy $b'_S := (b'_S(e_0|\theta_L) = 1,\ b'_S(e'|\theta_H) = 1)$ best responds to $w_2(e)$. Hence, it is rational. Thus, $\beta' = b'_S$, together with the prior p, induces the rational hypothesis $\pi_2(e')$, yielding the PBE belief $\mu^*(\theta_L|e') = 0$ for each $e' \in \{e_{n+1}, \ldots, e_N\}$.

Finally, we can suitably choose a second-order prior ρ such that

$$supp(\rho) = \{\pi_0, \pi_1(e)_{e\in\{e_1,\ldots,e_{n-1}\}}, \pi_2(e)_{e\in\{e_{n+1},\ldots,e_N\}}\},$$

$$\{\pi_0\} := \arg\max_{\pi\in supp(\rho)} \rho(\pi) \text{ and } \{\pi^{**}(e)\} := \arg\max_{\pi\in supp(\rho)} \rho_e(\pi) \text{ for each } e \in \mathcal{M}^\circ,$$

showing that an RHTE $(b_S^*, w^*, \rho, \mu_\rho^*)$ supporting the Separating PBE exists. ■

Proof of Proposition 3. We consider two cases. In Case 1, we show an Unintuitive PBE that is a rational equilibrium. In Case 2, we present an Intuitive PBE that fails to be a rational equilibrium.

Case 1. Consider the game in Figure 5 which has the family of PBE with pooling on N given by

$$b_S^*(N|\theta_L) = b_S^*(N|\theta_H) = 1,\ b_R^*(m|E) = b_R^*(e|N) = 1, \qquad \text{(PBE-3)}$$
$$\mu^*(\theta_L|N) = 1/3 \text{ and } \mu^*(\theta_L|E) \geq 1/2.$$

In Example 4, we have shown that the PBE with $\mu^*(\theta_L|E) = 1$ is supported by RHTE-3.

Now, we argue that RHTE-3 fails the Intuitive Criterion. According to the Intuitive Criterion, θ_H can be better off than his equilibrium payoff if he plays the out-of-equilibrium message E. That is, $I(E) = \{\theta_H\}$. This induces the out-of-equilibrium belief $\mu(\theta_H|E) = 1$. However, if the Receiver learns that E was chosen by θ_H, she will play e instead of m. Given that the Receiver responds e against E, type θ_H will indeed choose E. Thus, RHTE-3 fails the Intuitive Criterion.

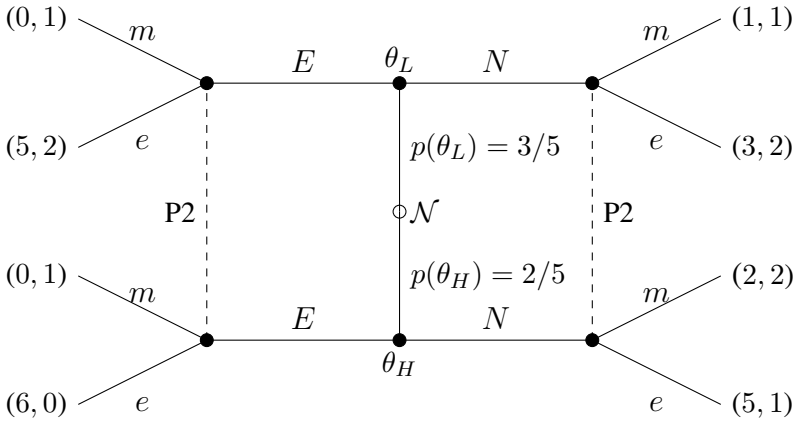

Figure 7: Intuitive PBE that violates rational consistency.

Case 2. Consider the game in Figure 7. It has the following family of PBEs with pooling on N:

$$b_S^*(N|\theta_L) = b_S^*(N|\theta_H) = 1,\ b_R^*(m|E) = b_R^*(e|N) = 1, \qquad \text{(PBE-4)}$$
$$\mu^*(\theta_L|N) = 3/5 \text{ and } \mu^*(\theta_L|E) \leq 1/2.$$

First, we show that there an RHTE that supports one of the pooling equilibria does not exist. Note that any strategy $b_R = (b_R(\cdot|E), b_R(\cdot|N))$ is rational. Thus, $\mathcal{B}_R = \mathcal{B}_R^\bullet$. To rationalize the Receiver's behavior off the path, $b_R^*(m|E)$, we need to determine the strategies for the Sender that best respond to some $b_R \in \mathcal{B}_R^\bullet$. Denote by $x := b_R(m|E)$ and $y := b_R(m|N)$ the probabilities that the Receiver plays m in response to E and to N, respectively. Then, θ_L will choose E if

$$2y - 5x \geq -2 \quad \Leftrightarrow \quad y \geq \frac{5}{2}x - 1. \tag{57}$$

