Testing Behavioral Hypotheses in Signaling Games*

Adam Dominiak† and Dongwoo Lee‡

May 2, 2021

Abstract

In this paper, we introduce an equilibrium concept for signaling games, called Focused Hypothesis Testing Equilibrium (HTE). This equilibrium notion incorporates Ortoleva's (2012) theory of belief updating on zero-probability events via selecting and updating hypotheses. Hypotheses are beliefs about strategies. If an equilibrium message is observed, the hypothesis that the sender plays his equilibrium strategy is selected and updated by Bayes' rule. However, if an out-of-equilibrium message is observed, this hypothesis is rejected. Then, a new hypothesis about a strategy that generates the observed message is selected and updated via Bayes' rule. Each Focused HTE is a Perfect Bayesian Equilibrium (PBE). Conversely, we show that each PBE is a Focused HTE, providing a novel justification for PBE beliefs. Hypotheses provide a useful tool to reason about off-path beliefs. We impose different behavioral restrictions on hypotheses and use stronger equilibrium notions as refinement criteria for PBE. We compare our refinements with the Intuitive Criterion. Our strongest refinement concept justifies off-path beliefs that are immune to the Stiglitz-Mailath critique of the Intuitive Criterion.

Keywords: Signaling games, Perfect Bayesian Equilibrium, Hypothesis Testing Equilibrium, belief updating, Bayes' rule, maximum likelihood updating, out-of-equilibrium beliefs, refinements.
JEL Classification: C72, D81, D83

*The authors are very grateful to Hans Haller, George Mailath, Gerelt Tserenjigmid, Matthew Kovach, Kevin He, and the audiences of the 28th International Conference on Game Theory and the 88th Southern Economic Association (SEA) Meetings, and seminar participants at the Australian National University for their valuable comments and fruitful discussions. This paper is based on Dongwoo Lee's Doctoral Dissertation (Chapter 3) written at the Department of Economics at Virginia Tech.

†Virginia Tech, Department of Economics, 3122 Pamplin Hall, 880 West Campus Drive, Blacksburg, VA 24061, USA. Email: dominiak@vt.edu.

‡**Corresponding Author:** China Center for Behavioral Economics and Finance, Southwestern University of Finance and Economics, Chengdu, Sichuan 610074, China. Email: dwlee05@gmail.com.

1 Introduction

Signaling games are an important class of dynamic games with incomplete information. Signaling refers to interactive situations in which one party uses observable actions of an informed opponent to make inferences about hidden information. Signaling games have been applied to explain a variety of economic phenomena, including job search (Spence, 1973), advertising (Neslon, 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).[1]

A Perfect Bayesian Equilibrium (PBE) is a standard solution concept for signaling games. A Sender observes his type, and chooses an optimal message. A Receiver observes the message but not the type, forms her posterior about the Sender's types, and best responds with an action. As the name indicates, posteriors are derived via Bayes' rule, whenever possible. However, the standard updating procedure poses a serious limitation. Since Bayes' rule does not specify how beliefs are derived at information sets with zero probability, PBE admits arbitrary, multiple off-path beliefs.

In this paper, we suggest a solution concept that admits belief updating for *all* messages, including out-of-equilibrium messages. In a nutshell, beliefs are derived by selecting and updating hypotheses about strategic behavior. We will argue that hypotheses provide a useful framework for reasoning about off-path beliefs, and thus facilitate development of various refinement criteria.

Our first goal is to introduce a solution concept called Focused Hypothesis Testing Equilibrium (HTE). This equilibrium notion incorporates a novel theory of belief updating on zero-probability events, the Hypothesis Testing model axiomatized by Ortoleva (2012). The key element of this model is a set of hypotheses. For signaling games, we define *hypothesis* as a belief of the Receiver about strategic behavior of the opponent player, combined with the prior information about types. For each message, the Receiver selects a hypothesis about the Sender's behavior that, in her view, generates the message and updates it via Bayes' rule to derive her posterior over types.

To illustrate the main idea, consider a labor-market game à la Spence (1973). A worker applying for a job has either low (L) or high (H) skill. An employer does not know the applicant's skill but knows the prior probability distribution over skills, denoted by $p(H)$ and $p(L)$. The applicant decides whether to invest in education or not. Based on the applicant's decision, the employer forms her posterior about skills, and assigns him to either an executive job or a manual one.[2]

Suppose that the employer believes that each worker type acquires education, i.e., each type signals education with probability one. Such a belief, combined with the prior p, defines a hypothesis, i.e., a probability distribution over a product space of types and messages. In this case,

[1] Riley (2001) provides a comprehensive survey of economic applications of signaling games.

[2] The players' payoffs are depicted in Figure 1. This game will be analyzed in Section 2.

the (pooling) hypothesis assigns probability $p(L)$ to the state in which the low-skilled type acquires education occurs and probability $p(H)$ to the state in which the high-skilled type acquires education occurs.[3] According to this hypothesis, it is certain that the job applicant acquires education. Therefore, when education is observed, the employer can select this hypothesis to derive her posterior about skills, which by applying Bayes' rule, coincides with the prior distribution.

When the employer observes no education, she concludes that her initial hypothesis was incorrect. She then selects a new hypothesis that is consistent with the signal. For instance, suppose the employer believes that skills are perfectly positively correlated with the levels of education. This belief, combined with the prior p, induces the (separating) hypothesis that assigns probability $p(H)$ to the state in which the high-skilled type acquires education and probability $p(L)$ to the state in which the low-skilled type acquires no education. By updating this hypothesis via Bayes' rule, the employer infers that the job applicant signaling no education must be of the low-skilled type.

In Focused HTE, the Receiver holds a prior over a set of hypotheses (i.e., a second-order prior). Before any information is revealed, the Receiver chooses an initial hypothesis (i.e., the most likely hypothesis with respect to her second-order prior). This hypothesis is about the equilibrium strategy for the Sender. It is updated via Bayes' rule to derive posteriors on the equilibrium path. However, if an out-of-equilibrium message is observed, the initial hypothesis is rejected. Then, the Receiver updates her second-order prior via Bayes' rule, and selects a new hypothesis (i.e., the most likely hypothesis according to the updated second-order prior). The new hypothesis is updated via Bayes' rule to derive posteriors given the out-of-equilibrium message. This updating procedure provides a system of posteriors that are well-defined for all information sets.[4]

The idea to derive posteriors from hypotheses about strategic behavior of opponent players dates back to Kreps and Wilson (1982). They informally suggested "hypothesis testing" as a tool to justify beliefs of sequential equilibrium. In particular, they argued that off-path beliefs should be *structurally consistent*. That is, at each information set, each belief should be derived from a single strategy that governed the previous moves via Bayesian updating. In another work, Kreps and Ramey (1987, p.1332) provided the following interpretation of structural consistency:

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

[3] Strictly speaking, the probabilities for these states are $1 \cdot p(L)$ and $1 \cdot p(H)$, respectively.

[4] In Ortoleva's (2012) model, an agent rejects her initial hypothesis and selects a new one if, according to the initial hypothesis, the observed event has probability equal or smaller than a threshold $\epsilon \geq 0$. In our equilibrium concept, the threshold is zero. The acronym "Focused" indicates that the Receiver considers only hypotheses that she actually uses.

> *some single "alternative hypothesis" as to the strategy governing prior play, such that under the hypothesis the information set is reached with positive probability."*

Focused HTE is a solution concept that formally incorporates structural consistency into signaling games via the Hypothesis Testing model of Ortoleva (2012).[5] In fact, our structural consistency is stronger since we require that hypotheses are consistent with the prior information about types.

Our first result establishes the relationship between Focused HTE and PBE. We show that Focused HTE and PBE are equivalent solution concepts. In particular, for each off-path belief of a given PBE, there exists a hypothesis that induces the belief. This result proves that PBE beliefs are structurally consistent in the spirit of Kreps and Wilson (1982) for signaling games.[6]

Our second goal is to use hypotheses as a tool to refine PBEs (equivalently, Focused HTEs). To this end, we will strengthen our equilibrium concept by imposing behavioral restrictions on hypotheses. Our first restriction requires hypotheses to be (second-order) rational. A rational hypothesis is a belief about rational strategies for the Sender (i.e., strategies that best respond to rational strategies for the Receiver).[7] The equilibrium notion is called Rational HTE.

In Rational HTE, posteriors are consistent with the mutual knowledge of players' rationality. We use the Rational HTE that supports a given PBE as an argument in favor of the equilibrium. A PBE is said to pass the Rational Hypothesis Testing (HT) refinement if there exists a Rational HTE that supports the PBE; otherwise, the equilibrium fails the refinement. We will show that Rational HTE substantially reduces the plethora of PBEs in the classical signaling game of Spence (1973).

We compare our refinement with the well-known Intuitive Criterion introduced by Cho and Kreps (1987). This refinement criterion does not build on any theory of belief updating. Instead, the idea is to eliminate beliefs that assign (strictly) positive probabilities to types that cannot benefit from sending an out-of-equilibrium message. In general, the Intuitive Criterion and Rational HTE are not nested.[8] However, there is a family of intuitive PBEs that passes the Rational HT refinement.[9] If, for each out-of-equilibrium message of a given PBE, there is a single type that could benefit from sending the message, then the PBE passes the Rational HT refinement. In this case, the Rational HTE is an intuitive PBE where the Receiver learns the single type that could "potentially" deviate.

Our second behavioral restriction requires that rational hypotheses are behaviorally consistent.

[5] 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 decisions of the Sender.

[6] However, for more general extensive-form games, Kreps and Ramey (1987) showed that sequential-equilibrium beliefs do not need to be structurally consistent.

[7] In Focused HTE, the initial hypothesis is always rational, while new hypotheses do not need to be rational.

[8] That is, there is a PBE that passes the Intuitive Criterion but fails the Rational HT refinement and vice versa; there is a PBE that passes the Rational HT refinement but fails the Intuitive Criterion.

[9] We use the term *intuitive* PBE as a reference to a PBE that passes the Intuitive Criterion.

A new hypothesis is behaviorally consistent with the initial hypothesis if - after updating it along the equilibrium path - it rationalizes the same behavior as the initial hypothesis. The corresponding equilibrium notion is called Behaviorally Consistent HTE.

Our main rationale for Behaviorally Consistent HTE is the Stiglitz-Mailath critique of the Intuitive Criterion (see Cho and Kreps, 1987, p.203 and Mailath, 1988). According to their critique, the Intuitive Criterion can select beliefs that cause inconsistencies in reasoning about behaviors on and off the equilibrium paths. Behaviorally consistent hypotheses rule out such inconsistencies. Therefore, we use the Behaviorally Consistent HTE supporting a PBE as an additional refinement criterion, and derive a condition under which a Rational HTE is Behaviorally Consistent HTE.

Finally, we show that our strongest solution concept is consistent with empirical findings. Brandts and Holt (1992) challenged the Intuitive Criterion from an experimental perspective. They ran a series of experiments to test the Intuitive-Criterion predictions.[10] For instance, in one of their sessions, a majority of subjects played in line with a PBE that fails the Intuitive Criterion. We show that Behaviorally Consistent HTE can account for the results reported by Brandts and Holt (1992).

This paper is organized as follows. In Section 2, we recapitulate the standard PBE notion. In Section 3, we formalize Focused HTE, and derive the equivalence result. In Section 4, we define Rational HTE, and introduce our first refinement criterion. In Section 5, we compare Rational HTE with the Intuitive Criterion. In Section 6, we solve the educational signaling game of Spence (1973). In Section 7, we elucidate the Stiglitz-Mailath critique, and define our second refinement based on Behaviorally Consistent HTE. In Section 8, we explain the experimental findings in Brandts and Holt (1992). In Section 9, we provide final remarks. Appendix A collects all proofs.

2 Perfect Bayesian Equilibrium

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 p 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 behavior strategy for the Sender is $b_S := (b_S(\cdot|\theta))_{\theta\in\Theta}$, a collection of type-contingent mixtures over messages. That is, $\sum_{m\in\mathcal{M}} b_S(m|\theta) = 1$ for each $\theta \in \Theta$, where $b_S(m|\theta)$ denotes the probability that θ sends m. $\mathcal{B}_S = [\Delta(\mathcal{M})]^\Theta$ denotes the set of all strategies for the Sender. When b_S is degenerate (i.e., for each $\theta \in \Theta$, $b_S(m|\theta) = 1$ for some $m \in \mathcal{M}$), he follows a pure strategy.

[10] In this paper, the games implemented by Brandts and Holt (1992) are presented in Figure 1 and Figure 4.

A behavior strategy for the Receiver is $b_R := (b_R(\cdot|m))_{m\in\mathcal{M}}$, a collection of message-contingent mixtures over actions. That is, $\sum_{a\in\mathcal{A}} b_R(a|m) = 1$ for each $m \in \mathcal{M}$, where $b_R(a|m)$ denotes the probability that a is chosen in response to m. $\mathcal{B}_R = [\Delta(\mathcal{A})]^{\mathcal{M}}$ is the set of all such strategies. If b_R is degenerate (i.e., for each $m \in \mathcal{M}$, $b_R(a|m) = 1$ for some $a \in \mathcal{A}$), she plays a pure strategy.

For each message $m \in \mathcal{M}$, $\mu(\cdot|m)$ denotes a posterior belief over types given m (i.e., a probability distribution over Θ). A family of posteriors is denoted by $\mu := \{\mu(\cdot|m)\}_{m\in\mathcal{M}}$.

A Perfect Bayesian Equilibrium (PBE) is a triple (b_S^*, b_R^*, μ^*) under the following conditions.

Definition 1 (PBE) *(b_S^*, b_R^*, μ^*) is a Perfect Bayesian Equilibrium if:*

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

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

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

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

Conditions (i) and (ii) ensure sequential rationality. Condition (iii) specifies how posteriors are determined. For each message on the path, posteriors are derived by Bayes' rule. For each out-of-equilibrium message, posteriors are determined arbitrarily. We denote by $\mathcal{M}^\circ$ the set of messages off the path.

A PBE is in *pure* strategies when b_S^* and b_R^* are degenerate behavior strategies.

Example 1 Consider the signaling game depicted in Figure 1. A worker has either type θ_L (low skill) or type θ_H (high skill). Knowing his type, the worker decides whether to invest in education (E) or no education (N). Given the signal, an employer assigns the worker to either an executive job (e) or a manual job (m). The prior probabilities over types are given by $p(\theta_L) = 1/3$ and $p(\theta_H) = 2/3$. Note that each worker type prefers the executive job regardless of his education status. Moreover, education is more costly for the low-skilled worker. For the employer, education is not productive since her payoff is unaffected by the signal. Thus, the employer prefers to match type θ_H with the executive job and type θ_L with the manual job.

There are two families of pooling PBEs.

PBE-1: In the first family, each type signals education with probability 1, i.e.,

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

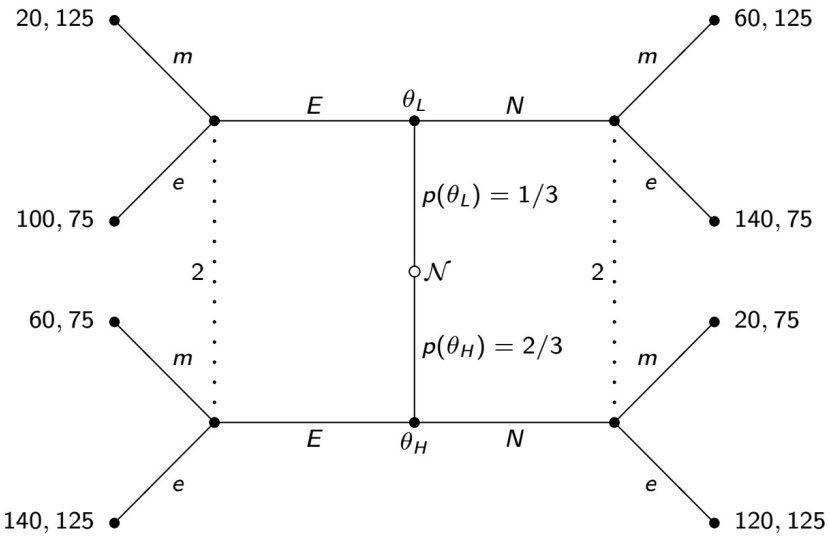

Figure 1: Labor-Market Game 1 from Brandts and Holt (1992).

PBE-2: In the second family, each type signals no education with probability 1, i.e.,

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

Note that there are multiple equilibria, due to the large amount of out-of-equilibrium beliefs.

In the next section, we introduce a solution concept that admits belief updating for all messages. The first goal is to introduce an equilibrium notion that allows us to explain all PBEs and off-path beliefs. Later, we will strengthen the equilibrium concept in order to reduce the number of PBEs.

3 Focused Hypothesis Testing Equilibrium

In this section, we introduce the notion of Focused Hypothesis Testing Equilibrium (HTE). We show that Focused HTE and PBE are equivalent solution concepts for the finite signaling games $\mathcal{G}$.

The main component of Focused HTE is a set of hypotheses. Hypotheses are beliefs about strategic behavior of the Sender. For each $\theta \in \Theta$, we denote by $\beta(\cdot|\theta)$ a type-contingent probability distribution over $\mathcal{M}$, representing the Receiver's belief about messages chosen by type θ. We denote by $\beta_R := (\beta_R(\cdot|\theta)_{\theta\in\Theta})$ a system of type-contingent beliefs. A system of beliefs (for short, belief) β_R combined with the prior information p about types defines a *hypothesis*.