Similarly, θ_H will choose E if

$$3y - 6x \geq -1 \quad \Leftrightarrow \quad y \geq 2x - \frac{1}{3}. \tag{58}$$

Note that $2x - \frac{1}{3} > \frac{5}{2}x - 1$ for any $x \in [0,1]$. Hence, (57) is satisfied with strict inequality for each $(x,y) \in [0,1] \times [0,1]$ that satisfies (58). This means that θ_L strictly prefers E over N whenever θ_H weakly prefers E. That is, for each $b_R \in \mathcal{B}_R^\bullet$, if

$$\sum_{a \in \mathcal{A}} u_S(\theta_H, E, a) b_R(a|E) \quad \geq \quad \sum_{a \in \mathcal{A}} u_S(\theta_H, N, a) b_R(a|N), \text{then} \tag{59}$$
$$\sum_{a \in \mathcal{A}} u_S(\theta_L, E, a) b_R(a|E) \quad > \quad \sum_{a \in \mathcal{A}} u_S(\theta_L, N, a) b_R(a|N). \tag{60}$$

Hence, there is no $b_S \in \mathcal{B}_S^\bullet$ such that $b_S(E|\theta_H) > b_S(E|\theta_L)$. Thus, by updating a rational hypothesis defined by $\beta \in \mathcal{B}_S^\bullet$ and the prior p, we have $\mu_\rho(\theta_L|E) \geq 3/5$. However, for each $\mu_\rho(\theta_L|E) \geq 3/5$, the Receiver chooses e instead of m. Thus, there is no RHTE supporting PBE-4.[40]

Now, we show that PBE-4 passes the Intuitive Criterion. According to the Intuitive Criterion, both types θ_L and θ_H could be better off than their equilibrium payoff by choosing E. That is, $I(E) = \{\theta_L, \theta_H\}$. In this case, the Intuitive Criterion admits all beliefs over $I(E) = \{\theta_L, \theta_H\}$, including any belief, such that $\mu(\theta_L|E) \leq 1/2$. We know that no player has an incentive to deviate to E as long as $\mu(\theta_L|E) \leq 1/2$. Therefore, each PBE-4 passes the Intuitive Criterion. ■

[40] Interestingly enough, PBE-4 can neither be supported by Ortoleva's HTE. In this signaling game, all Receiver's strategies are rational (i.e., $\mathcal{B}_R = \mathcal{B}_R^\bullet$). Therefore, Ortoleva's HTE and RHTE coincide.

Proof of Lemma 2. Let (b_S^*, b_R^*, μ^*) be a PBE and m° be an unsent message. Since $I(m^\circ) \neq \emptyset$, there is an action $a^\circ \in BR(\Theta, m^\circ)$ and a set of types $I(m^\circ; a^\circ) \subseteq I(m^\circ)$ such that

$$u_S^*(\theta) \leq u_S(\theta, m^\circ, a^\circ) \text{ for each } \theta \in I(m^\circ; a^\circ). \tag{61}$$

Hence, we can define a rational strategy b_R° as follows: For each $m \in \mathcal{M} \setminus \{m^\circ\}$,

$$b_R^\circ(a|m) = b_R^*(a|m) \text{ for each } a \in \mathcal{A}, \tag{62}$$

and for $m = m^\circ$,

$$b_R^\circ(a|m) = \begin{cases} 0, & \text{for } a \in \mathcal{A} \setminus \{a^\circ\}, \\ 1, & \text{for } a = a^\circ. \end{cases} \tag{63}$$

Note that b_R° is identical to the equilibrium strategy b_R^* except when the Receiver observes m°. Since $b_R^\circ(a^\circ|m^\circ) = 1$ is a best response to some belief over Θ, b_R° is rational; i.e, $b_R^\circ \in \mathcal{B}_R^\bullet$.

For the Sender, let b_S° be a best-response to b_R° that generates m°. That is, if $\theta \in \Theta \setminus I(m^\circ; a^\circ)$,

$$b_S^\circ(m|\theta) = b_S^*(m|\theta) \text{ for each } m \in \mathcal{M}, \tag{64}$$

and if $\theta \in I(m^\circ; a^\circ)$,

$$b_S^\circ(m|\theta) = \begin{cases} 0, & \text{for } m \in \mathcal{M} \setminus \{m^\circ\}, \\ 1, & \text{for } m = m^\circ. \end{cases} \tag{65}$$

The Receiver's belief $\beta = b_S^\circ$ together with the prior p defines the rational hypothesis π_{m° given by

$$\pi_{m^\circ}(m, \theta) := \begin{cases} p(\theta), & \text{if } (m, \theta) \in (\{m^\circ\} \times I(m^\circ; a^\circ)), \\ b_S^*(m|\theta)p(\theta), & \text{if } (m, \theta) \in (\mathcal{M} \times \Theta \setminus I(m^\circ; a^\circ)). \end{cases} \tag{66}$$

By updating π_{m° given m°, we get $\mu_\rho(\theta|m^\circ) > 0$ for each $\theta \in I(m^\circ; a^\circ)$. Proceeding in this way for each $m^\circ \in \mathcal{M}^\circ$, we construct a rational hypothesis π_{m° under which m° is feasible. ■

Proof of Theorem 1. Let (b_S^*, b_R^*, μ^*) be an Intuitive PBE. First, we show that for each message m° off the path, there exists a rational hypothesis π_{m° that induces the PBE belief in Equation (28). Second, we construct a rational hypothesis that induces the PBE belief on the path.

Step 1. Fix an out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$. By assumption, there is a response a° to

m° such that (i) $I(m^\circ; a^\circ) \neq \emptyset$ and (ii) the PBE belief conditional on m° (i.e., (28)) is given by

$$\mu^*(\theta|m^\circ) = \begin{cases} \frac{p(\theta)}{\sum_{\theta' \in I(m^\circ;a^\circ)} p(\theta')} & \text{if } \theta \in I(m^\circ; a^\circ), \\ 0 & \text{if } \theta \notin I(m^\circ; a^\circ). \end{cases}$$

Then, we apply the arguments presented in the proof of Lemma 2 to construct (i) a rational strategy b_R° for the Receiver (as in (62) and (63)), (ii) a rational strategy b_S° for the Sender (as in (64) and (65)), and (iii) a rational hypothesis π_{m° under which m° is feasible (as in (74)), which is given by

$$\pi_{m^\circ}(m, \theta) := \begin{cases} p(\theta), & \text{if } (m, \theta) \in \{m^\circ\} \times I(m^\circ; a^\circ), \\ b_S^*(m|\theta)p(\theta), & \text{if } (m, \theta) \in \mathcal{M} \times \Theta \setminus I(m^\circ; a^\circ). \end{cases} \tag{67}$$

Since $\pi_{m^\circ}(m^\circ, \Theta) > 0$, by updating π_{m° given m°, we get $\mu_\rho(\theta|m^\circ)$ for any $\theta \in \Theta$ given by

$$\begin{aligned} \mu_\rho(\theta|m^\circ) &= \frac{\pi_{m^\circ}(m^\circ, \theta)}{\sum_{\theta' \in \Theta} \pi_{m^\circ}(m^\circ, \theta')} = \frac{\pi_{m^\circ}(m^\circ, \theta)}{\sum_{\theta' \in I(m^\circ;a^\circ)} \pi_{m^\circ}(m^\circ, \theta')} \quad (\because b_S^*(m^\circ|\theta) = 0, \ \forall \theta \notin I(m^\circ; a^\circ)) \\ &= \frac{\pi_{m^\circ}(m^\circ, \theta)}{\sum_{\theta' \in I(m^\circ;a^\circ)} p(\theta')}, \end{aligned}$$

Thus, $\mu_\rho^*(\cdot|m^\circ) = \mu^*(\cdot|m^\circ)$, as requested.

Step 2. Consider b_S^*. Since $b_R^* \in \mathcal{B}_R^\bullet$ and b_S^* best responds to b_R^*, b_S^* is rational (i.e., $b_S^* \in \mathcal{B}_S^\bullet$). Hence, $\beta = b_S^*$ and p define the rational hypothesis π^*; i.e., for each $(m, \theta) \in \mathcal{M} \times \Theta$,

$$\pi^*(m, \theta) = b_S^*(m|\theta)p(\theta). \tag{68}$$

Finally, we can choose a second-order prior ρ with $supp(\rho) = \{\pi^*, \pi_{m^\circ}^{**}\}_{m^\circ \in \mathcal{M}^\circ}$, such that

$$\{\pi^*\} := \arg\max_{\pi \in supp(\rho)} \rho(\pi) \ \text{ and } \ \{\pi_{m^\circ}^{**}\} := \arg\max_{\pi \in supp(\rho)} \rho_{m^\circ}(\pi) \ \text{ for each } m^\circ \in \mathcal{M}^\circ, \tag{69}$$

showing that there exists an RHTE, $(b_S^*, b_R^*, \rho, \mu_\rho^*)$, supporting the Intuitive PBE. ■