Definition 2 (Hypothesis) *A* hypothesis π *is the probability distribution on* $\mathcal{M} \times \Theta$ *induced by beliefs* $\beta_R \in \mathcal{B}_S$ *and the prior probability distribution* $p \in \Delta(\Theta)$*; i.e., for every* $(m, \theta) \in \mathcal{M} \times \Theta$*:*

$$\pi(m, \theta) = \beta_R(m|\theta)p(\theta). \tag{1}$$

A hypothesis π ascribes probability $\pi(m,\theta)$ to the state: "type θ signals m." A hypothesis π is consistent with m if $\pi(m,\theta) > 0$, that is, $\beta_R(m|\theta)p(\theta) > 0$; the Receiver believes that her opponent plays a strategy according to which type θ signals m with a strictly positive probability.[11] Note that, by construction, each hypothesis is consistent with the prior information p. That is,

$$\pi(\mathcal{M},\theta) = \sum_{m\in\mathcal{M}} \pi(m,\theta) = \sum_{m\in\mathcal{M}} \beta_R(m|\theta)p(\theta) = p(\theta). \tag{2}$$

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

Example 2a In the signaling game of Figure 1, there are four simple hypotheses:

1) $\pi_1 := \{\pi_1(N,\theta_L) = 1/3, \pi_1(E,\theta_H) = 2/3\}$ where $\beta_R(N|\theta_L) = 1$ and $\beta_R(E|\theta_H) = 1$,

2) $\pi_2 := \{\pi_2(E,\theta_L) = 1/3, \pi_2(N,\theta_H) = 2/3\}$ where $\beta_R(E|\theta_L) = 1$ and $\beta_R(N|\theta_H) = 1$,

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

4) $\pi_4 := \{\pi_4(N,\theta_L) = 1/3, \pi_4(N,\theta_H) = 2/3\}$ where $\beta_R(N|\theta_L) = 1$ and $\beta_R(N|\theta_H) = 1$.

According to π_1, the employer believes that workers separate; the high-skilled worker signals E, while the low-skilled worker signals N. According to π_2, the employer believes that workers "reversely" separate; the high-skilled worker signals N, while the low-skilled worker signals E. According to π_3 (resp., π_4), the employer believes that the worker types pool on E (resp., on N).

The employer can believe that the low-skilled worker chooses N with probability 1, while the high-skilled type mixes between N and E with probabilities $\lambda \in (0,1)$ and $1-\lambda$, respectively. Such a belief, together with p, induces the following non-simple hypothesis:

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

To specify how the Receiver selects and updates hypotheses, we apply the Focused Hypothesis Testing model axiomatized by Ortoleva (2012). It is a theory of dynamic choice that admits belief updating on zero-probability events. Below, we elucidate how his theory works in our setup.

Consider 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., $|supp(\rho)| \in \mathbb{N}$).

[11] Recall that we assume $p(\theta) > 0$ for each $\theta \in \Theta$.

We assume that ρ induces a strict partial order over $supp(\rho)$. Before any information is revealed, the Receiver selects an initial hypothesis π^*. It is the most likely hypothesis with respect to ρ, i.e.,

$$\{\pi^*\} := \arg\max_{\pi \in supp(\rho)} \rho(\pi). \tag{3}$$

Upon arrival of a message m, the Receiver conducts a test. If π^* is consistent with m, she accepts π^*, and updates it via Bayes' rule. However, if π^* is inconsistent with m (i.e., $\pi^*(m, \Theta) = 0$), the Receiver rejects π^*. Then, she updates her second-order prior ρ, given m, via Bayes' rule, which is denoted by ρ_m. We assume that ρ_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.

$$\{\pi_m^{**}\} := \arg\max_{\pi \in supp(\rho)} \rho_m(\pi) \quad \text{where} \quad \rho_m(\pi) = \frac{\pi(m, \Theta)\rho(\pi)}{\sum_{\pi' \in supp(\rho)} \pi'(m, \Theta)\rho(\pi')}, \tag{4}$$

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 $\pi \in supp(\rho)$ that is consistent with m (i.e., $\pi(m, \Theta) > 0$).

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

$$supp(\rho) := \{\pi^*\} \cup \bigcup_{\substack{m \in \mathcal{M} \ s.t. \\ \pi^*(m,\Theta)=0}} \{\pi_m^{**}\}, \tag{5}$$

where π^* is the most likely hypothesis with respect to ρ and π_m^{**} is the most likely hypothesis with respect to ρ_m for a zero-probability message m according to π^*. This is the essence of Ortoleva's Focused Hypothesis Testing model, which we now incorporate into our solution concepts.[12]

A Focused HTE consists of a strategy profile (b_S^*, b_R^*), a focused second-order prior ρ, and a family of posteriors $\mu_\rho^* = \{\mu_\rho^*(\cdot|m)\}_{m \in \mathcal{M}}$ derived via the Hypothesis Testing model.

[12] We consider a special case of the Focused Hypothesis Testing model (see Ortoleva, 2012, Definition 4). In this general version, an agent rejects her initial belief if the conditioning 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 (see Ortoleva, 2012, Definition 3). In our setup, the initial hypothesis is rejected at information sets with zero probability (i.e., $\epsilon = 0$). Therefore, our equilibrium notion incorporates, strictly speaking, the Minimal Focused Hypothesis Testing model. One advantage of this model specification is that it admits a unique representation (see Ortoleva, 2012, Proposition 2).

Definition 3 (Focused HTE) $(b_S^*, b_R^*, \rho, \mu_\rho^*)$ *is a Focused Hypothesis Testing Equilibrium if:*

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

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

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

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

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

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

Conditions (i) and (ii) ensure sequential rationality. Conditions (iii) and (iv) ensure that posteriors are well-defined. For each message, the Receiver best replies with respect to the posterior derived from a hypothesis about the opponent's strategies that - in her view - generate the message.

On the equilibrium path (i.e., for each $m \in \mathcal{M}$ with $\pi^*(m,\Theta) > 0$), posteriors are derived from the initial hypothesis π^* according to which the Receiver believes that the Sender plays his equilibrium strategy (i.e., $\beta_R^* = b_S^*$). Off the equilibrium path (i.e., for each $m^\circ \in \mathcal{M}^\circ$ such that $\pi^*(m^\circ,\Theta) = 0$), π^* is rejected. Then, a new hypothesis $\pi_{m^\circ}^{**}$ is selected and updated conditional on m°. According to $\pi_{m^\circ}^{**}$, the Receiver believes that her opponent plays a (non-equilibrium) strategy that generates m° (i.e., $\beta_R(m^\circ|\theta) > 0$ for some $\theta \in \Theta$, implying that $\pi^{**}(m^\circ,\Theta) > 0$).

Remark 1 There are two reasons for assuming focused second-order priors. First, a focused ρ is what matters for behavior. An analyst can only observe the Receiver's beliefs after different messages (on-path and off-path). However, the analyst cannot empirically test whether the Receiver has a focused or unfocused second-order prior. Thus, if we start with an equilibrium with an unfocused second-order prior, one could construct another equilibrium with the same strategy profile and the same belief for each message, such that the new equilibrium has a focused second-order prior. Thus, it is legitimate to assume that the Receiver behaves *as if* she had a focused second-order prior.

Second, a focused ρ, which is a strict partial order, guarantees that beliefs associated with an equilibrium $(b_S^*, b_R^*, \rho, \mu_\rho^*)$ 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 $\rho'(\pi) < \rho'(\pi')$. Therefore, the system of posteriors associated with ρ' is the same as the system of posteriors associated with ρ; that is, $\{\mu_\rho(\cdot|m)\}_{m\in\mathcal{M}} = \{\mu_{\rho'}(\cdot|m)\}_{m\in\mathcal{M}}$. In this way, we avoid multiplicity of equilibria due to multiplicity of strict partial orders, i.e., ρs.

Below, we demonstrate two Focused HTEs for the labor-market game of Figure 1.

Example 2b There are two Focused HTEs with simple hypotheses presented in Example 2a.

FHTE-1: In the first Focused HTE, each type signals education, i.e.,

$$b_S^*(E|\theta_L) = b_S^*(E|\theta_H) = 1,\ b_R^*(e|E) = b_R^*(m|N) = 1,$$
$$supp(\rho) = \{\pi_1, \pi_3\} \text{ such that } \rho(\pi_1) < \rho(\pi_3),\ \ \mu_\rho^*(\theta_L|E) = 1/3 \ \text{ and } \ \mu_\rho^*(\theta_L|N) = 1.$$

Initially, the employer selects π_3 believing that both types choose E (i.e., $\pi^* = \pi_3$). Updating, π_3 given E, yields the prior distribution p. Off the path, the pooling hypothesis π_3 is rejected. The employer selects π_1 according to which she believes that workers separate (i.e., $\pi_N^{**} = \pi_1$). By updating π_1, the employer infers that N is chosen by the low-skilled type (i.e., $\mu_\rho^*(\theta_L|N) = 1$).

FHTE-2: In the second Focused HTE, each type signals no education, i.e.,

$$b_S^*(N|\theta_L) = b_S^*(N|\theta_H) = 1,\ b_R^*(m|E) = b_R^*(e|N) = 1,$$
$$supp(\rho) = \{\pi_2, \pi_4\} \text{ such that } \rho(\pi_2) < \rho(\pi_4),\ \ \mu_\rho^*(\theta_L|N) = 1/3 \ \text{ and } \ \mu_\rho^*(\theta_L|E) = 1.$$

When E is observed, the initial hypothesis π_4 is discarded and π_2 is selected (i.e., $\pi_E^{**} = \pi_2$). According to π_2, the employer believes that workers "reversely" separate. By updating π_2 conditional on E, the employer infers that E is chosen by the low-skilled worker (i.e., $\mu_\rho^*(\theta_L|E) = 1$).

Note that FHTE-1 and FHTE-2 coincide with the two pooling PBEs in which the employer infers the worker type from signals off the path (see PBE-1 and PBE-2). However, there are more Focused HTEs. As our first result proves, each PBE can be explained by a Focused HTE.

Theorem 1 *Let (b_S^*, b_R^*, μ^*) be a PBE and $\mathcal{M}^\circ$ be the set of out-of-equilibrium messages. Then, there exists a Focused HTE, $(b_S^*, b_R^*, \rho, \mu_\rho^*)$, that supports the PBE, i.e.,*

$$(i)\ \ \mu_\rho^*(\cdot|m) = \mu^*(\cdot|m) \textit{ for each equilibrium message } m, \textit{ and}$$
$$(ii)\ \ \mu_\rho^*(\cdot|m^\circ) = \mu^*(\cdot|m^\circ) \textit{ for each out-of-equilibrium message } m^\circ \in \mathcal{M}^\circ.$$

Theorem 1 provides an explanation for the origin of PBE beliefs. Each PBE belief can be derived from a hypothesis about strategic behavior of the Sender via Ortoleva's Hypothesis Testing model.

A few remarks are in order.

Remark 2 On the one hand, Theorem 1 is not surprising. We use mixed behavior (i.e., non-simple

hypotheses) to explain the plethora of PBE beliefs. On the other hand, non-simple hypotheses are necessary for the existence of a Focused HTE, even if pure strategies are played. If one assume simple hypotheses, a Focused HTE may not exist although a pure PBE exists (see Appendix C). In each pure Focused HTE, posteriors on the path are derived from a simple hypothesis according to which the Sender plays his (pure) equilibrium strategy. However, off-path beliefs are derived from hypotheses that do not need to be simple. In this case, the Receiver believes that out-of-equilibrium messages are outcomes of mixed strategies, although pure strategies are played in equilibrium. Non-simple hypotheses are necessary to justify each PBE by a Focused HTE.

Remark 3 Kreps and Wilson (1982) argued that off-path beliefs should be structurally consistent. A belief at an information set is structurally consistent if there exists a single behavior strategy under which the information set can be reached with a positive probability and from which the belief can be derived via Bayes' rule. Kreps and Ramey (1987) showed that sequential-equilibrium beliefs do not need to be structurally consistent in some extensive-form games. Theorem 1 shows that PBE beliefs are structurally consistent in signaling games. In fact, our structural consistency is stronger, as we require posteriors to be consistent with the prior information about types.

Remark 4 To model structural consistency, Kreps and Wilson (1982) informally suggested to use a sequence of hypotheses with a lexicographic order à la Blume, Brandenburger, and Dekel (1991). It is worth noting, however, that the lexicographic probability system (LPS) is very restrictive when it comes to modeling strategic behavior in signaling games. The reason is the following: In general, LPS lacks a well-defined procedure how to choose among the higher-order probabilities (hypotheses), unless all probability distributions have disjoint supports. In this special case, called the *lexicographic conditional probability system*, one could think of the following procedure: For each out-of-equilibrium message, the Receiver rejects the first-order (initial) hypothesis, and chooses a higher-order hypothesis whose support entails the message. However, hypotheses with non-overlapping supports are very restrictive behaviorally. For instance, in games with two types and two messages, as in Figure 1, if both types pool on one message in equilibrium, the only possible explanation for the out-of-equilibrium message is again pooling behavior, thus eliminating the families PBE-1 and PBE-2. To encompass broader behaviors, hypotheses need to overlap. However, in this case, it is not clear how to select among, possibly many, higher-order hypotheses that are consistent with the same out-of-equilibrium message.[13] In the Hypothesis Testing model, new hypotheses are uniquely selected in the maximum-likelihood fashion, allowing for hypotheses with overlapping supports. For this reason, the Hypothesis Testing model provides a convenient way to incorporate structural consistency into signaling games.

[13] In this case, it is common to select a probability distribution of the lowest order that assigns a strictly positive probability to the observed event. In signaling games, this procedure will again restrict the feasible off-path behaviors.

Finally, we note that Focused HTE and PBE are equivalent solution concepts. Since each Focused HTE is PBE by Definition 3, Theorem 1 implies the following corollary.

Corollary 1 *PBE and Focused HTE are equivalent solution concepts.*

In the next sections, we strengthen our solution concept by imposing additional conditions on hypotheses in order to reduce the number of PBEs (equivalently, the number of Focused HTEs).

4 Rational Hypothesis Testing Equilibrium

In this section, we introduce a stronger equilibrium notion, and suggest it as a refinement for PBE.

The idea is to derive out-of-equilibrium beliefs from hypotheses reflecting rational behavior. In Focused HTE, the initial hypothesis is about (second-order) rational behavior. According to this hypothesis, the Sender plays his (equilibrium) strategy that best responds to the Receiver's (equilibrium) strategy. However, new hypotheses do not need to reflect best-responding behavior. Since strategies that generate off-path behavior must differ from the Sender's equilibrium strategy, some of them may be irrational. To refine PBE, we eliminate beliefs that can only be derived from irrational strategies. That is, we require hypotheses to be about rational strategies.

A behavior strategy $b_R \in \mathcal{B}_R$ for the Receiver 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; that is,

$$b_R(a|m) > 0 \text{ implies } a \in BR(\Theta, m) := \bigcup_{\mu(\cdot|m)\in\Delta(\Theta)} BR(\mu, m).^{14} \tag{6}$$

Denote by $\mathcal{B}_R^\bullet$ the set of (first-order) rational strategies for the Receiver.

A behavior strategy $b_S \in \mathcal{B}_S$ for the Sender is (second-order) rational if it is a best response to some strategy $b_R \in \mathcal{B}_R^\bullet$ of the Receiver; that is, for each $\theta \in \Theta$ and $m \in \mathcal{M}$,

$$b_S(m|\theta) > 0 \text{ implies } m \in \arg\max_{m'\in\mathcal{M}} \sum_{a\in\mathcal{A}} u_S(\theta, m', a)\, b_R(a|m'). \tag{7}$$

Denote by $\mathcal{B}_S^\bullet$ the set of (second-order) rational strategies for the Sender.

A message m is rational if there is a strategy $b_S \in \mathcal{B}_S^\bullet$ such that $b_S(m|\theta) > 0$ for some $\theta \in \Theta$. A message $m^d \in \mathcal{M}^d$ is (strictly) dominated if for each $\theta \in \Theta$, choosing m^d is a never-best response (i.e., for each $\theta \in \Theta$, any strategy $b_S \in \mathcal{B}_S$ such that $b_S(m^d|\theta) > 0$ is a never-best response.)

A *rational hypothesis* is a belief about strategies for the Sender that best respond to some rational strategies of the Receiver combined with the prior information about types. We slightly

[14] $BR(\mu, m) := \arg\max_{a\in\mathcal{A}} \sum_{\theta\in\Theta} \mu(\theta|m) u_R(\theta, m, a)$.

abuse our notation, and denote by $\bar{\beta}_R := (\bar{\beta}_R(\cdot|\theta))_{\theta\in\Theta} \in \mathcal{B}_S^\bullet$ a system of type-contingent beliefs of the Receiver about rational choices by the Sender. A system of beliefs $\bar{\beta}_R \in \mathcal{B}_S^\bullet$, together with the prior probability distribution p on Θ, defines a rational hypothesis as follows.

Definition 4 (Rational Hypothesis) *A rational hypothesis is the probability distribution π on $\mathcal{M}\times\Theta$ induced by beliefs $\beta_R \in \mathcal{B}_S^\bullet$ and the prior information p on Θ; i.e., for every $(m,\theta) \in \mathcal{M}\times\Theta$,*

$$\pi(m,\theta) = \bar{\beta}_R(m|\theta)p(\theta). \tag{8}$$

A rational hypothesis π is called *simple* if $\bar{\beta}_R$ is a system of degenerate beliefs (i.e., for each $\theta \in \Theta$, $\bar{\beta}_R(m|\theta) = 1$ for some $m \in \mathcal{M}$). According to a simple-rational hypothesis, the Receiver believes that her opponent chooses a pure strategy that best responds to some of her rational strategies.

A Focused HTE with $supp(\rho)$ that contains only rational hypotheses is called *Rational HTE.*

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

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

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

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

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

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

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

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

In Rational HTE, posteriors are derived from rational hypotheses. On the equilibrium path, posteriors are derived from the initial hypothesis, according to which the Receiver believes her opponent plays his equilibrium strategy b_S^*. Since b_S^* best responds to b_R^*, the initial hypothesis is rational. Off the equilibrium path, each posterior is derived from an alternative hypothesis, according to which

the Receiver believes that the Sender follows a (non-equilibrium) strategy that best responds to some of her rational strategies. However, there might be a very costly message that none of the types will ever play. Thus, there will be no rational strategy generating the message. Since strictly dominated messages do not affect equilibrium behavior, we allow for arbitrary beliefs for such messages in order to make Rational HTE applicable for a larger class of signaling games.

Rational HTE is consistent with the players' mutual knowledge of rationality, given that $\mathcal{M}^d = \emptyset$. That is, each player is rational, and knows that the opponent is rational. This implies that each player believes that the opponent follows strategies that best reply to (first-order) rational strategies. That is, the Receiver believes that each message is chosen as a best reply to some of her rational strategies. Rational HTE is a more stringent solution concept than PBE. For this reason, we will use the Rational HTE supporting a given PBE as an argument in favor of the equilibrium.[15]

There are two important remarks.

Remark 5 In signaling games for which all strategies for the Sender are rational, i.e., $\mathcal{B}_S = \mathcal{B}_S^\bullet$, a Rational HTE always exists. Moreover, the set of Rational HTEs coincides with the set of PBEs.[16] However, if $\mathcal{B}_R^\bullet \subset \mathcal{B}_R$, a Rational HTE may not exist (see Appendix C)) and if it does, the set of Rational HTEs is a subset of PBEs, thus refining the multiple PBEs.[17]

Remark 6 Ortoleva (2012) suggested an equilibrium called Hypothesis Testing Equilibrium. In line with our equilibrium concept, he assumed that the initial hypothesis is rejected when a zero-probability message arrives (i.e., $\epsilon = 0$). Besides this, there are two substantial differences to our model. First, he considered equilibria in pure strategies, and assumed simple hypotheses. We allow for mixed equilibrium behavior and non-simple hypotheses. Second, and more importantly, Ortoleva's hypothesis notion is weaker than ours, as it is about first-order rational behavior. That is, his hypothesis is about a pure strategy for the Sender that best responds to *some* - not necessarily rational - pure strategy of the Receiver (see Ortoleva, 2012, Section IV). Hence, off-path beliefs may be justified by strategies that only best respond to never-best responses. We eliminate such beliefs by requiring second-order rationality.[18] For this reason, all formal results of the following sections apply to Ortoleva's Hypothesis Testing Equilibrium. Recently, Sun (2019) extended Ortoleva's equilibrium notion by allowing for $\epsilon \geq 0$, and showed that a strictly positive threshold

[15] One could require higher-order rationality to construct hypotheses. The resulting equilibrium would provide a stronger refinement criterion. However, many refinement concepts in the economic literature such as the Intuitive Criterion of Cho and Kreps (1987) and the novel Rationality Compatible Equilibrium of Fudenberg and He (2020) implicitly assume (second-order) rationality. Our goal is to compare Rational HTE with the Intuitive Criterion (see Section 5). Moreover, we are unaware of any interesting examples where the additional level of rationality would play a role. Therefore, we consider second-order rationality. Readers interested in a solution concept for signaling games that rely on rationalizability are referred to Sobel, Stole, and Zapater (1990) and Battigalli (2006).

[16] This fact follows from the proof of Theorem 1 if one replaces $\mathcal{B}_S$ in the proof by the set of rational strategies $\mathcal{B}_S^\bullet$.

[17] In Appendix D, we provide conditions for existence of a Rational HTE for finite (monotone) signaling games.

[18] In Section 6, we show that Rational HTE and Ortoleva's HTE yield different predictions for the Spence game.

may lead to behavior that is inconsistent with the sequential equilibrium of Cho and Kreps (1987).

A PBE is said to pass the Rational Hypothesis Testing (HT) refinement if there exists a Rational HTE supporting the PBE; otherwise the equilibrium fails the refinement.

Definition 6 (Rational HT Refinement) *A PBE, (b_S^*, b_R^*, μ^*), passes the Rational Hypothesis Testing refinement if there exists a Rational HTE, $(b_S^*, b_R^*, \rho, \mu_\rho^*)$ with $\mu_\rho^* := \{\mu_\rho^*(\cdot|m)\}_{m\in\mathcal{M}}$, such that*

$$\mu_\rho^*(\cdot|m) = \mu^*(\cdot|m) \text{ for all } m \in \mathcal{M}.$$

If m is strictly dominated, then any probability distribution $\mu^*(\cdot|m) \in \Delta(\Theta)$ is admissible.

For a given PBE, (b_S^*, b_R^*, μ^*), the refinement-algorithm operates in two steps. In the first step, for each message $m \in \mathcal{M}$, we examine if there exists a rational hypothesis that is consistent with m. Denote by Π_m the set of rational hypotheses consistent with m (i.e. $\pi(m, \Theta) > 0$). If $\Pi_m = \emptyset$, because m is strictly dominated, all beliefs are admissible. If $\Pi_m \neq \emptyset$, the second step applies. In this step, we examine if there is a hypothesis in Π_m that induces the PBE belief, $\mu^*(\cdot|m)$; that is,

$$\text{for some } \pi \in \Pi_m, \frac{\pi(m,\theta)}{\pi(m,\Theta)} = \mu^*(\theta|m) \text{ for each } \theta \in \Theta.$$

Accordingly, a PBE might fail the Rational HT refinement for two reasons. First, there is an out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$ for which $\Pi_{m^\circ} = \emptyset$ and m° is not strictly dominated. Second, $\Pi_{m^\circ} \neq \emptyset$. However, none of the rational hypotheses induces the PBE belief (i.e., for all $\pi \in \Pi_{m^\circ}$, $\mu_\rho(\cdot|m^\circ) \neq \mu^*(\cdot|m^\circ)$). This means that the PBE belief is inconsistent with the prior information p about types under any rational strategy for the Sender that generates m°.[19]

The following example shows that PBE-1, but not PBE-2, passes the Rational HT refinement.

Example 3a Consider PBE-1 for the labor-market game depicted in Figure 1 (see Example 1). For any $\lambda \in [1/2, 1]$, consider the rational strategy $b_S(\lambda)$ for the worker, defined by

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

Note that $b_S(\lambda)$ best responds to $b_R(m|E) = 1/4$, $b_R(e|E) = 3/4$, and $b_R(e|N) = 1$. The employer's belief $\bar{\beta}_R = b_S(\lambda)$, together with p, induces the following rational hypothesis:

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

[19] Example 3b shows that PBE-2 fails the Rational HT refinement for this reason.

Since π_λ is consistent with the out-of-equilibrium message N, by updating π_λ, we obtain

$$\mu^*_\rho(\theta_L|N) = \frac{1}{3-2\lambda} \geq \frac{1}{2} \text{ for each } \lambda \in [1/2, 1].$$

Therefore, for each $\lambda \in [1/2, 1]$, we have the Rational HTE with pooling on E:

$$b^*_S(E|\theta_L) = b^*_S(E|\theta_H) = 1,\ b^*_R(e|E) = b^*_R(m|N) = 1,$$
$$supp(\rho) = \{\pi_\lambda, \pi_3\} \text{ such that } \rho(\pi_\lambda) < \rho(\pi_3),\ \mu^*_\rho(\theta_L|E) = 1/3 \text{ and } \mu^*_\rho(\theta_L|N) = 1/(3-2\lambda),$$

showing that the whole family PBE-1 passes the Rational HT refinement.

Example 3b Consider PBE-2. For $\gamma \in [0, 1]$, consider the rational strategy $b_S(\gamma)$ for the worker:

$$b_S(E|\theta_L) = 1-\gamma,\ b_S(N|\theta_L) = \gamma \in [0, 1], \text{ and } b_S(E|\theta_H) = 1.$$

Note that b_S is rational, as it best responds to $b_R(e|E) = 1$, $b_R(m|N) = 1/2$ and $b_R(e|N) = 1/2$. The employer's belief $\bar{\beta}'_R = b_S(\gamma)$, together with p induces, the following rational hypothesis:

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

By updating π_γ given E, we obtain

$$\mu_\rho(\theta_L|E) = \frac{\pi_\gamma(E, \theta_L)}{\pi_\gamma(E, \Theta)} = \frac{(1-\gamma)\frac{1}{3}}{1-\gamma\frac{1}{3}} \leq \frac{1}{3} \text{ for each } \gamma \in [0, 1].$$

The employer infers that E is more likely to be chosen by the high-skilled worker, and assigns this worker to the executive job e instead of m. Thus, PBE-2 fails the Rational HT refinement.

In this section, we have argued that one can apply Rational HTE as a refinement criterion for PBE. In the next section, we compare Rational HTE with the Intuitive Criterion.

5 Rational HTE versus the Intuitive Criterion

In this section, we compare Rational HTE with the Intuitive Criterion of Cho and Kreps (1987).

Contrary to our approach, the Intuitive Criterion does not build on a theory of belief updating. Instead, the refinement is payoff-based. The idea is to eliminate off-path beliefs that assign a positive probability to types that do not have any incentive to deviate from their equilibrium strategy.

Let us briefly recall the refinement operates. For a PBE, (b^*_S, b^*_S, μ^*), denote by $u^*_S(\theta)$ the

expected equilibrium payoff for type θ. For an out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$, let

$$BR(\Theta, m^\circ) := \bigcup_{\mu(\cdot|m^\circ)\in\Delta(\Theta)} BR(\mu, m^\circ) \tag{9}$$

be the set of best responses for the Receiver to m° with respect to beliefs over Θ.[20] Let $T(m^\circ) \subseteq \Theta$ be the set of types that cannot improve upon their equilibrium payoff by playing m°, no matter how the Receiver responds. That is, for each type $\theta \in T(m^\circ)$ and all $a \in BR(\Theta, m^\circ)$:

$$u_S^*(\theta) > u_S(\theta, m^\circ, a). \tag{10}$$

Each type in $I(m^\circ) := \Theta \setminus T(m^\circ)$ could be better off than his equilibrium payoff by playing m°. A PBE fails the Intuitive Criterion if there exists a type that would be better off by choosing m°, presupposed the Receiver best responds to m° with respect to beliefs that assign a zero probability to each type that has no incentive to deviate; otherwise the PBE survives the refinement criterion.

Definition 7 (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 $a \in BR(I(m^\circ), m^\circ)$,*

$$u_S^*(\theta) < u_S(\theta, m^\circ, a), \tag{11}$$

where $BR(I(m^\circ), m^\circ)$ is the set of best responses induced by the beliefs concentrated on $I(m^\circ)$.

If a PBE passes the Intuitive Criterion, for each type, there is a best reply for the Receiver with respect to a posterior over $I(m^\circ)$, which makes all types better off by playing their equilibrium strategies (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). For a singleton set $I(m^\circ) = \{\theta_{m^\circ}\}$, the Receiver learns the single type that could benefit by playing m°. In this case, the Intuitive Criterion outcome is the PBE with $\mu(\theta_{m^\circ}|m^\circ) = 1$ for each $m^\circ \in \mathcal{M}^\circ$. We call a PBE that passes (resp. fails) the intuitive condition the *intuitive* (resp. *unintuitive*) PBE.

The example below illustrates how the Intuitive Criterion works for the game in Figure 1.

Example 4 Consider the family of PBEs with pooling on E (i.e., PBE-1 in Example 1). Given the equilibrium payoff, the low-skilled type 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 could be played only by the low-skilled worker, there is no type that has an incentive to play N. The pooling PBE with the out-of-equilibrium belief $\mu^*(\theta_L|N) = 1$ passes the Intuitive Criterion.[21]

[20] $BR(\mu, m^\circ) := \arg\max_{a\in\mathcal{A}} \sum_{\theta\in\Theta} u_R(\theta, m^\circ, a)\mu(\theta|m^\circ)$.

[21] PBE-2 fails the Intuitive Criterion. Only the high-skilled type could benefit by playing the out-of-equilibrium message E. Thus, the employer learns that E is chosen by type θ_H (i.e., $\mu^*(\theta_H|E) = 1$), and best responds with e. This, however, makes the high-skilled worker signal E instead of N.

Recall that, in Example 3a, we showed that this PBE passes the Rational HT refinement.

In general, the refinements based on Rational HTE and on the Intuitive Criterion are not nested. That is, there is a PBE that passes the Rational HT refinement but fails the Intuitive Criterion, and *vice versa*; there is a PBE that passes the Intuitive Criterion but fails the Rational HT refinement.

Proposition 1 *Rational HTE and Intuitive PBE are not nested.*

However, there are signaling games for which each intuitive PBE is a Rational HTE. More specifically, if for each out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$ of an intuitive PBE, there exists a single type that could benefit from choosing m°, then there exists a Rational HTE supporting the PBE.[22] In this case, the Rational HTE justifies the same off-path beliefs as the Intuitive Criterion.

Theorem 2 *Let (b_S^*, b_R^*, μ^*) be a PBE that passes the Intuitive Criterion. If for each out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$, $I(m^\circ)$ is a singleton (i.e., $I(m^\circ) = \{\theta_{m^\circ}\}$ for some $\theta_{m^\circ} \in \Theta$), then there exists a Rational HTE $(b_S^*, b_R^*, \rho, \mu_\rho^*)$ that supports the PBE where for each $m^\circ \in \mathcal{M}^\circ$,*

$$\mu^*(\theta_{m^\circ}|m^\circ) = \mu_\rho^*(\theta_{m^\circ}|m^\circ) = 1. \tag{12}$$

The single-type condition guarantees that there exists a pure strategy $b_S \in \mathcal{B}_S^\bullet$ that generates the out-of-equilibrium message m° (i.e., $b_S(m^\circ|\theta_{m^\circ}) = 1$). The Receiver's belief that her opponent follows this strategy (i.e., $\bar{\beta}_R = b_S$), combined with the prior p, induces a rational hypothesis π that is consistent with m° (i.e, $\pi(m^\circ, \theta^\circ) = p(\theta^\circ) > 0$). By updating π_{m°, given m°, the Receiver thus learns θ_{m°, yielding the off-path belief admitted by the Intuitive Criterion (i.e., $\mu^*(\theta_{m^\circ}|m^\circ) = 1$).

The single-type condition implies that the Intuitive-Criterion outcome is unique. However, there may be more than one Rational HTE supporting an intuitive PBE. To guarantee uniqueness, an auxiliary condition is needed. If, in addition to the single-type condition, each out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$ is a never-best response for each type that cannot be better off by playing m° than his equilibrium payoff (i.e., each $\theta \in T(m^\circ)$), then there is a unique Rational HTE supporting the intuitive PBE. The uniqueness condition is stated in the following corollary.

Corollary 2 *Consider an intuitive PBE with $I(m^\circ)$ being a singleton for each message $m^\circ \in \mathcal{M}^\circ$. If, for each out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$, it is true that*

$$m^\circ \notin \arg\max_{m \in \mathcal{M}} \sum_{a \in \mathcal{A}} u_S(\theta, m, a) b_R(a|m), \tag{13}$$

for any $b_R \in \mathcal{B}_R^\bullet$ and each $\theta \in T(m^\circ)$, then the Rational HT refinement outcome is unique.

[22] The converse of Theorem 2 does not work (see PBE-II derived in Section 7). PBE-II satisfies the single-type condition of Theorem 2. However, while PBE-II passes the Rational HT refinement, it fails the Intuitive Criterion.

The Intuitive Criterion has the same limitation as the PBE if there are more than one types that have an incentive to deviate for some message m° (i.e., $|I(m^\circ)| \geq 1$). In this case, the Intuitive Criterion admits arbitrary beliefs over $I(m^\circ)$. When $I(m^\circ) = \Theta$, the Intuitive Criterion does not reduce the set of beliefs at all, while the Rational HT refinement might reduce it (see Appendix B). For this reason, a variety of other refinement concepts has been suggested in the economic literature (e.g., Cho, 1987; Banks and Sobel, 1987; Mailath, Okuno-Fujiwara, and Postlewaite, 1993; Eső and Schummer, 2009; Fudenberg and He, 2018, 2020). A comprehensive comparison between other popular refinements and Rational HTE would go beyond the scope of this paper.