Proof of Corollary 1. Consider an Intuitive PBE, (b_S^*, b_R^*, μ^*). Fix a message $m^\circ \in$ off the path. By assumption, $I(m^\circ) = \{\theta_{m^\circ}\}$. Thus, $\mu^*(\theta_{m^\circ}|m^\circ) = 1$ is the PBE belief. Since $I(m^\circ) = \{\theta_{m^\circ}\}$, there exists $a^\circ \in BR(\Theta, m^\circ)$ such that $I(m^\circ) = I(m^\circ; a^\circ) \neq \emptyset$; i.e., $u_S^*(\theta_{m^\circ}) \leq u_S(\theta_{m^\circ}, m^\circ, a^\circ)$. Moreover, given $a^\circ \in BR(\Theta, m^\circ)$, $\mu^*(\theta_{m^\circ}|m^\circ) = 1$ satisfies Condition (28) in Theorem 1. Thus, by Theorem 1, an RHTE supporting the Intuitive PBE exists. ■

Proof of Proposition 4. Let (b_S^*, b_R^*, μ^*) be a given PBE and $m^\circ \in \mathcal{M}^\circ$ a message off the path. Let $(b_S^{**}, b_R^{**}, \mu^{**})$ be an alternative PBE, where $K(m^\circ) \neq \emptyset$, that does not defeat the given PBE. Given the equilibrium strategies b_R^* and b_R^{**}, we construct the following strategy b'_R for the Receiver:

$$b'_R(a|m) = b_R^*(a|m) \text{ for each } a \in A \text{ if } m \in \mathcal{M} \setminus \{m^\circ\}, \text{ and} \tag{70}$$

$$b'_R(a|m) = b_R^{**}(a|m) \text{ for each } a \in A \text{ if } m = m^\circ. \tag{71}$$

Since b'_R is identical to b_R^* for each $m \in \mathcal{M} \setminus \{m^\circ\}$ and to b_R^{**} for m°, b'_R is rational.

For $\theta \in \Theta$, let $u_S^*(\theta)$ be the equilibrium expected payoff for θ by playing $b_S^*(\cdot|\theta)$, and $u^{**}(\theta)$ by playing $b_S^{**}(\cdot|\theta)$ (see (24)). Each $\theta \notin K(m^\circ)$ does not have an incentive to deviate to m° given b'_R.

By assumption, each $\theta \in K(m^\circ)$ is strictly better off by deviating to m° given b'_R; i.e., $u^{**}(\theta) > u^*(\theta)$. Since the alternative PBE does not defeat the given PBE, for each $\theta \in K(m^\circ)$,

$$\mu^*(\theta|m^\circ) = \frac{p(\theta)\xi(\theta)}{\sum_{\theta' \in K(m^\circ)} p(\theta')\xi(\theta')} = \frac{p(\theta)}{\sum_{\theta' \in K(m^\circ)} p(\theta')} \tag{72}$$

since $\xi(\theta) = 1$. Now, we construct a best response b'_S by the Sender to b'_R as follows: For each $\theta \notin K(m^\circ)$, $b'_S(m|\theta) = b_S^*(m|\theta)$ for each $m \in \mathcal{M}$, and for each $\theta \in K(m^\circ)$,

$$b'_S(m|\theta') = \begin{cases} 0, & \text{for } m \in \mathcal{M} \setminus \{m^\circ\} \\ 1, & \text{for } m = m^\circ. \end{cases} \tag{73}$$

The belief $\beta = b'_S$ together with the prior p defines the rational hypothesis π_{m° given by

$$\pi_{m^\circ}(m, \theta) := \begin{cases} p(\theta), & \text{if } (m, \theta) \in (\{m^\circ\} \times K(m^\circ)), \\ b_S^*(m|\theta)p(\theta), & \text{if } (m, \theta) \in (\mathcal{M} \times \Theta \setminus K(m^\circ)). \end{cases} \tag{74}$$

By updating π_{m° given m° via Bayes' rule, we obtain

$$\mu_\rho(\theta|m^\circ) = \frac{p(\theta)}{\sum_{\theta' \in K(m^\circ)} p(\theta')} \text{ for each } \theta \in K(m^\circ). \tag{75}$$

Hence, $\mu_\rho(\theta|m^\circ) = \mu^*(\theta|m^\circ)$ for each $\theta \in K(m^\circ)$.