6 Educational Signaling Game

In this section, we solve the classic signaling game of Spence (1973), which is known to have a continuum of PBEs. We show that Rational HTE can substantially reduce the number of PBEs.

We consider a finite version of the Spence model. As in previous games, 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 $\theta_L < \theta_H$). The prior probability distribution p on Θ is $p(\theta_L) = 1-\alpha$ and $p(\theta_H) = \alpha \in (0,1)$.

The worker knows his type θ, and chooses an education level e from $\mathcal{M} = \{e_0, e_1, \ldots, e_N\}$, where $e_0 := 0 < e_1 < \ldots < e_N$.[23] The worker's payoff is given by

$$u_S(\theta, e, w) = w - \frac{e}{\theta} \quad \text{for } \theta \in \Theta, \tag{14}$$

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 2 depicts type-dependent indifference curves (the red one for θ_L and the blue one for θ_H).

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

$$u_R(\theta, e, w) = -(\theta - w)^2 \quad \text{for } \theta \in \Theta. \tag{15}$$

The rational employer offers a wage that is equal to expected productivity. That is, the best response for 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 her posterior belief over θ.[24] Note that $\mathcal{A} = \mathbb{R}_+$. However, $w(e) \in [\theta_L, \theta_H]$. We denote

[23] To simplify our analysis, we assume that $\mathcal{M}$ is sufficiently large, yet finite. In 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 $\mathcal{M} = \mathbb{R}_+$, the reader is referred to Fudenberg and Tirole (1991, Chapter 8, p.329).

[24] One can justify $w(e) = \mathbb{E}(\theta|e)$ by considering a perfect competition in a market with many rational employers. Jeong (2019) studies imperfect competition among employers in the context of job market signaling. In particular, he investigates how wage offers change, depending on the degree of competition.

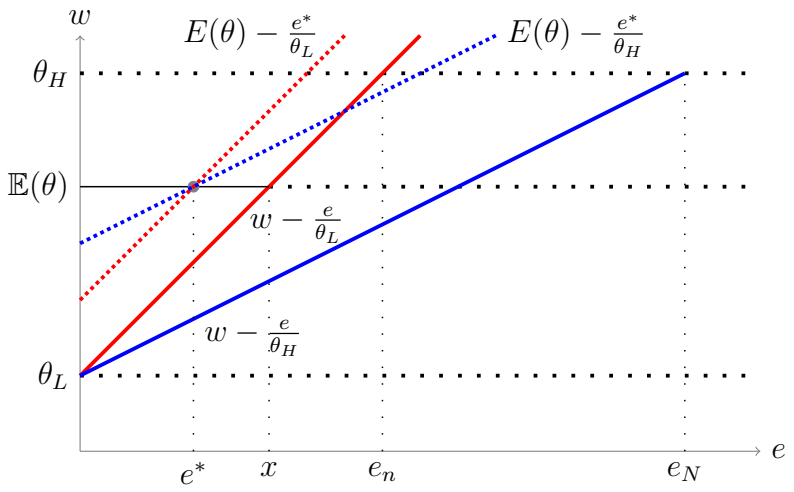

Figure 2: Pooling PBE with education level e^* and wage $w^* = \mathbb{E}(\theta)$.

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 use $w(e)$ as a short form to denote the pure strategy $b_R(w(e)|e) = 1$.

6.1 Pooling PBE

In a pooling equilibrium, both worker types choose the same education level e^*. Note that the payoff of the low-productivity type from signaling e^* at the average-productivity $w^* = \mathbb{E}(\theta)$ must be greater than his payoff from signaling no education at the minimum wage $w = \theta_L$. Formally, e^* must satisfy

$$\mathbb{E}(\theta) - \frac{e^*}{\theta_L} \geq \theta_L, \quad \text{or equivalently,} \quad e^* \leq \underbrace{\alpha(\theta_H - \theta_L)\theta_L}_{:=x} \tag{16}$$

specifying an upper bound x for education levels that can be explained 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

$$\mathbb{E}(\theta) - \frac{e^*}{\theta_L} \geq w(e) - \frac{e}{\theta_L}, \tag{17}$$

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

$$\mathbb{E}(\theta) - \frac{e^*}{\theta_H} \geq w(e') - \frac{e'}{\theta_H}. \tag{18}$$

Hence, there is a plethora of pooling PBEs. Note that there are multiple education levels that can be supported by a pooling equilibrium. Furthermore, each pooling PBE admits various wage schemes

due to arbitrary off-path beliefs. We will focus on the following family of pooling PBEs:

Observation 1 *For each education level e such that $0 \leq e^* \leq x := \alpha(\theta_H - \theta_L)\theta_L$, the strategy profile (b_S^*, w_R^*) and beliefs $\mu^* := (\mu^*(\cdot|e)_{e\in\mathcal{M}})$ such that*

(i) $b_S^*(e^*|\theta_L) = b_S^*(e^*|\theta_H) = 1$,

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

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

In each PBE, the employer believes that any education below the equilibrium level e^* is chosen by the low-productivity type, and the employer offers the minimum wage. Any education equal to, or larger than e^* but below e_n, does not convey any information about types. Thus, the employer pays the average productivity. The employer believes that any education above e_n is chosen by the high-productivity type with a probability ranging from zero to one, making any wage ranging from θ_L to θ_H admissible.[25] The level of education e_n makes the low-productivity type indifferent between choosing no education for $w = \theta_L$ and choosing e_n for $w = \theta_H$ (see Figure 2).

Note that for each pooling PBE, there is an out-of-equilibrium message that only the high-productivity type could be better off than his equilibrium payoff. If the employer believes that θ_H chooses such a message, she will offer the highest wage and then type θ_H will indeed deviate from the equilibrium. For this reason, each pooling PBE fails the Intuitive Criterion in the Spence game.

Are there pooling PBEs that can be justified by a Rational HTE? There are indeed. In the next section, we will derive the set of education levels that can be supported by a pooling Rational HTE. Moreover, we will show that the "size" of the set depends on the prior information about types.

6.2 Pooling Rational HTE

Let us first elucidate why a given PBE may fail the Rational HT refinement. Consider an out-of-equilibrium education level $e \in \{e_n, \ldots, e_N\}$.[26] Note that the low-productivity type has no incentive to choose such e even if he were be paid the highest wage $w(e) = \theta_H$. For each $e \in \{e_n, \ldots, e_N\}$, the payoff of the low-productivity type is lower than his payoff for no education at

[25] For the sake of simplicity, we consider a family of PBEs that covers all levels of education on which the workers can pool and all possible wages paid for $e \geq e_n$. Note that the family does not cover all admissible wages for $e \leq e_n$, such that $e \neq e^*$ (e.g., for each e such that $e^* \leq e < e_n$, $w(e)$ such that $\mathbb{E}(\theta) \leq w(e) \leq \theta_L$ is admissible).

[26] The lower and upper bounds of this set are $e_n := \theta_L(\theta_H - \theta_L)$ and $e_N := \theta_H(\theta_H - \theta_L)$.

the minimum wage $w(e) = \theta_L$. Thus, any strategy where type θ_L chooses $e \in \{e_n, \ldots, e_N\}$ is a never-best response.[27] This means that there does not exist a rational hypothesis according to which θ_L chooses any education level $e \in \{e_n, \ldots, e_N\}$ with a positive probability. Thus, each pooling PBE with $\mu(\theta_L|e) > 0$ for some $e \in \{e_n, \ldots, e_N\}$ fails the Rational HT refinement.

What levels of education can be explained as a pooling Rational HTE? Let us consider simple hypotheses. Since only the high-productivity type can choose $e \in \{e_n, \ldots, e_N\}$ as a best response, $\mu(\theta_H|e) = 1$ is the only off-path belief that can be justified by a rational hypothesis. This means that 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 $e \in \{e_n, \ldots, e_N\}$ at the highest wage θ_H. That is, e^* must satisfy

$$\mathbb{E}(\theta) - \frac{e^*}{\theta_H} \geq \theta_H - \frac{e_n}{\theta_H}, \quad \text{or equivalently,} \quad e^* \leq \underbrace{(\theta_H - \theta_L)(\theta_L - (1-\alpha)\theta_H)}_{:=y}, \tag{19}$$

which specifies an upper bound y for education that can be supported by a pooling Rational HTE. Note that a Rational HTE exists if and only if $y \geq 0$, or equivalently, $\alpha \geq \frac{\theta_H - \theta_L}{\theta_H}$. In other words, the number of PBEs passing the Rational HT refinement is a function of α, the fraction of high-productivity types in the market. The larger α, the more pooling Rational HTEs exist. If $\alpha < \frac{\theta_H - \theta_L}{\theta_H}$, none of the pooling PBEs can be justified by a Rational HTE (see Figure 3).

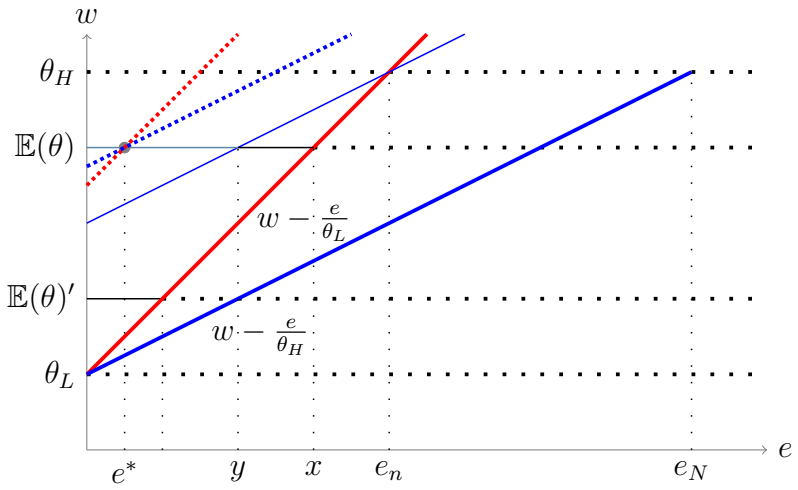

Figure 3: Refinement of pooling PBE at $\mathbb{E}(\theta)$ but no refinement at $\mathbb{E}(\theta)'$.

Below, we present a family of pooling equilibria, each passing the Rational HT refinement:

Proposition 2 *For each education level e^* such that $0 \leq e^* \leq y := (\theta_H - \theta_L)\big(\theta_L - (1-\alpha)\theta_H\big)$, the strategy profile (b_S^*, w_R^*) and beliefs $\mu^* := (\mu^*(\cdot|e)_{e \in \mathcal{M}})$ such that*

[27] Even under wage $w(e)$, such that $w(e) = \theta_H$ for $e > e_n$ and $w(e) = \theta_L$ for $e \leq e_n$, θ_L will not choose $e > e_n$.

(i) $b_S^*(e^*|\theta_L) = b_S^*(e^*|\theta_H) = 1$,

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

constitute a pooling PBE that can be supported by a Rational HTE.

In Rational HTE, the employer believes that any education level below e^* is chosen by the low-productivity type, and pays the minimum wage. For any education level between e^* and e_n, the employer cannot infer the type, and pays the average productivity. The employer believes that any education level above e_n is chosen by the high-productivity type, and pays the highest wage.

Note that Rational HTE refines pooling PBE in two dimensions. The first dimension is the level of education.[28] The second dimension is off-path belief (equivalently, wage).

Finally, let us briefly remark on the role of rationality for the refinement outcome. To this end, assume that hypotheses reflect first-order rationality as in Ortoleva (2012) (see Remark 6). That is, hypotheses are about strategies for the worker that best respond to some (not-necessarily rational) strategy of the employer. For each education level $e \in \{e_n, \ldots, e_N\}$, there is a strategy according to which each type best responds with e (e.g., when the employer pays the wage equal to $\theta_L + \varepsilon + \frac{e}{\theta_L}$ for e, where $\varepsilon > 0$, and θ_L otherwise). Hence, for each $e \in \{e_n, \ldots, e_N\}$, we can construct a hypothesis that justifies the off-path belief $\mu^*(\theta_H|e) = \alpha$, which allows us to explain each pooling PBE of Observation 1 in which the average productivity $\mathbb{E}(\theta)$ is offered off the path.

However, paying average productivity is inconsistent with mutual knowledge of rationality. The rational employer offers the expected productivity (i.e., $w(e) = \mathbb{E}(\cdot|e)$ for each e). Knowing that the employer is rational, the worker will never play a strategy according to which the low-productivity worker signals $e \in \{e_n, \ldots, e_N\}$. On the other hand, knowing that the worker is rational - in the sense of never playing a dominated strategy - the employer must know that each $e \in \{e_n, \ldots, e_N\}$ can only be chosen by the high-productivity type. Consequently, the rational employer will offer the highest wage $w(e) = \theta_H$ for e, and her opponent knows this fact.

Under mutual knowledge of rationality, each pooling PBE that fails the Rational HT refinement but can be justified by first-order rational hypotheses is not robust. Consider such a pooling PBE, i.e., one in which an education level e^* such that $y < e^* \leq x$ and the average productivity $w(e) = \mathbb{E}(\theta)$ is paid for each $e \in \{e_n, \ldots, e_N\}$ (see Figure 2). Knowing that the employer is rational, the payoff of the high-productivity worker from e_n at wage θ_H is larger than his equilibrium payoff,

[28] Interestingly enough, there is experimental evidence in favor of such equilibria. Especially, Kübler, Müller, and Normann (2008) found pooling behavior at lower education levels in the framework of the Spence model.

providing him an incentive to signal e_n instead of e^*. Rational HTE rules out such incentives to deviate, making the equilibrium robust due to its consistency with mutual knowledge of rationality.

In sum, our requirement to justify beliefs by second-order rational hypotheses significantly reduces the number of pooling PBEs in the signaling game of Spence (1973). A Rational HTE, provided it exists, is one in which the employer offers the highest wage for each out-of-equilibrium-message that only the high-productivity type has an incentive to choose (i.e., each $e \in \{e_n, \ldots, e_N\}$).

7 Behaviorally Consistent Hypotheses

In this section, we recapitulate the Stiglitz-Mailath critique of the Intuitive Criterion. Then, we introduce an equilibrium concept with out-of-equilibrium beliefs that are immune to their critique.

The idea is to derive out-of-equilibrium beliefs from hypotheses that are "consistent" with the initial hypothesis in the sense that they both rationalize the same behaviors on the equilibrium path.

In Rational HTE, the initial hypothesis rationalizes the Receiver's behavior on the path. Off the path, the initial hypothesis is rejected. The Receiver may select a new hypothesis which - after updating it along the equilibrium path - will rationalize a different action than her equilibrium action. In this case, the hypothesis is said to be *behaviorally inconsistent* with the initial hypothesis. However, behavioral inconsistency can provide an argument against the new hypothesis that resembles the argument used by Stiglitz (see Cho and Kreps, 1987, p.203) and Mailath (1988) in their critique of the posteriors admitted by the Intuitive Criterion.

Let us briefly recall the Stiglitz-Mailath critique by means of the following example.

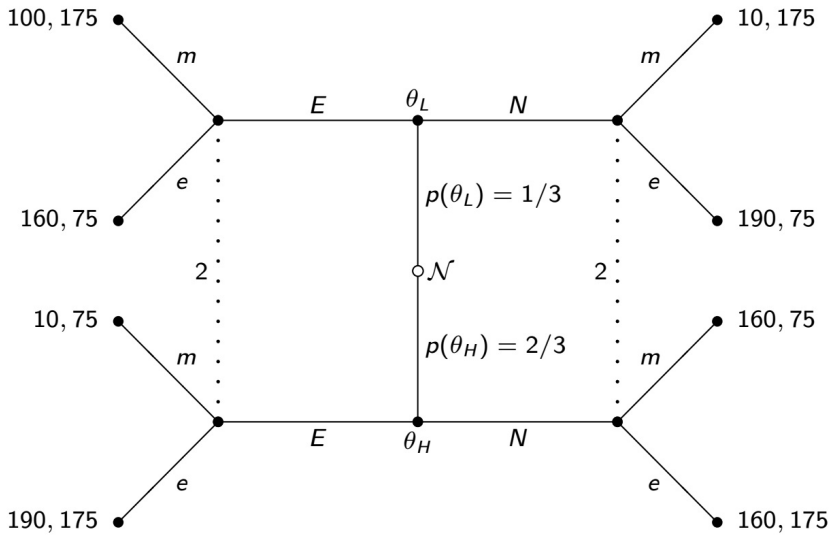

Figure 4: Labor-Market Game 2 in Brandts and Holt (1992)

Example 5a Consider the game depicted in Figure 4. There are two families of pooling PBEs.

PBE-I: In the first family, both types pool on E; i.e,

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

PBE-II: In the second family, both types pool on N; i.e.

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

Consider PBE-II. The Intuitive Criterion asserts that only the high-skilled worker could benefit by choosing the out-of-equilibrium message E. That is, $I(E) = \{\theta_H\}$. Therefore, if E is observed, the employer infers that education is chosen by the high-skilled type (i.e., $\mu(\theta_H|E) = 1$). However, given this belief, the employer prefers to assign the worker to the executive job e instead of matching him with the manual job m. Therefore, the pooling PBE fails the Intuitive Criterion.[29]

Suppose the employer reasons further. Since the worker signaling E receives the executive job, the high-skilled type is strictly better off by choosing E instead of N. If the employer reasons consistently, then she infers that only the low-skilled worker can choose N. Therefore, she will best respond by matching the worker signaling N with the manual job m. This, in turn, will induce the low-skilled worker to signal E. This chain of reasoning provides an argument against $\mu(\theta_H|E) = 1$, the posterior admitted by the Intuitive Criterion. As a consequence, we may discard this equilibrium, although the "intuitive" belief used against PBE-II is implausible itself. This is the essence of the Stiglitz-Mailath critique of the Intuitive Criterion.