Finally, we can choose a second-order prior ρ with $supp(\rho) = \{\pi^*, \pi_{m^\circ}^{**}\}_{m^\circ \in \mathcal{M}^\circ}$ such that

$$\{\pi^*\} := \arg\max_{\pi \in supp(\rho)} \rho(\pi) \text{ and } \{\pi_{m^\circ}^{**}\} := \arg\max_{\pi \in supp(\rho)} \rho_{m^\circ}(\pi) \text{ for each } m^\circ \in \mathcal{M}^\circ,$$

where π^* is the initial hypothesis, defined as, e.g., in the proof of Theorem 1. Thus, the RHTE

$(b_S^*, b_R^*, \rho, \mu_\rho)$, constructed above, supports the given PBE (b_S^*, b_R^*, μ^*); completing the proof. ■

Proof of Theorem 2. Consider a strategically stable PBE (b_S^*, b_R^*, μ^*) with $\mu^*(\cdot|m^\circ) \in \Delta(H(m^\circ))$ for each unsent message m°. Since $H(m^\circ) \neq \emptyset$, there exists $b_R^\bullet(\cdot|m^\circ) \in MBR(\Theta, m^\circ)$ such that

$$u^*(\theta) = \sum_{a\in\mathcal{A}} u_S(t, m^\circ, a) b_R^\bullet(a|m^\circ) \text{ for all } \theta \in H(m^\circ), \text{ and} \tag{76}$$

$$u^*(\theta) > \sum_{a\in\mathcal{A}} u_S(t, m^\circ, a) b_R^\bullet(a|m^\circ) \text{ for all } \theta \notin H(m^\circ). \tag{77}$$

When the Receiver replies to m° to according to $b_R^\bullet$, the Sender's types in $H(m^\circ)$ may voluntarily defect to m°. Each type $\theta \notin H(m^\circ)$ strictly prefers the equilibrium strategy $b_S^*(\cdot|\theta)$, yielding $u^*(\theta)$.

Given $\mu^* \in \Delta(H(m^\circ))$ and $b_R^\bullet(\cdot|m^\circ)$, we construct a rational hypothesis π_{m° as follows: First, we define a rational strategy b'_R for the Receiver. That is, for each message $m \in \mathcal{M}$,

$$\text{if } m \neq m^\circ, \quad \text{let } b'_R(a|m) = b_R^*(a|m) \quad \text{for each } a \in \mathcal{A}, \text{ and} \tag{78}$$

$$\text{if } m = m^\circ, \quad \text{let } b'_R(a|m) = b_R^\bullet(a|m) \quad \text{for each } a \in \mathcal{A}. \tag{79}$$

Since b_R^* and $b_R^\bullet(\cdot|m^\circ)$ are rational, so is b'_R. We now define a rational strategy b'_S for the Sender: For each $\theta \notin H(m^\circ)$, let

$$b'_S(m|\theta) = b_S^*(m|\theta) \text{ for each } m \in \mathcal{M}, \tag{80}$$

and for each $\theta \in H(m^\circ)$,

$$b'_S(m|\theta) = (1 - \alpha(\theta)) b_S^*(m|\theta) \text{ for } m \neq m^\circ \text{and } b'_S(m^\circ|\theta) = \alpha(\theta) \in [0, 1]. \tag{81}$$

Thus, b'_S rationalizes the voluntary defection to m° by each $\theta \in H(m^\circ)$ if $\alpha(\theta) > 0$ for all $\theta \in H(m^\circ)$. We allow that $\alpha(\theta) = 0$ for some (not all) $\theta \in H(m^\circ)$ for a reason that will be clear later.

Now, $\beta' = b'_S$ together with the prior p defines the rational hypothesis π_{m° given by

$$\pi_{m^\circ}(m, \theta) = \begin{cases} b_S^*(m|\theta)p(\theta), & \text{for each } (m, \theta) \in \mathcal{M} \times \Theta \setminus H(m^\circ) \\ \big(1 - \alpha(\theta)\big) b_S^*(m|\theta)p(\theta), & \text{for each } (m, \theta) \in \mathcal{M} \times H(m^\circ),\ m \neq m^\circ, \\ \alpha(\theta)p(\theta), & \text{for each } (m, \theta) \in \mathcal{M} \times H(m^\circ),\ m = m^\circ. \end{cases} \tag{82}$$

Hence, m° is feasible under π_{m° (i.e., $\pi_{m^\circ}(m^\circ, H(m^\circ)) > 0$) since $\alpha(\theta) > 0$ for some $\theta \in H(m^\circ)$. Let μ_ρ be the Bayesian update of π given m°. It remains to be shown that $\mu_\rho = \mu^*$.