To eliminate such "implausible" beliefs, we require that new hypotheses are behaviorally consistent. Let π^* be an initial hypothesis. A rational hypothesis is said to be *behaviorally consistent* with π^* if - after updating it along the equilibrium path - it rationalizes the same behavior as π^*.

Definition 8 (Behaviorally Consistent Hypothesis) *Let π^* be an initial hypothesis. A rational hypothesis π is* behaviorally consistent *with π^* if for each $m \in \mathcal{M}$, such that $\pi^*(m, \Theta) > 0$ and*

$$a^* \in \arg\max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \frac{\pi^*(m, \theta)}{\pi^*(m, \Theta)} u_R(\theta, m, a), \tag{20}$$

it is true that $\pi(m, \Theta) > 0$ and

$$a^* \in \arg\max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \frac{\pi(m, \theta)}{\pi(m, \Theta)} u_R(\theta, m, a). \tag{21}$$

Definition 9 (Behaviorally Consistent Hypothesis Testing Equilibrium) *$(b_S^*, b_R^*, \rho, \mu_\rho^*)$ is a Behaviorally Consistent Hypothesis Testing Equilibrium if $(b_S^*, b_R^*, \rho, \mu_\rho^*)$ satisfies Conditions (i) through*

[29] PBE-I passes the Intuitive Criterion yielding $\mu(\theta_L|N) = 1$.

(v) in Definition 5, and $supp(\rho)$ contains only behaviorally consistent hypotheses.

Behaviorally-consistent hypotheses are rational, and ensure that beliefs are immune to the Stiglitz-Mailath critique. For this reason, we apply Behaviorally Consistent HTE as an additional refinement criterion for PBE. The Behaviorally Consistent HT refinement is defined in the same way as the Rational HT refinement (see Definition 6). A PBE is said to pass the Behaviorally Consistent HT refinement if there exists a Behaviorally Consistent HTE that supports the PBE; otherwise the equilibrium fails the refinement.

Example 5b We show that PBE-II passes the Behaviorally Consistent HT refinement. Consider the following simple-rational hypotheses:

1) $\pi_1' := \{\pi_1(N,\theta_L) = 1/3,\ \pi_1(N,\theta_H) = 2/3\}$ given $\bar{\beta}_R := (\bar{\beta}_S(N|\theta_L) = \bar{\beta}_S(N|\theta_H) = 1)$,

2) $\pi_2' := \{\pi_2(E,\theta_L) = 1/3,\ \pi_2(N,\theta_H) = 2/3\}$ given $\bar{\beta}_R' := (\bar{\beta}_S'(E|\theta_L) = \bar{\beta}_S'(N|\theta_H) = 1)$,

where $\bar{\beta}_R$ and $\bar{\beta}_R'$ best respond against $b_R(m|E) = b_R(e|N) = 1$ and $b_R'(m|E) = b_R'(m|N) = 1$, respectively. Hence, $(b_S^*, b_R^*, \rho, \mu_\rho^*)$ such that

$$b_S^*(N|\theta_L) = b_S^*(N|\theta_H) = 1,\ \ b_R^*(m|E) = b_R^*(e|N) = 1,$$
$$supp(\rho) = \{\pi_1', \pi_2'\} \text{ such that } \rho(\pi_2') < \rho(\pi_1'),\ \mu_\rho^*(\theta_L|N) = 1/3 \text{ and } \mu_\rho^*(\theta_L|E) = 1,$$

is the Behaviorally Consistent HTE that supports the PBE with pooling on N and $\mu^*(\theta_L|E) = 1$. By updating the initial hypothesis π_1', the employer infers that N is more likely to be chosen by the high-skilled type. Therefore, the job applicant is matched with the executive job e. According to π_2', the employer believes that workers "reversely" separate; i.e., the high-skilled worker signals N while the low-skilled worker signals E. By updating π_2' on the equilibrium path, the employer infers that signal N is chosen by the high-skilled worker. Therefore, π_2' rationalizes the same action as the initial hypothesis π_1', showing that π_2' is behaviorally consistent with π_1'.

The employer has no reason to deviate from her equilibrium strategy if she uses π_2' instead on π_1' on the path. At this stage, the inconsistency in reasoning about optimal behaviors on and off the equilibrium paths is prevented. Since the out-of-equilibrium belief is immune to the Stiglitz-Mailath critique, we argue that there is no reason to refute the pooling PBE with $\mu(\theta_L|E) = 1$.

Can we justify each intuitive PBE by a Behaviorally Consistent HTE? From Theorem 2, we know that if a PBE passes the Intuitive Criterion and if the single-type condition is satisfied, then the PBE passes the Rational HT refinement, yielding the same off-path belief as the Intuitive Criterion. However, the intuitive PBE does not need to pass the Behaviorally Consistent HT refinement unless an auxiliary condition is satisfied. The auxiliary condition is presented in the following proposition.

Proposition 3 *Let (b_S^*, b_R^*, μ) be a PBE that passes the Intuitive Criterion and for each out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$, $I(m^\circ)$ is a singleton (i.e., $I(m^\circ) = \{\theta_{m^\circ}\}$ for each m°). If the Receiver's best-response correspondence is single-valued on the equilibrium path; that is, for each equilibrium message m, it is true that $b_R^*(a_m|m) = 1$ for some $a_m \in \mathcal{A}$ and*

$$\{a_m\} = \arg\max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \mu(\theta|m) u_R(\theta, m, a), \tag{22}$$

then there exists a Behaviorally Consistent HTE $(b_S^*, b_R^*, \rho, \mu_\rho^*)$ that supports the intuitive PBE.

Example 5c Consider the PBE with pooling on E and $\mu^*(\theta_L|N) = 1$ (see PBE-I, Example 5a). This equilibrium passes the Intuitive Criterion with $I(N) = \{\theta_L\}$. Moreover, the best-reply correspondence of the employer is single-valued. Let $\bar{\beta}_R$ be the employer's belief that workers follow their equilibrium strategy, i.e., $\bar{\beta}_R = b_S^*$, such that $b_S^* := (b_S^*(E|\theta_L) = b_S^*(E|\theta_H) = 1)$. This belief, combined with p, induces the initial hypothesis $\pi^* = \{\pi(E, \theta_L) = 1/3,\ \pi(E, \theta_H) = 2/3\}$.

For any $\varepsilon \in (0, 1)$, consider the rational strategy $b_S(\varepsilon)$ given by

$$b_S(E|\theta_L) = 1 - \varepsilon,\ b_S(N|\theta_L) = \varepsilon, \text{ and } b_S(E|\theta_H) = 1,$$

which best responds to $b_R = (b_R(e|E) = 1,\ b_R(m|N) = 1/6,\ b_R(e|N) = 5/6)$. Note that given b_R, the low-skilled worker is indifferent between choosing E and N, while the high-skilled worker will best respond with E. Combining the employer's beliefs $\bar{\beta}_R = b_S(\varepsilon)$ with p induces:

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

Note that π_ε yields $\mu_\rho(\theta_L|N) = 1$. Moreover, $\mu_\rho(\theta_H|E) = 2/(3 - \epsilon) > 1/2$ for any $\varepsilon \in (0, 1)$. Thus, $b_R^*(e|E) = 1$ is the best-response strategy with respect to the updated hypothesis π_ε given E, showing that π_ε is behaviorally consistent with the initial hypothesis π^*. Hence, for each $\varepsilon \in (0, 1)$,

$$b_S^*(E|\theta_L) = b_S^*(E|\theta_H) = 1,\ \ b_R^*(e|E) = b_R^*(m|N) = 1,$$
$$supp(\rho) = \{\pi^*, \pi_\varepsilon\} \text{ such that } \rho(\pi_\varepsilon) < \rho(\pi^*),\ \ \mu_\rho^*(\theta_L|E) = 1/3 \text{ and } \mu_\rho^*(\theta_L|N) = 1,$$

is the Behaviorally Consistent HTE supporting the pooling PBE with $\mu^*(\theta_L|N) = 1$.

8 Experimental Findings of Brandts and Holt

In this section we show that the Behaviorally Consistent HTE is consistent with empirical findings.

Brandts and Holt (1992) ran an experiment testing predictions of the Intuitive Criterion for the labor-market games analyzed in the previous sections (i.e., Figure 1 and Figure 4.) Interestingly, subjects behaved consistently with PBEs that pass the Behaviorally Consistent HT refinement.

Consider the labor-market game in Figure 1. This game has the intuitive PBE-1 and the unintuitive PBE-2. Since PBE-1, but not PBE-2, can be supported by a Behaviorally Consistent HTE,[30] our prediction coincides with the Intuitive Criterion. Brandts and Holt (1992) reported that 102 out of 128 subjects' decisions matched with the intuitive PBE-1, and only 7 decisions matched with the unintuitive PBE-2. This result is consistent with our prediction.[31]

There is another interesting finding. Brandts and Holt (1992) analyzed the behavior of Senders. They found some evidence for the new hypothesis π_1 of the Behaviorally Consistent HTE supporting PBE-1 (see footnote 30). According to π_1, the low-skilled type signals N and the high-skilled type signals E. Brandts and Holt (1992) reported that 84 out of 84 high-skilled subjects played E. However, 24 out of 44 low-skilled subjects played the out-of-equilibrium message N.

> "*This type-dependence is consistent with the out-of-equilibrium beliefs that support the intuitive [. . .] equilibrium*" (see Brandts and Holt, 1992, p.1357).

A majority of Receivers seemed to believe that N is sent by low-skilled types in accordance with π_1. Then, 17 out of 24 Receivers who observed N responded with the equilibrium action m. Hence, the reported type-dependence and subjects' replies to N suggest that π_1 is a reasonable hypothesis.

Consider now the game in Figure 4, with the intuitive PBE-I and the unintuitive PBE-II.[32] For this game, our predictions differ from the Intuitive Criterion predictions. Then, PBE-I and PBE-II pass the Behaviorally Consistent HT refinement (see Examples 5b, and 5c). Interestingly, a majority of subjects behaved consistently with the unintuitive PBE-II. As Brandts and Holt (1992) reported, only 23 out of 144 subjects' decisions matched with the intuitive PBE-I, while 84 out of 144 decisions matched with the unintuitive PBE-II.

As in the previous case, Brandts and Holt (1992) found some evidence for the new hypothesis π_2' of the Behaviorally Consistent HTE supporting the unintuitive PBE-I. According to π_2', Senders "reversely" separate, i.e., the low-skilled type chooses E, while the high-skilled type chooses N. Brandts and Holt (1992) found that 72 out of 99 high-skilled Senders played N while 20 out of 45

[30] In Example 3b, we showed that PBE-1 is supported by the Rational HTE with the initial hypothesis π_3 and the new hypothesis π_1. By updating π_1 on the equilibrium path, we obtain $\mu_\rho(\theta_H|E) = 1$, and the employer will match the job applicant with the executive job e. Therefore, π_1 rationalizes the same behavior as π_3, showing that π_1 is behaviorally consistent with π_3. Thus, PBE-1 passes the Behaviorally Consistent HT refinement.
Regarding PBE-2, we have shown that this family already fails the Rational HT refinement (see Example 3b).

[31] This game was implemented in their Treatment 1. The results of this treatment are summarized in Table 3 (parts (a) and (c)) (see Brandts and Holt, 1992, p.1358).

[32] This game was implemented in their Treatment 5. Results of this treatment are summarized in Table 4 (parts (a) and (b)) (see Brandts and Holt, 1992, p.1363).

low-skilled Senders played E. Notably, a significant number of Receivers believed that Senders do "reversely" separate. Then, 24 out of 47 Receivers who observed E responded with m.

In sum, the Behaviorally Consistent HTE can explain the experimental results of Brandts and Holt (1992) better than the Intuitive Criterion, making our solution concept empirically relevant.[33]

9 Conclusion

In this paper, we have suggested solution concepts for signaling games. Our equilibrium notions admit belief updating at information sets with zero probability. Off-path beliefs are derived from hypotheses about strategic behavior of the Sender, together with the prior information about types.

We have argued that hypotheses offer a useful formal tool to develop belief-driven refinements. To reduce the number of Perfect Bayesian Equilibria, we have suggested two refinement criteria.

Our first refinement requires that beliefs are derived from rational hypotheses. This criterion ensures that off-path beliefs are consistent with mutual knowledge of rationality. Our second refinement is more stringent. It requires that beliefs are derived from behaviorally consistent hypotheses, ensuring that off-path beliefs are immune to the Stiglitz-Mailath critique of the Intuitive Criterion.

Our refinement criteria provide an alternative approach to equilibrium selection. Equilibrium selection based on the Intuitive Criterion has been criticized by many authors including Mailath (1988), van Damme (1989), Mailath, Okuno-Fujiwara, and Postlewaite (1993) as being "implausible" due to inconsistency in reasoning between behaviors on and off the equilibrium paths. The refinement based on behaviorally consistent hypotheses does not only eliminate such inconsistencies but it delivers predictions that are consistent with experimental findings on equilibrium behavior. Thus, we believe that our equilibrium notions are worth further exploration and application.

A Proofs

Proof of Theorem 1. Consider a PBE, (b_S^*, b_R^*, μ^*). Let $\mathcal{M}^\circ$ be the set of out-of-equilibrium messages and $\{\mu^*(\cdot|m^\circ)\}_{m^\circ \in \mathcal{M}^\circ}$ be the family of equilibrium beliefs. The proof consists of two steps. In Step 1, we construct a hypothesis that induces the equilibrium beliefs on the path. In Step 2, for each $m^\circ \in \mathcal{M}^\circ$, we construct a hypothesis that induces the off-path belief $\mu^*(\cdot|m^\circ)$.

Step 1. Consider the equilibrium strategy b_S^*. Let β_R be the Receiver's belief, such that $\beta_R = b_S^*$.

[33] Other studies have tested the Intuitive Criterion. For instance, Banks, Camerer, and Porter (1994) found evidence in favor of intuitive PBE. However, implementing similar games as Banks, Camerer, and Porter (1994), Brandts and Holt (1993) could not find unequivocal support for intuitive PBEs. Instead, Brandts and Holt (1993) replicated similar patterns of equilibrium behaviors as reported in Brandts and Holt (1992). Over a series of treatments, a majority of subjects behaved consistently with an unintuitive PBE that is also consistent with Behaviorally Consistent HTE.

This belief and the prior p on Θ define the following hypothesis: for each $(m,\theta) \in \mathcal{M} \times \Theta$,

$$\pi^*(m,\theta) = b_S^*(m|\theta)p(\theta). \tag{23}$$

By Condition (iii) in Definition 1, for any $m \in \mathcal{M}$ such that $\pi^*(m,\Theta) > 0$, we have

$$\mu^*(\theta|m) = \mu_\rho(\theta|m) = \frac{\pi^*(m,\theta)}{\pi^*(m,\Theta)} \quad \text{for each } \theta \in \Theta, \tag{24}$$

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

Step 2. Fix an out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$. Consider the following strategy b_S for the Sender:

$$b_S(m|\theta) = \begin{cases} \left(\frac{\mu^*(\theta|m^\circ)}{p(\theta)}\right)\frac{1}{X}, & \text{for } m = m^\circ, \\ 1 - \left(\frac{\mu^*(\theta|m^\circ)}{p(\theta)}\right)\frac{1}{X}, & \text{for some } m \in (\mathcal{M} \setminus \{m^\circ\}), \\ 0, & \text{otherwise,} \end{cases} \tag{25}$$

where $X := \sum_{\theta\in\Theta} \frac{\mu^*(\theta|m^\circ)}{p(\theta)}$. Since $X \geq \frac{\mu^*(\theta|m^\circ)}{p(\theta)}$ 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 $\mu^*(\cdot|m^\circ)$ play m° with a strictly positive probability (i.e., $b_S(m^\circ|\theta) > 0$ for each $\theta \in \Theta$, such that $\mu^*(\theta|m^\circ) > 0$).

Now, $\beta_R = b_S$ and the prior p induce the following hypothesis π_{m°: for every $(m,\theta) \in \mathcal{M} \times \Theta$,

$$\pi_{m^\circ}(m,\theta) = \beta_R(m|\theta)p(\theta). \tag{26}$$

When updating π_{m° conditional on m°, by (25) and (26), we have

$$\mu_\rho(\theta|m^\circ) = \frac{\pi_{m^\circ}(m^\circ,\theta)}{\pi_{m^\circ}(m^\circ,\Theta)} = \frac{\frac{\mu^*(\theta|m^\circ)}{\sum_{\theta\in\Theta}\frac{\mu^*(\theta|m^\circ)}{p(\theta)}}}{\frac{1}{\sum_{\theta\in\Theta}\frac{\mu^*(\theta|m^\circ)}{p(\theta)}}} \text{ for every } \theta \in \Theta, \tag{27}$$

yielding the off-path belief that coincides with the PBE belief. That is,

$$\mu_\rho(\theta|m^\circ) = \mu^*(\theta|m^\circ) \text{ for each } \theta \in \Theta. \tag{28}$$

Hence, there exists a hypothesis π_{m° that induces the out-of-equilibrium belief $\mu^*(\cdot|m^\circ)$ for m°.