Since $\mu^*(\cdot|m^\circ) \in \Delta(H(m^\circ))$, either $supp(\mu^*(\cdot|m^\circ)) = H(m^\circ)$ or $supp(\mu^*(\cdot|m^\circ)) \subset H(m^\circ)$.

In either case, we can determine a vector of weights $(\alpha(\theta))_{\theta\in H(m^\circ)} \in [0,1]^{H(m^\circ)}$ for which

$$\mu_\rho(\theta|m^\circ) := \frac{\pi(m^\circ,\theta)}{\pi(m^\circ,\Theta)} = \frac{\alpha(\theta)p(\theta)}{\sum_{\theta'\in H}\alpha(\theta')p(\theta')} = \mu^*(\theta|m^\circ) \text{ for all } \theta\in\Theta. \tag{83}$$

In particular, for each $\theta \in H(m^\circ)$, let $\alpha(\theta)$ be given by

$$\alpha(\theta) := \frac{\mu^*(\theta|m^\circ)\sum_{\theta'\in H(m^\circ)\setminus\{\theta\}}\alpha(\theta')p(\theta')}{(1-\mu^*(\theta|m^\circ))p(\theta)}. \tag{84}$$

Thus, we have shown that there is a rational hypothesis π_{m° that justifies $\mu^*(\cdot|m^\circ) = \mu_\rho(\cdot|m^\circ)$ where $\mu_\rho(\cdot|m^\circ) \in co\text{-}hull(\mu_\rho(\cdot|m^\circ), \Delta(H(m^\circ)))$ for each unsent message $m^\circ \in \mathcal{M}^\circ$.

Finally, we can choose a second-order prior ρ with $supp(\rho) = \{\pi^*, \pi^{**}_{m^\circ}\}_{m^\circ\in\mathcal{M}^\circ}$ such that

$$\{\pi^*\} := \underset{\pi\in supp(\rho)}{\arg\max}\,\rho(\pi) \text{ and } \{\pi^{**}_{m^\circ}\} := \underset{\pi\in supp(\rho)}{\arg\max}\,\rho_{m^\circ}(\pi) \text{ for each } m^\circ\in\mathcal{M}^\circ,$$

where π^* is the initial hypothesis defined as, e.g., in the proof of Theorem 1. Thus, the constructed RHTE $(b^*_S, b^*_R, \rho, \mu_\rho)$ constitutes the strategically stable PBE (b^*_S, b^*_R, μ^*); completing the proof. ■

To prove Proposition 5, we will invoke Lemma 3 which itself provides an interesting remark. It shows that $H(m^\circ)$ and $I(m^\circ)$ are related to each other in the following sense:

Lemma 3 *For each PBE and each unsent message m°, $H(m^\circ)\neq\emptyset$ if and only if $I(m^\circ)\neq\emptyset$.*

Proof of Lemma 3. Let (b^*_S, b^*_R, μ^*) be a PBE and m° an unsent message. Suppose $H(m^\circ)\neq\emptyset$. By definition of $H(m^\circ)$, for some $b^\bullet_R(\cdot|m^\circ)\in MBR(\Theta, m^\circ)$, we have

$$u^*(\theta) = \sum_{a\in\mathcal{A}} u_S(t,m^\circ,a)b^\bullet_R(a|m^\circ) \text{ for all } \theta\in H(m^\circ), \text{ and} \tag{85}$$

$$u^*(\theta) > \sum_{a\in\mathcal{A}} u_S(t,m^\circ,a)b^\bullet_R(a|m^\circ) \text{ for all } \theta\notin H(m^\circ). \tag{86}$$

Hence, if $\theta\in H(m^\circ)$ then $\theta\in I(m^\circ)$. Thus, $I(m^\circ)\neq\emptyset$.