Since we chose m° arbitrarily, for each out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$, there is a hypothesis π_{m° that will induce $\mu^*(\cdot|m^\circ)$. Let $\{\pi_{m^\circ}\}_{m^\circ\in\mathcal{M}^\circ}$ be the family of such hypotheses.

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

$$\{\pi^*\} := \arg\max_{\pi\in supp(\rho)} \rho(\pi) \text{ and } \{\pi^{**}_{m^\circ}\} := \arg\max_{\pi\in supp(\rho)} \rho_{m^\circ}(\pi) \text{ for each } m^\circ \in \mathcal{M}^\circ, \tag{29}$$

(i.e., π^* is the most likely hypothesis w.r.t ρ and $\pi^{**}_{m^\circ}$ is the most likely hypothesis w.r.t ρ_{m° for $m^\circ \in \mathcal{M}^\circ$). Thus, there exists a Focused HTE $(b^*_S, b^*_R, \rho, \mu^*_\rho)$ supporting the PBE (b^*_S, b^*_R, μ^*). ■

Proof of Proposition 1. To prove that the refinement criteria are not nested, we provide two cases. In Case 1, we show a PBE that passes the Rational HT refinement but fails the Intuitive Criterion. In Case 2, we present a PBE that fails the former but passes the latter criterion.

Case 1. Consider the PBEs with pooling on N for the labor-market game in Figure 4 (i.e., PBE-II):

$$b^*_S(N|\theta_L) = b^*_S(N|\theta_H) = 1,\ b^*_R(m|E) = 1,\ r^*(e|N) = 1,\ \ \mu^*(\theta_L|N) = 1/3 \text{ and } \mu^*(\theta_L|E) \geq 1/2.$$

Consider the rational hypotheses π'_1 and π'_2 depicted in Example 5b. Then, the Rational HTE,

$$b^*_S(N|\theta_L) = b^*_S(N|\theta_H) = N,\ b^*_R(m|E) = 1,\ b^*_R(e|N) = 1,$$
$$supp(\rho) = \{\pi'_1, \pi'_2\} \text{ such that } \rho(\pi'_2) < \rho(\pi'_1),\ \ \mu^*_\rho(\theta_L|N) = 1/3 \text{ and } \mu^*_\rho(\theta_L|E) = 1,$$

supports the PBE with pooling on N and the out-of-equilibrium belief $\mu^*(\theta_L|E) = 1$.

Now, we argue that PBE-II 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, showing that PBE-II fails the Intuitive Criterion.

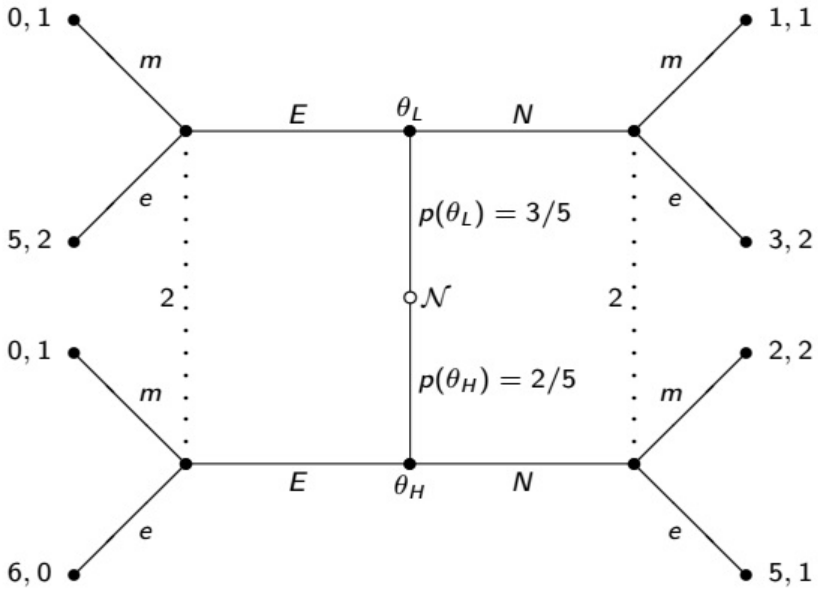

Figure 5: An intuitive PBE that fails the Rational HT refinement

Case 2. Consider the game in Figure 5. It has the following family of PBEs with pooling on N:

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

First, we show that there does not exist any Rational HTE that supports any of the PBEs. 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 support the Receiver's off-the-equilibrium behavior $b_R^*(m|E)$, we need to determine the strategies for the Sender that best respond against $b_R \in \mathcal{B}_R^\bullet$. 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 in response to N, respectively. Then, θ_L will choose E if

$$2y - 5x \geq -2 \quad \Leftrightarrow \quad y \geq \frac{5}{2}x - 1. \tag{30}$$

Similarly, θ_H will choose E if

$$3y - 6x \geq -1 \quad \Leftrightarrow \quad y \geq 2x - \frac{1}{3}. \tag{31}$$

Note that $2x - \frac{1}{3} > \frac{5}{2}x - 1$ for any $x \in [0,1]$. Hence, for each $(x,y) \in [0,1] \times [0,1]$ that satisfies (31), (30) is satisfied with strict inequality. This means that type θ_L strictly prefers E to N whenever type θ_H weakly prefers E. That is, for any $b_R \in \mathcal{B}_R^\bullet$, whenever

$$\sum_{a \in \mathcal{A}} u_S(\theta_H, E, a) b_R(a|E) \geq \sum_{a \in \mathcal{A}} u_S(\theta_H, N, a) b_R(a|N), \tag{32}$$

we have

$$\sum_{a \in \mathcal{A}} u_S(\theta_L, E, a) b_R(a|E) > \sum_{a \in \mathcal{A}} u_S(\theta_L, N, a) b_R(a|N). \tag{33}$$

Hence, there is no rational strategy $b_S \in \mathcal{B}_S^\bullet$ such that $b_S(E|\theta_H) > b_S(E|\theta_L)$. Thus, by updating any hypothesis based on a system of beliefs $\bar{\beta}_R \in \mathcal{B}_S^\bullet$, we will have $\mu_\rho(\theta_L|E) \geq 3/5$. However, given such posteriors, the Receiver will choose e instead of m. Hence, there does not exist any Rational HTE supporting any of the PBEs. Therefore, all PBEs fail the Rational HT refinement.

Now, we show that each pooling PBE 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 the out-of-equilibrium message 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 from the equilibrium strategy as long as $\mu(\theta_L|E) \leq 1/2$. Therefore, the whole family of pooling PBEs passes the Intuitive Criterion. ■

Proof of Theorem 2. Consider a PBE, (b_S^*, b_R^*, μ^*). Fix an out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$. Let a^* be a response to m° played with a strictly positive probability by the Receiver (i.e.,

$b_R^*(a^*|m^\circ) > 0$). By assumption, there is a single type that could benefit by playing m°. That is, $I(m^\circ) = \{\theta_{m^\circ}\}$ for some $\theta_{m^\circ} \in \Theta$. Moreover, the PBE passes the Intuitive Criterion. This means that a^* is optimal with respect to $\mu^*(\theta_{m^\circ}|m^\circ) = 1$, the posterior admitted by the Intuitive Criterion.

The proof consists of two steps. In Step 1, we show that for each out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$, there exists a rational hypothesis π_{m° that is consistent with m° (i.e., $\pi_{m^\circ}(m^\circ, \theta) > 0$ for some $\theta \in \Theta$). Moreover, π_{m° induces the out-of-equilibrium belief $\mu^*(\theta_{m^\circ}|m^\circ) = 1$. In Step 2, we construct a rational hypothesis that induces the equilibrium beliefs on the path.

Step 1. Denote by $u_S^*(\theta)$ the expected equilibrium payoff for type θ. By the single-type condition, there exists a best response against m° to some belief over Θ. That is, there is $a^\circ \in BR(\Theta, m^\circ)$ such that

$$u_S^*(\theta_{m^\circ}) \leq u_S(\theta_{m^\circ}, m^\circ, a^\circ), \tag{34}$$

and

$$u_S^*(\theta) > u_S(\theta, m^\circ, a^\circ) \text{ for } \theta \in \Theta \setminus \{\theta_{m^\circ}\}. \tag{35}$$

That is, when the Receiver chooses a° in response to m°, only type θ_{m° could benefit by playing m°, thus deviating from the equilibrium strategy b_S^*.

Now, we construct a rational strategy $b_R^\circ \in \mathcal{B}_R^\bullet$ for the Receiver. For each $m \in \mathcal{M} \setminus \{m^\circ\}$,

$$b_R^\circ(a|m) = b_R^*(a|m) \text{ for each } a \in \mathcal{A}, \tag{36}$$

and otherwise

$$b_R^\circ(a|m^\circ) = \begin{cases} 0, & \text{for } a \in \mathcal{A} \setminus \{a^\circ\}, \\ 1, & \text{for } a = a^\circ. \end{cases} \tag{37}$$

The strategy b_R° coincides with the equilibrium strategy b_R^* for all messages except m°. By the single-type condition, a° is a best-response to m°. Hence, b_R° is rational, i.e., $b_R^\circ \in \mathcal{B}_R^\bullet$.

By construction of b_R°, m° is a best response to $b_R^\circ(\cdot|m^\circ)$ only for type θ_{m°. That is,

$$m^\circ \in \arg\max_{m \in \mathcal{M}} \sum_{a \in \mathcal{A}} u_S(\theta, m, a)\, b_R^\circ(a|m) \text{ for } \theta = \theta_{m^\circ}, \tag{38}$$

and

$$m^\circ \notin \arg\max_{m \in \mathcal{M}} \sum_{a \in \mathcal{A}} u_S(\theta, m, a)\, b_R^\circ(a|m) \text{ for } \theta \in \Theta \setminus \{\theta_{m^\circ}\}. \tag{39}$$

Note that the equilibrium strategy $b_S^*(\cdot|\theta)$ is a best response to b_R° for any $\theta \in \Theta \setminus \{\theta_{m^\circ}\}$. Now, consider a strategy b_S° for the Sender, which is defined as follows: for each $\theta \in \Theta \setminus \{\theta_{m^\circ}\}$,

$$b_S^\circ(m|\theta) = b_S^*(m|\theta) \text{ for each } m \in \mathcal{M}, \tag{40}$$

and if $\theta = \theta_{m^\circ}$,

$$b_S^\circ(m|\theta) = \begin{cases} 0, & \text{for } m \in \mathcal{M} \setminus \{m^\circ\}, \\ 1, & \text{for } m = m^\circ. \end{cases} \tag{41}$$

Since b_S° best responds against b_R°, there exists a rational strategy $b_S^\circ \in \mathcal{B}_S^\bullet$ that generates m°.

Now, $\bar{\beta}_R = b_S^\circ$, together with the prior p on Θ, defines the following rational hypothesis π_{m°: For each $(m, \theta) \in \mathcal{M} \times \Theta$:

$$\pi_{m^\circ}(m, \theta) := \begin{cases} p(\theta_{m^\circ}), & \text{if } (m, \theta) = (m^\circ, \theta_{m^\circ}), \\ b_S^*(m|\theta)p(\theta), & \text{if } (m, \theta) \neq (m^\circ, \theta_{m^\circ}). \end{cases} \tag{42}$$

Since $\pi(m^\circ, \theta_{m^\circ}) = p(\theta_{m^\circ}) > 0$, π_{m° is consistent with m°. By updating π_{m°, we obtain

$$\mu_\rho(\theta_{m^\circ}|m^\circ) = \frac{\pi_{m^\circ}(m^\circ, \theta_{m^\circ})}{\pi_{m^\circ}(m^\circ, \Theta)} = 1, \tag{43}$$

showing that π_{m° justifies the out-of-equilibrium belief of the Intuitive Criterion, $\mu^*(\theta_{m^\circ}|m^\circ) = 1$.

Since m° was chosen arbitrarily, we can construct a rational hypothesis for any out-of-equilibrium message in $\mathcal{M}^\circ$. That is, for any $m^\circ \in \mathcal{M}^\circ$, there exists a rational hypothesis π_{m° that induces $\mu^*(\theta_{m^\circ}|m^\circ) = 1$. Let $\{\pi_{m^\circ}\}_{m^\circ \in \mathcal{M}^\circ}$ be the collection of such rational hypotheses.

Step 2. Consider the equilibrium strategy b_S^*. Since $b_R^* \in \mathcal{B}_R^\bullet$ and b_S^* best responds to b_R^*, b_S^* is rational (i.e., $b_S^* \in \mathcal{B}_S^\bullet$). Then, $\bar{\beta}_R = b_S^*$, together with p on Θ, induces the rational hypothesis π^*: For each $(m, \theta) \in \mathcal{M} \times \Theta$,

$$\pi^*(m, \theta) = b_S^*(m|\theta)p(\theta). \tag{44}$$

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

$$\{\pi^*\} := \arg\max_{\pi \in supp(\rho)} \rho(\pi) \text{ and } \{\pi_{m^\circ}^{**}\} := \arg\max_{\pi \in supp(\rho)} \rho_{m^\circ}(\pi) \text{ for each } m^\circ \in \mathcal{M}^\circ. \tag{45}$$

Hence, there exists a Rational HTE $(b_S^*, b_R^*, \rho, \mu_\rho^*)$ supporting the intuitive PBE, (b_S^*, b_R^*, μ^*). ■

Proof of Corollary 2. Let (b_S^*, b_R^*, μ^*) be an intuitive PBE and $\mathcal{M}^\circ$ be the set of off-path messages. By the single-type condition, for each $m^\circ \in \mathcal{M}^\circ$, $I(m^\circ) = \{\theta_{m^\circ}\}$ for some $\theta_{m^\circ} \in \Theta$. By Theorem 2 (Step 1), there always exists a family of rational hypotheses $\{\pi_{m^\circ}\}_{m^\circ \in \mathcal{M}^\circ}$; each π_{m° being consistent with m° (i.e., $\pi_{m^\circ}(m^\circ, \theta_{m^\circ}) > 0$). It remains to show that any such family $\{\pi_{m^\circ}\}_{m^\circ \in \mathcal{M}^\circ}$ induces the unique family of out-of-equilibrium beliefs $(\mu_\rho(\theta_{m^\circ}|m^\circ) = 1)_{m^\circ \in \mathcal{M}^\circ}$.

By Condition (13), playing $m^\circ \in \mathcal{M}^\circ$ by some type $\theta \in T(m^\circ)$ is a never-best response. Hence, there does not exist a rational strategy $b_R \in \mathcal{B}_R^\bullet$ against which some type in $T(m^\circ)$ could best respond with $m^\circ \in \mathcal{M}^\circ$. Recall that $T(m^\circ)$ is the set of types that cannot improve upon

their equilibrium payoff by choosing m°. This implies that there does not exist a rational strategy $b_S \in \mathcal{B}_S^\bullet$ according to which some type in $T(m^\circ)$ signals m°. Hence, for all $b_S \in \mathcal{B}_S^\bullet$, $b_S(m^\circ|\theta) = 0$ for each $\theta \in T(m^\circ)$ and $m^\circ \in \mathcal{M}^\circ$. Consequently, for each $m^\circ \in \mathcal{M}^\circ$ and $\theta \in T(m^\circ)$, there does not exist any rational hypothesis π, such that $\pi(m^\circ, \theta) > 0$. Thus, the behavior strategy b_S° defined in (40) and (41) is the only rational strategy for the Sender, such that $b_S^\circ(m^\circ|\theta_{m^\circ}) > 0$. Hence, any rational hypothesis $\bar{\pi}_{m^\circ}$ that is consistent with $m^\circ \in \mathcal{M}^\circ$ must be such that

$$\bar{\pi}_{m^\circ}(m^\circ, \theta_{m^\circ}) = \bar{\beta}_R(m^\circ|\theta_{m^\circ})p(\theta_{m^\circ}) = p(\theta_{m^\circ}), \tag{46}$$

where $\bar{\beta}_R(m^\circ|\theta_{m^\circ}) = b_S(m^\circ|\theta_{m^\circ}) = 1$ and $I(m^\circ) = \{\theta_{m^\circ}\}$. Moreover, $\bar{\pi}_{m^\circ}(m^\circ, \Theta) = p(\theta_{m^\circ})$. Hence, by updating any such $\bar{\pi}_{m^\circ}$ conditional on m°, we have

$$\mu_\rho(\theta_{m^\circ}|m^\circ) = \frac{\bar{\pi}_{m^\circ}(m^\circ, \theta_{m^\circ})}{\bar{\pi}_{m^\circ}(m^\circ, \Theta)} = 1. \tag{47}$$

Hence, any family of rational hypotheses $\{\pi_{m^\circ}\}_{m^\circ \in \mathcal{M}^\circ}$ constructed in Step 1 of the proof of Theorem 2 yields the unique family of off-the-path beliefs $(\mu_\rho(\theta_{m^\circ}|m^\circ) = 1)_{m^\circ \in \mathcal{M}^\circ}$. Thus, the Rational HTE $(b_S^*, b_R^*, \rho, \mu_\rho^*)$ with $supp(\rho) = \{\pi^*, \pi_{m^\circ}^{**}\}_{m^\circ \in \mathcal{M}^\circ}$ supporting the intuitive PBE is unique. Hence, the Rational HT refinement is unique. ■

Proof of Proposition 2. We show that the following PBEs pass the Rational HT refinement.