Conversely, suppose $I(m^\circ)\neq\emptyset$. For the equilibrium strategy b^*_R, it holds true that

$$u^*(\theta) \geq \sum_{a\in\mathcal{A}} u_S(t,m^\circ,a)b^*_R(a|m^\circ) \text{ for all } \theta\in I(m^\circ), \text{ and} \tag{87}$$

$$u^*(\theta) > \sum_{a\in\mathcal{A}} u_S(t,m^\circ,a)b^*_R(a|m^\circ) \text{ for all } \theta\notin I(m^\circ). \tag{88}$$

If Equation (87) is binding, $H(m^\circ)\neq\emptyset$ is immediate. Assume that Equation (87) is not binding.

Since $I(m^\circ) \neq \emptyset$, there is some $b'_R(\cdot|m^\circ)$ such that

$$u^*(\theta) \geq \sum_{a\in\mathcal{A}} u_S(t, m^\circ, a) b'_R(a|m^\circ) \text{ for all } \theta \in I(m^\circ), \text{ and} \tag{89}$$

$$u^*(\theta) > \sum_{a\in\mathcal{A}} u_S(t, m^\circ, a) b'_R(a|m^\circ) \text{ for all } \theta \notin I(m^\circ). \tag{90}$$

The continuity of expected utility in probabilities ensures that $b^\bullet_R(\cdot|m^\circ)$ and a non-empty set $J(m^\circ) \subseteq I(m^\circ)$ exist such that

$$u^*(\theta) = \sum_{a\in\mathcal{A}} u_S(t, m^\circ, a) b^\bullet_R(a|m^\circ) \text{ for all } \theta \in J(m^\circ), \text{ and} \tag{91}$$

$$u^*(\theta) > \sum_{a\in\mathcal{A}} u_S(t, m^\circ, a) b^\bullet_R(a|m^\circ) \text{ for all } \theta \notin J(m^\circ). \tag{92}$$

Hence, $H(m^\circ) = J(m^\circ) \neq \emptyset$, completing the proof. ■

Proof of Proposition 5. Kohlberg and Mertens (1986) showed that a strategically stable PBE exists. If some PBE has $I(m^\circ) \neq \emptyset$ for each unsent message m°, then $H(m^\circ) \neq \emptyset$ by Lemma 3. Thus, by Theorem 2, there exists a rational equilibrium that is strategically stable. ■

Proof of Proposition 6. Let (b^*_S, b^*_R, μ^*) be a PBE that passes the $D1$ Criterion. By Monotonicity, all Sender's types have the same ranking over the Receiver's actions. Thus, each $\theta \in H(m^\circ)$ strictly benefits from the deviation to m° whenever each $\theta \notin H(m^\circ)$ weakly benefits from the deviation. Hence, each $D1$ equilibrium has μ^* such that $\mu^*(\theta|m^\circ) = 0$ for $\theta \notin H(m^\circ)$; i.e., $supp(\mu^*(\cdot|m^\circ)) \subseteq H(m^\circ)$. As shown in the proof of Theorem 2, all beliefs in $\Delta(H(m^\circ))$ can be justified by a rational hypothesis. For this reason, the $D1$ equilibrium is a rational equilibrium. ■

References

ASHEIM, G. B., AND A. PEREA (2005): "Sequential and quasi-perfect rationalizability in extensive games," *Games and Economic Behavior*, 53(1), 15–42.

AUMANN, R. J., AND A. BRANDENBURGER (1995): "Epistemic Conditions for Nash Equilibrium," *Econometrica*, 63, 1161–1180.

BANKS, J., C. CAMERER, AND D. PORTER (1994): "An Experimental Analysis of Nash Refinements in Signaling Games," *Games and Economic Behavior*, 6(1), 1–31.

BANKS, J. S. (1990): "A Model of Electoral Competition with Incomplete Information," *Journal of Economic Theory*, 50(2), 309–325.

BANKS, J. S., AND J. SOBEL (1987): "Equilibrium Selection in Signaling Games," *Econometrica*, 55(3), 647–661.

BATTIGALLI, P. (2006): "Rationalization in Signaling Games: Theory and Applications," *International Game Theory Review*, 08, 67–93.

BATTIGALLI, P., AND M. SINISCALCHI (2002): "Strong Belief and Forward Induction Reasoning," *Journal of Economic Theory*, 106(2), 356–391.

BERNHEIM, B. D. (1994): "A Theory of Conformity," *Journal of Political Economy*, 102(5), 841–877.

BHATTACHARYA, S. (1979): "Imperfect Information, Dividend Policy, and the Bird in the Hand Fallacy," *Bell Journal of Economics*, 10(1), 259–270.

BRANDTS, J., AND C. A. HOLT (1992): "An Experimental Test of Equilibrium Dominance in Signaling Games," *American Economic Review*, 82(5), 1350–1365.

BRANDTS, J., AND C. A. HOLT (1993): "Adjustment Patterns and Equilibrium Selection in Experimental Signaling Sames," *International Journal of Game Theory*, 22(3), 279–302.

CHO, I.-K. (1987): "A Refinement of Sequential Equilibrium," *Econometrica*, 55(6), 1367–1389.

CHO, I.-K., AND D. M. KREPS (1987): "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102(2), 179–221.

CHO, I.-K., AND J. SOBEL (1990): "Strategic Stability and Uniqueness in Signaling Games," *Journal of Economic Theory*, 50(2), 381–413.

ESŐ, P., AND J. SCHUMMER (2009): "Credible Deviations from Signaling Equilibria," *International Journal of Game Theory*, 38(3), 411–430.

FARRELL, J. (1993): "Meaning and Credibility in Cheap-Talk Games," *Games and Economic Behavior*, 5, 514–531.

FUDENBERG, D., AND K. HE (2020): "Payoff Information and Learning in Signaling Games," *Games and Economic Behavior*, 120, 96–120.

FUDENBERG, D., AND D. K. LEVINE (1993): "Self-Confirming Equilibrium," *Econometrica*, 61(3), 523–545.

FUDENBERG, D., AND J. TIROLE (1991a): *Game Theory*. MIT Press: Cambridge, Massachusetts.

FUDENBERG, D., AND J. TIROLE (1991b): "Perfect bayesian equilibrium and sequential equilibrium," *Journal of Economic Theory*, 53(2), 236260.

GAL-OR, E. (1989): "Warranties as a Signal of Quality," *Canadian Journal of Economics*, 22(1), 50–61.

GALPERTI, S. (2019): "Persuasion: The Art of Changing Worldviews," *American Economic Review*, 109(3), 996–1031.

JOHN, K., AND J. WILLIAMS (1985): "Dividends, Dilution, and Taxes: A Signalling Equilibrium," *Journal of Finance*, 40(4), 1053–1070.

KOHLBERG, E., AND J.-F. MERTENS (1986): "On the Strategic Stability of Equilibria," *Econometrica*, 54(5), 1003–1037.

KREPS, D. M., AND G. RAMEY (1987): "Structural Consistency, Consistency, and Sequential Rationality," *Econometrica*, 55(6), 1331–1348.

KREPS, D. M., AND R. WILSON (1982): "Sequential Equilibria," *Econometrica*, 50(4), 863–894.

KÜBLER, D., W. MÜLLER, AND H.-T. NORMANN (2008): "Job-Market Signaling and Screening: An Experimental Comparison," *Games and Economic Behavior*, 64(1), 219–236.

LIU, S., AND H. PEI (2020): "Monotone Equilibria in Signaling Games," *European Economic Review*, p. 103408.

LOHMANN, S. (1995): "Information, Access, and Contributions: A Signaling Model of Lobbying," *Public Choice*, 85(3-4), 267–284.

MAILATH, G. J., M. OKUNO-FUJIWARA, AND A. POSTLEWAITE (1993): "Belief-Based Refinements in Signalling Games," *Journal of Economic Theory*, 60(2), 241 – 276.

MILGROM, P., AND J. ROBERTS (1982): "Limit Pricing and Entry under Incomplete Information: An Equilibrium Analysis," *Econometrica*, 50(2), 443–459.

——— (1986): "Price and Advertising Signals of Product Quality," *Journal of Political Economy*, 94(4), 796–821.

MILLER, R. M., AND C. R. PLOTT (1985): "Product Quality Signaling in Experimental Markets," *Econometrica*, 53(4), 837–872.

NELSON, P. (1974): "Advertising as Information," *Journal of Political Economy*, 82(4), 729–754.

ORTOLEVA, P. (2012): "Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News," *American Economic Review*, 102(6), 2410–36.

RILEY, J. G. (1979): "Informational equilibrium," *Econometrica*, 47(2), 331–359.

SOBEL, J., L. STOLE, AND I. ZAPATER (1990): "Fixed-Equilibrium Rationalizability in Signaling Games," *Journal of Economic Theory*, 52, 304 – 331.

SPENCE, M. (1973): "Job Market Signaling," *Quarterly Journal of Economics*, 87(3), 355–374.

SUN, L. (2019): "Hypothesis Testing Equilibrium in Signalling Games," *Mathematical Social Sciences*, 100, 29–34.