(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, $\bar{\beta}_R = b_S^*$, together with p, induces the following rational hypothesis:

$$\pi_0^*(e, \theta) = b_S^*(e|\theta)p(\theta) \ \text{ for each } (e, \theta) \in \mathcal{M} \times \Theta. \tag{48}$$

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

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

$$\mathcal{P}(\mathcal{M}^\circ) = \big\{ \underbrace{\{e_0, \ldots, e_{i-1}\}}_{\text{Case 1}}, \underbrace{\{e_{i+1}, \ldots, e_n\}}_{\text{Case 2}}, \underbrace{\{e_{n+1}, \ldots, e_N\}}_{\text{Case 3}} \big\}, \tag{49}$$

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:

$$w_1(e) = \begin{cases} w' & \text{if } e = e', \\ \theta_H, & \text{if } e = e_n, \\ \theta_L, & \text{elsewhere,} \end{cases} \quad \text{and} \quad \mu(\theta_L|e) = \begin{cases} \frac{\theta_H - w'}{\theta_H - \theta_L}, & \text{if } e = e', \\ 0, & \text{if } e = e_n \\ 1, & \text{elsewhere.} \end{cases} \tag{50}$$

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,

$$w' - \frac{e'}{\theta_L} > \theta_L - \frac{0}{\theta_L}, \text{ or equivalently, } w' > \theta_L + \frac{e'}{\theta_L}. \tag{51}$$

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.,

$$w' - \frac{e'}{\theta_H} < \theta_H - \frac{e_n}{\theta_H}, \text{ or equivalently, } w' < \theta_H + \frac{e'}{\theta_H} - \frac{e_n}{\theta_H}. \tag{52}$$

By (51) and (52), $\mu(\theta_L|e')$ is defined by

$$\frac{e_n - e'}{\theta_H(\theta_H - \theta_L)} < \mu(\theta_L|e') := \frac{\theta_H - w'}{\theta_H - \theta_L} < 1 - \frac{e'.}{\theta_L(\theta_H - \theta_L)}. \tag{53}$$

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, $\bar{\beta}_R = b_S$, together with the prior p, induces the simple-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:

$$w_2(e) = \begin{cases} \theta_H, & \text{if } e = e', \\ \theta_L, & \text{elsewhere,} \end{cases} \quad \text{and} \quad \mu(\theta_L|e) = \begin{cases} 0, & \text{if } e = e', \\ 1, & \text{elsewhere.} \end{cases} \tag{54}$$

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, $\bar{\beta}'_R = b'_S$, together with the prior p, induces the simple-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:

$$w_3(e) = \begin{cases} \theta_H, & \text{if } e = e', \\ \theta_L, & \text{elsewhere,} \end{cases} \quad \text{and} \quad \mu(\theta_L|e) = \begin{cases} 0, & \text{if } e = e', \\ 1, & \text{elsewhere.} \end{cases} \tag{55}$$

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, $\bar{\beta}''_R = b''_S$, together with the prior p, induces the simple-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

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

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

showing that there exists a Rational HTE $(b^*_S, w^*, \rho, \mu^*_\rho)$ supporting the pooling PBE with e^*_i. Therefore, each pooling PBE with e^*_i, such that $e_0 \leq e^*_i \leq y$, passes the Rational HT refinement.■

Proof of Proposition 3. Consider an intuitive PBE, (b^*_S, b^*_R, μ^*). Let $\mathcal{M}^\circ$ be the set of out-of-equilibrium messages. Moreover, for each $m^\circ \in \mathcal{M}^\circ$, $I(m^\circ) = \{\theta_{m^\circ}\}$ for some $\theta_{m^\circ} \in \Theta$ (by the single-type condition). The proof consists of two steps. In Step 1, for each $m^\circ \in \mathcal{M}^\circ$, we show that there exists a rational hypothesis π_{m° that is consistent with m°. In Step 2, we show that π_{m° is behaviorally consistent with the initial hypothesis π^*, which rationalizes the equilibrium behavior on the path. Moreover, π_{m° will induce the off-path belief admitted by the Intuitive Criterion.

Step 1. Consider the rational hypothesis π^* defined by $\bar{\beta}_R = b^*_S$ and p: for each $(m, \theta) \in \mathcal{M} \times \Theta$:

$$\pi^*(m, \theta) = b^*_S(m|\theta)p(\theta). \tag{56}$$

Note that π^* induces the posteriors on the equilibrium path. It will serve as the initial hypothesis.

Now, fix an out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$. We construct a rational hypothesis that is consistent with m°. Denote by $u^*_S(\theta)$ the expected equilibrium payoff for the Sender if his type is θ. By the single-type condition, there exists a best response against m° to some belief over Θ. That is, there is $a^\circ \in BR(\Theta, m^\circ)$ such that

$$u^*_S(\theta_{m^\circ}) \leq u_S(\theta_{m^\circ}, m^\circ, a^\circ), \tag{57}$$

and

$$u^*_S(\theta) > u_S(\theta, m^\circ, a^\circ) \text{ for } \theta \in \Theta \setminus \{\theta_{m^\circ}\}, \tag{58}$$

However, in equilibrium, type θ_{m° does not play m°. Hence, it is true that

$$u_S^*(\theta_{m^\circ}) \geq u_S(\theta_{m^\circ}, m^\circ, a^*) \text{ for any } a^* \in \mathcal{A}, \text{ such that } b_R^*(a^*|m^\circ) > 0. \tag{59}$$

From (57) and (59), we know that there also exists a rational strategy $b_R^\circ \in \mathcal{B}_R^\bullet$ for the Receiver, such that $b_R^\circ(\cdot|m) = b_R^*(\cdot|m)$ for $m \in \mathcal{M} \setminus \{m^\circ\}$ and $b_R^\circ(\cdot|m^\circ) \in \Delta(\mathcal{A})$ that satisfies[34]

$$u_S^*(\theta_{m^\circ}) = \sum_{a\in\mathcal{A}} u_S(\theta_{m^\circ}, m^\circ, a) b_R^\circ(a|m^\circ). \tag{60}$$

By construction of b_R°, type θ_{m° is indifferent between playing his equilibrium strategy $b_S^*(\cdot|\theta_{m^\circ})$ and the out-of-equilibrium message m° in response to b_S°. Any other type $\theta \in \Theta \setminus \{\theta_{m^\circ}\}$ plays the equilibrium strategy $b_S^*(\cdot|\theta)$.

Now, consider the following strategy $b_{S,\varepsilon}^\circ$ for the Sender. Let $\varepsilon \in (0,1)$. For any $\theta \in \Theta \setminus \{\theta_{m^\circ}\}$,

$$b_{S,\varepsilon}^\circ(m|\theta) = b_S^*(m|\theta) \text{ for each } m \in \mathcal{M}, \tag{61}$$

and for $\theta = \theta_{m^\circ}$,

$$b_{S,\varepsilon}^\circ(m|\theta_{m^\circ}) = \begin{cases} (1-\varepsilon) b_S^*(m|\theta_{m^\circ}), & \text{for each } m \in \mathcal{M} \setminus \{m^\circ\}, \\ \varepsilon, & \text{for } m = m^\circ. \end{cases} \tag{62}$$

Since $\sum_{m\in\mathcal{M}} b_S^*(m|\theta_{m^\circ}) = \sum_{m\in\mathcal{M}\setminus\{m^\circ\}} b_S^*(m|\theta_{m^\circ}) = 1$, we have $\sum_{m\in\mathcal{M}} b_{S,\varepsilon}^\circ(m|\theta_{m^\circ}) = 1$. Hence, $b_{S,\varepsilon}^\circ$ is a well-defined probability distribution on $\mathcal{M}$. Note that $b_{S,\varepsilon}^\circ$ best responds to b_R°. Hence, there exists a rational strategy for the Sender (i.e., $b_{S,\varepsilon}^\circ \in \mathcal{B}_S^\bullet$) that generates m°.

Now, we can construct the rational hypothesis $\pi_{m^\circ}(\varepsilon)$ induced by $\bar{\beta}_R = b_{S,\varepsilon}^\circ$ and p:

$$\pi_{m^\circ}(\varepsilon)(m,\theta) = b_{S,\varepsilon}^\circ(m|\theta) p(\theta). \tag{63}$$

By construction (62), $\pi_{m^\circ}(\varepsilon)$ is consistent with m° (i.e., $\pi_{m^\circ}(\varepsilon)(m^\circ, \Theta) > 0$).

Step 2. Since $\mathcal{A}$ is finite, the Receiver's best-response correspondence is single-valued on the equilibrium path (i.e., $b_R^*(a_m^*|m) = 1$ for each $m \in \mathcal{M}$) and continuous in ϵ on $(0,1)$, there is a sufficiently small $\varepsilon^* \in (0,1)$ for which $\pi_{m^\circ}(\varepsilon^*)$ satisfies the following identity:

$$\arg\max_{a\in\mathcal{A}} \sum_{\theta\in\Theta} \frac{\pi^*(m,\theta)}{\pi^*(m,\Theta)} u_R(\theta,m,a) = \{a_m^*\} = \arg\max_{a\in\mathcal{A}} \sum_{\theta\in\Theta} \frac{\pi_{m^\circ}(\varepsilon^*)(m,\theta)}{\pi_{m^\circ}(\varepsilon^*)(m,\Theta)} u_R(\theta,m,a)$$

[34] It is possible that a° is an equilibrium action (i.e, $b_R^*(a^\circ|m^\circ) > 0$) when (57) is binding. In this case, b_R^* and b_R° are the same.

for each message m on the equilibrium path (i.e., $\pi^*(m|\Theta) > 0$). That is, on the path, $\pi_{m^\circ}(\varepsilon^*)$ rationalizes the same action as π^*. This proves that $\pi_{m^\circ}(\varepsilon^*)$ is behaviorally consistent with π^*.

To show that $\pi_{m^\circ}(\varepsilon^*)$ induces the same posteriors as the Intuitive Criterion, we update $\pi_{m^\circ}(\varepsilon^*)$ conditional on m°, yielding

$$\mu^*_\rho(\theta|m^\circ) = \frac{\pi_{m^\circ}(\varepsilon^*)(m^\circ,\theta)}{\pi_{m^\circ}(\varepsilon^*)(m^\circ,\Theta)} = \frac{b^\circ_{S,\varepsilon^*}(m^\circ|\theta)p(\theta)}{\sum_{\theta'\in\Theta} b^\circ_{S,\varepsilon^*}(m^\circ|\theta')p(\theta')} \text{ for every } \theta\in\Theta. \tag{64}$$

Since $b^\circ_{S,\varepsilon^*}(m^\circ|\theta_{m^\circ}) = \epsilon > 0$ and $b^\circ_{S,\varepsilon^*}(m^\circ|\theta) = 0$ for any type $\theta \neq \theta_{m^\circ}$, we have

$$\mu^*_\rho(\theta|m^\circ) = \begin{cases} 0, & \text{for } \theta\in\Theta\setminus\{\theta_{m^\circ}\}, \\ 1, & \text{for } \theta = \theta_{m^\circ}, \end{cases} \tag{65}$$

showing that $\pi_{m^\circ}(\varepsilon^*)$ justifies the posterior, $\mu(\theta_{m^\circ}|m^\circ) = 1$, admitted by the Intuitive Criterion.

Furthermore, since the PBE with $\mu(\theta_{m^\circ}|m^\circ) = 1$ passes the Intuitive Criterion, the rational hypothesis $\pi_{m^\circ}(\varepsilon^*)$ rationalizes the out-of-equilibrium behavior of the PBE (i.e., $b^*_R(\cdot|m^\circ)$).

Step 3. Since m° was chosen arbitrarily, we can construct a behaviorally consistent hypothesis $\pi_{m^\circ}(\varepsilon^*(m^\circ))$ for each out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$.[35] Finally, we can suitably choose a second-order prior ρ with $supp(\rho) = \{\pi^*, \pi_{m^\circ}(\varepsilon^*(m^\circ))\}_{m^\circ\in\mathcal{M}^\circ}$, such that

$$\{\pi^*\} := \arg\max_{\pi\in supp(\rho)} \rho(\pi) \text{ and } \{\pi_{m^\circ}(\varepsilon^*(m^\circ))\} := \arg\max_{\pi\in supp(\rho)} \rho_{m^\circ}(\pi) \text{ for each } m^\circ\in\mathcal{M}^\circ, \tag{66}$$

where π^* is the initial hypothesis w.r.t ρ and $\pi_{m^\circ}(\varepsilon^*(m^\circ))$ is the most likely hypothesis w.r.t ρ_{m° for $m^\circ \in \mathcal{M}^\circ$. Hence, we derived a Behaviorally Consistent HTE, $(b^*_S, b^*_R, \rho, \mu^*_\rho)$, that supports the intuitive PBE with $\mu(\theta_{m^\circ}|m^\circ) = 1$ for each for $m^\circ \in \mathcal{M}^\circ$. ■

B The Intuitive Criterion and Arbitrary Beliefs.

In this Appendix, we show that the Intuitive Criterion does not reduce the number of PBEs, while the Rational HT refinement does.

Consider the game depicted in Figure 6. There is a family of PBEs with pooling on N:

$$b^*_S(N|\theta_L) = b^*_S(N|\theta_H) = 1,\ b^*_R(m|E) = 1,\ r^*(e|N) = 1,\ \mu^*(\theta_L|N) = 4/5\ \mu^*(\theta_L|E) \geq 1/2.$$

The Intuitive Criterion asserts that $I(E) = \{\theta_L, \theta_H\}$ and $T(E) = \emptyset$. Therefore, any probability distribution over $\Theta = \{\theta_L, \theta_H\}$ is admitted. As a consequence, each PBE belief $\mu^*(\theta_L|E) \geq 1/2$

[35] Note that $\varepsilon^*(m^\circ)$ might depend on $m^\circ \in \mathcal{M}^\circ$.

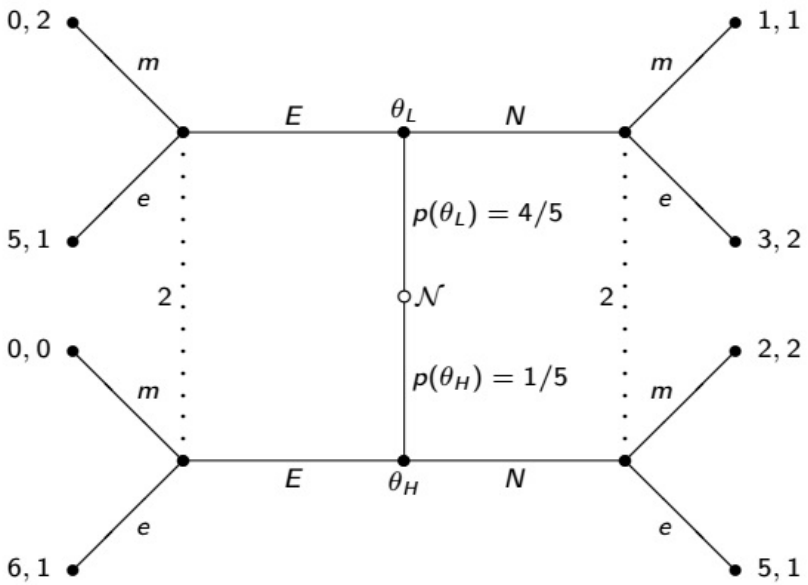

Figure 6: The Intuitive Criterion and Arbitrary Beliefs

is consistent with the Intuitive Criterion, showing each PBE passes the Intuitive Criterion.

Next, we apply the Rational HT refinement. Note that any strategy $b_R = (b_R(\cdot|E), b_R(\cdot|N)) \in [\Delta(\{e, m\})]^2$ is rational, thus, $\mathcal{B}_R^\bullet = \mathcal{B}_R$. Then, for any $b_R \in \mathcal{B}_R^\bullet$, it is true that whenever

$$\sum_{a\in\mathcal{A}} u_S(\theta_H, E, a) b_R(a|E) \geq \sum_{a\in\mathcal{A}} u_S(\theta_H, N, a) b_R(a|N), \tag{67}$$

we have

$$\sum_{a\in\mathcal{A}} u_S(\theta_L, E, a) b_R(a|E) > \sum_{a\in\mathcal{A}} u_S(\theta_L, N, a) b_R(a|N). \tag{68}$$

This implies that 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 any rational hypothesis based on a system of message-contingent beliefs $\bar{\beta}_R \in \mathcal{B}_S^\bullet$, we have

$$\mu_\rho(\theta_L|E) \geq 4/5. \tag{69}$$

Thus, any PBE with $\mu^*(\theta_L|E) \in [1/2, 4/5)$ fails the Rational HTE refinement. ■

C Non-Existence of Equilibria with Simple Hypotheses

In this Appendix, we provide an example for a signaling game for which (i) a Focused HTE with respect to simple hypotheses does not exist and for which (ii) a Rational HTE does not exist.

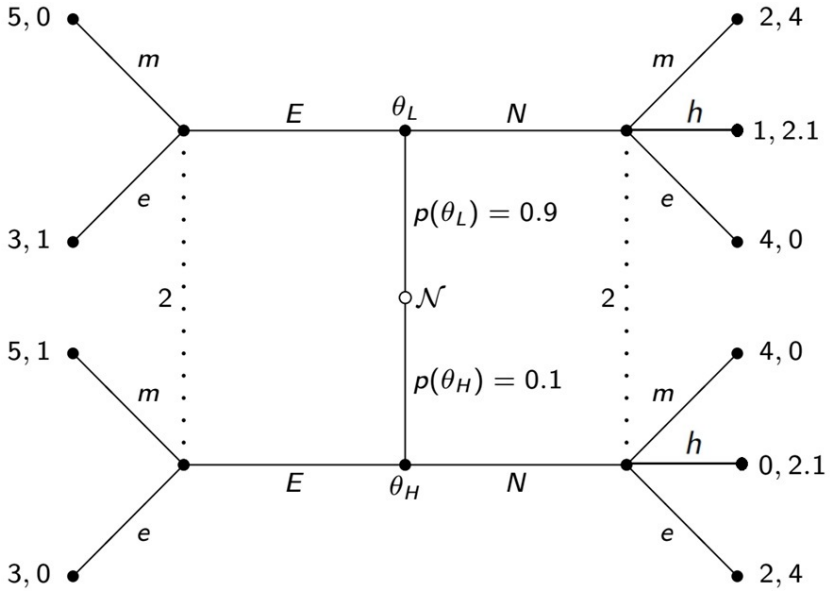

Figure 7: Game with pooling PBE but no Rational HTE.

Consider the game depicted in Figure 7. This game has the following family of pooling PBEs:

$$b_S^*(E|\theta_L) = b_S^*(E|\theta_H) = 1,\ b_R^*(e|E) = 1,\ b_R^*(h|N) = 1,$$
$$\mu^*(\theta_L|E) = 0.9, \text{ and } \mu^*(\theta_L|N) \in [0.475,\ 0.525].$$

Note that (b_S^*, b_R^*) is the only equilibrium strategy profile. Hence, the family of PBEs is unique.

To show (i), consider the four simple hypotheses in this game:

1) $\pi_1 := \{\pi_1(N, \theta_L) = 0.9, \pi_1(E, \theta_H) = 0.1\}$ if $\beta_R(N|\theta_L) = 1$ and $\beta_R(E|\theta_H) = 1$,

2) $\pi_2 := \{\pi_2(E, \theta_L) = 0.9, \pi_2(N, \theta_H) = 0.1\}$ if $\beta_R(E|\theta_L) = 1$ and $\beta_R(N|\theta_H) = 1$,

3) $\pi_3 := \{\pi_3(E, \theta_L) = 0.9, \pi_3(E, \theta_H) = 0.1\}$ if $\beta_R(E|\theta_L) = 1$ and $\beta_R(E|\theta_H) = 1$,

4) $\pi_4 := \{\pi_4(N, \theta_L) = 0.9, \pi_4(N, \theta_H) = 0.1\}$ if $\beta_R(N|\theta_L) = 1$ and $\beta_R(N|\theta_H) = 1$.

Note that π_3 rationalizes the Receiver's best response on the equilibrium path, i.e., given E. However, none of the other hypotheses π_1, π_2 and π_4 rationalizes the Receiver's best response off the path. Thus, a Focused HTE restricted to simple hypotheses does not exist for this game.

To show (ii), note that π_3 is rational, and supports the Receiver's best response on the path. However, there is no rational strategy according to which both types θ_L and θ_H signal N. For such a strategy to exist, the Receiver needs to (optimally) randomize between m and e in response to N. This, however, is not rational as h strictly dominates any mixture between m and e. Hence, there is no rational hypothesis that justifies $\mu^*(\theta_L|N)$, showing that there is no Rational HTE. ■

D Existence of Rational HTE

In this Appendix, we provide sufficiency conditions for existence of a Rational HTE. We consider finite signaling games under conditions that resemble the properties of monotone signaling games with continuous spaces (e.g., see Mailath, 1987; Cho and Sobel, 1990; Kreps and Sobel, 1994).

We assume that Θ, $\mathcal{M}$ and $\mathcal{A}$ are finite, partially ordered sets of real numbers. That is,

$$\begin{aligned}\Theta &= \{\theta_1, \theta_2, \ldots, \theta_T\} \quad \text{where } \theta_t \in \mathbb{R} \text{ for } t = 1, \ldots, T;\\ \mathcal{M} &= \{m_1, m_2, \ldots, m_L\} \quad \text{where } m_l \in \mathbb{R} \text{ for } l = 1, \ldots, L;\\ \mathcal{A} &= \{a_1, a_2, \ldots, a_K\} \quad \text{where } a_k \in \mathbb{R} \text{ for } k = 1, \ldots, K.\end{aligned}$$

For the Sender, we assume that u_S satisfies Monotonicity and Single-Crossing Property.

(i) (Monotonicity) $u_S(\theta, m, a)$ is strictly decreasing in m and strictly increasing in a for any θ.

(ii) (Single-Crossing Property) For each $a \in A$, all $\theta, \theta' \in \Theta$ and $m, m' \in \mathcal{M}$, such that $\theta' > \theta$ and $m' > m$, $u_S(\theta, m, a) \leq u_S(\theta, m', a')$ implies $u_S(\theta', m, a) < u_S(\theta', m', a')$.

For the Receiver, we assume that her best-reply correspondence is message-independent, single-valued, and increasing in θ. Moreover, the "highest" type θ_T has an incentive to signal m_L.

(iii) For each $m \in \mathcal{M}$ and $\mu := \mu(\cdot \mid m) \in \Delta(\Theta)$, $BR(\mu, m) = BR(\mu)$. Moreover, $BR(\mu(\theta) = 1)$ is increasing in θ, and $BR(\mu(\theta) = 1)$ is single-valued for each $\theta \in \Theta$.

(iv) For m_1, m_L and θ_T, $u_S(\theta_T, m_L, BR(\mu(\theta_T) = 1)) \geq u_S(\theta_T, m_1, BR(\mu(\theta_1) = 1))$.

Denote by $\mathcal{G}_M$ the family of signaling games that satisfy Conditions (i) through (iv).

We show that under another mild condition imposed on the prior, p, a pooling Rational HTE exists. Denote by $\overline{m}$ the "highest" message that makes type θ_1 better off than choosing the "lowest" message m_1 if the Receiver believes that $\overline{m}$ is chosen by θ_H while m_1 is chosen by θ_1; i.e.,

$$\overline{m} := \max\left\{m \in \mathcal{M} : u_S(\theta_1, m_1, BR(\mu(\theta_1) = 1)) \leq u_S(\theta_1, m, BR(\mu(\theta_T) = 1))\right\}. \tag{70}$$

Notice that any message $m' \leq \overline{m}$ is rational for type θ_1 in the sense that θ_1 best replies by choosing m' if the Receiver behaves according to the following strategy:

$$b_R(BR(\mu(\theta_T) = 1)|m') = 1 \;\text{ and }\; b_R(BR(\mu(\theta_1) = 1)|m) = 1 \;\text{ for any } m \neq m'.$$

For a given $\overline{m}$, denote by $\underline{a}$ the "lowest" response that satisfies

$$\underline{a} := \min\left\{a \in \mathcal{A} : u_S(\theta_T, \overline{m}, BR(\mu(\theta_T) = 1)) \leq u_S(\theta_T, m_1, a)\right\}. \tag{71}$$

Since $\underline{a} \leq BR(\mu(\theta_T) = 1)$, $\underline{a}$ is well-defined. Note that for any $a \geq \underline{a}$, type θ_T prefers m_1 to $\overline{m}_1$. Moreover, Monotonicity and Single-Crossing Property imply that this is true for all types.

Lemma 1 *If the Receiver's best reply to m_1 is $a \geq \underline{a}$, then any $\theta \in \Theta$ prefers m_1 over $m > \overline{m}$.*

Proof. Let $\{a\} = BR(\mu(\theta_T) = 1)$ be the Receiver's best response against m_1. If $a \geq \underline{a}$, type θ_T prefers choosing m_1 over $\overline{m}$ by Definition (71). By Single-Crossing Property, this is true for all the other types $\theta_t < \theta_T$. Since $BR(\mu(\theta_T) = 1)$ is the highest response by the rational Receiver, Monotonicity implies that m_1 is preferred to any message $m > \overline{m}$ for all $\theta \in \Theta$. ■

For our existence result, we need to ensure that all types have potentially an incentive to pool on some message. To guarantee this, we assume that the prior probability distribution p is "skewed" towards "higher" types so that the Receiver's best response with respect to p is "higher" than $\underline{a}$.[36]

(v) (Skewness) The prior probability distribution $p \in \Delta(\Theta)$ is such that $BR(p) \geq \underline{a}$.

We can now prove existence of a pooling Rational HTE for each signaling games in $\mathcal{G}_M$.[37]

Proposition 4 *Under Conditions (i)-(v), there exists a pooling Rational HTE.*

Proof. Consider a strategy profile (b_S^*, b_R^*), such that $b_S^*(m_1|\theta) = 1$ for any $\theta \in \Theta$ and

$$b_R^*(BR(p)|m') = 1, \; b_R^*(BR(\mu(\theta_T) = 1)|m'') = 1, \tag{72}$$

for any $m', m'' \in \mathcal{M}$, such that $m_1 \leq m' \leq \overline{m}$ and $\overline{m} < m'' \leq m_L$, where $\overline{m}$ is defined as in Equation (70). The Receiver's strategy is optimal with respect to $\mu^* := \{\mu^*(\cdot|m)\}_{m \in \mathcal{M}}$, where

$$\mu^*(\theta_t|m') = p(\theta_t) \text{ for any } \theta_t \in \Theta \text{ and } m_1 \leq m' \leq \overline{m}, \tag{73}$$

$$\mu^*(\theta_T|m'') = 1 \text{ for any } \overline{m} < m'' \leq m_L. \tag{74}$$

By Condition (iii) and Lemma 1, (b_S^*, b_R^*, μ^*) is a PBE. Hence, it remains to show that μ^* can be justified by a set of rational hypotheses.

We distinguish two cases. In Case 1, $\overline{m} = m_L$ and in Case 2, $\overline{m} < m_L$.

Case 1. Consider $\overline{m} = m_L$. In this case, type θ_1 can choose any message m in $\mathcal{M}$ as a best response. Specifically, consider a rational strategy b_R' of the Receiver, such that

$$b_R'(BR(\theta_T) = 1|m') = 1 \text{ and } b_R'(BR(\theta_1) = 1|m) = 1 \text{ for any } m \neq m',$$

[36] For instance, the version of the signaling game of Spence (1973) satisfies Conditions (i)-(v) (see Section 6).

[37] Under an additional (richness) condition imposed on the set of messages, we can show that a separating Rational HTE exists for each signaling game in $\mathcal{G}_M$ (see our Online Appendix).

where $m_1 \leq m' \leq m_L$. By Monotonicity, type θ_1 will choose m' as a best response. Moreover, Single-Crossing Property implies that this is true for all types. Thus, against b'_R, each $\theta_t \in \Theta$ chooses m' as a best response. Let b'_S be such a best response strategy (i.e., $b'_S(m'|\theta_t) = p(\theta_t)$ for each $\theta_t \in \Theta$). The Receiver's belief $\bar{\beta}_R = b'_S$ and p induce the following rational hypothesis $\pi_{m'}$:

$$\pi_{m'}(m, \theta) = b'_S(m|\theta)p(\theta) \text{ for any } (m, \theta) \in \mathcal{M} \times \Theta. \tag{75}$$

By updating $\pi_{m'}$ on m', we thus obtain $\mu^*_\rho(\theta_t|m_1) = p(\theta_t)$ for any $\theta_t \in \Theta$. Thus, $\mu^*_\rho(\theta_t|m') = p(\theta_t)$ for any $\theta_t \in \Theta$ and m', such that $m_1 \leq m' \leq m_L$.

Case 2. Consider $\overline{m} < m_L$. For any m', such that $m_1 \leq m' \leq \overline{m}$, we apply Case 1 to construct the rational hypothesis $\pi_{m'}$ that is consistent with m' as defined in Equation (75).

Now, we construct a rational hypothesis consistent with m'', where $\overline{m} < m'' \leq m_L$. Condition (iv) implies that type θ_T chooses m_L as a best response to the following rational strategy b_R: $b_R := (b_R(BR(\mu(\theta_T) = 1)|m_L) = 1$ and $b_R(BR(\mu(\theta_1) = 1)|m) = 1)$ for any $m \neq m_L$.

Consider another rational strategy b''_R for the Receiver,

$$b''_R(BR(\mu(\theta_T) = 1)|m'') = 1 \text{ and } b''_R(\cdot|m) \in \Delta(\mathcal{A}) \text{ for any } m \neq m'', \tag{76}$$

such that

$$u_S(\theta_T, m'', BR(\mu(\theta_T) = 1)) = \sum_a u_S(\theta_T, m_1, a)b''_R(a|m). \tag{77}$$

By construction of $\underline{a}$, we have $u_S(\theta_T, \overline{m}, BR(\mu(\theta_T) = 1)) \leq u_S(\theta_T, m_1, \underline{a})$. Since Monotonicity implies $u_S(\theta_T, m'', BR(\mu(\theta_T) = 1)) < u_S(\theta_T, \overline{m}, BR(\mu(\theta_T) = 1))$ for $m'' > \overline{m}$, it is true that

$$\sum_a u_S(\theta_T, m_1, a)b''_R(a|m) < u_S(\theta_T, m_1, \underline{a}).$$

By Condition (iii), $\sum_a ab''_R(a|m) \in (a_1, a_T)$, where $\{a_1\} = BR(\mu(\theta_1) = 1)$ and $\{a_T\} = BR(\mu(\theta_T) = 1)$. Therefore, b''_R exists.

Since θ_T is indifferent between choosing m'' and m_1 to b''_R, by Single-Crossing Property, any type θ_t such that $\theta_t < \theta_T$ plays m_1 against b''_R. Thus, the rational strategy b''_S for the Sender,

$$b''_S := (b_S(m_1|\theta_t) = 1\ ,\ b_S(m''|\theta_T) = 1) \text{ for } \theta_t \neq \theta_T, \tag{78}$$

best responds to b''_R. Hence, $\bar{\beta}_R = b''_S$ and p induce the following rational hypothesis $\pi_{m''}$:

$$\pi_{m''}(m, \theta) = b''_S(m|\theta)p(\theta) \text{ for each } (m, \theta) \in \mathcal{M} \times \Theta. \tag{79}$$

By updating $\pi_{m''}$ on m'', we obtain $\mu^*_\rho(\theta_T|m'') = 1$. Thus, $\mu^*_\rho(\theta_T|m'') = 1$ for any $\theta_t \in \Theta$ and m'', such that $\overline{m} < m'' \leq m_L$. For both cases 1 and 2, we can suitably choose a second-order prior ρ, such that $supp(\rho) = \{\pi_{m_1}, \pi_{m^\circ}\}_{m \in \mathcal{M}^\circ}$, where

$$\{\pi^*_{m_1}\} := \underset{\pi \in supp(\rho)}{\arg\max}\, \rho(\pi) \text{ and } \{\pi_{m^\circ}\} := \underset{\pi \in supp(\rho)}{\arg\max}\, \rho_{m^\circ}(\pi) \text{ for each } m^\circ \in \mathcal{M}^\circ, \tag{80}$$

showing that there exists a Rational HTE $(b^*_S, b^*_R, \rho, \mu^*_\rho)$ supporting the PBE, (b^*_S, b^*_R, μ^*). ■

References

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.

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.

BLUME, L., A. BRANDENBURGER, AND E. DEKEL (1991): "Lexicographic Probabilities and Equilibrium Refinements," *Econometrica*, 59(1), 81–98.

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.

FUDENBERG, D., AND K. HE (2018): "Learning and Type Compatibility in Signaling Games," *Econometrica*, 86(4), 1215–1255.

——— (2020): "Payoff Information and Learning in Signaling Games," *Games and Economic Behavior*, 120, 96–120.

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

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.

JEONG, D. (2019): "Job Market Signaling with Imperfect Competition among Employers," *International Journal of Game Theory*, 48(4), 1139–1167.

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

KREPS, D., AND J. SOBEL (1994): "Signalling," in *Handbook of Game Theory with Economic Applications*, ed. by R. Aumann, and S. Hart, vol. 2, pp. 849 – 867. Elsevier, Amsterdam.

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.

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

MAILATH, G. J. (1987): "Incentive Compatibility in Signaling Games with a Continuum of Types," *Econometrica*, 55(6), 1349–1365.

——— (1988): "A Reformulation of a Criticism of the Intuitive Criterion and Forward Induction," *Working Paper*.

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.

NESLON, 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. (2001): "Silver Signals: Twenty-Five Years of Screening and Signaling," *Journal of Economic literature*, 39(2), 432–478.

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.

VAN DAMME, E. (1989): "Stable Equilibria and Forward Induction," *Journal of Economic Theory*, 48(2), 476 – 496.