Testing Behavioral Hypotheses in Signaling Games

Adam Dominiak†1 and Dongwoo Lee‡2

1,2 Department of Economics, Virginia Tech

November 27, 2018

Abstract

This paper develops an equilibrium concept for signaling games, called Hypothesis Testing Equilibrium (HTE). Our equilibrium incorporates Ortoleva’s (2012) theory of belief updating on zero-probability events by selecting and updating hypotheses. A hypothesis is a receiver’s belief about a sender’s strategy. In HTE, 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 HTE is a Perfect Bayesian Equilibrium (PBE). Moreover, we show that each PBE and its posterior beliefs can be explained by HTE, providing a novel justification for PBE beliefs. Hypotheses facilitate reasoning about out-of-equilibrium beliefs. Therefore, we strengthen our equilibrium by using stronger hypotheses as a refinement criterion for PBE. We compare our refinement with the Intuitive Criterion. We also suggest a further refinement criterion under which the selected beliefs 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, out-of-equilibrium beliefs, refinements.

JEL Classification: C72, D81, D83

†Email: dominiak@vt.edu
‡Email: dwlee05@vt.edu

*The authors are deeply indebted to Hans Haller, George Mailath, Gerelt Tserenjigmid, Matthew Kovach, Kevin He, the audiences of the 28th International Conference on Game Theory at Stony Brook and 88th Southern Economic Association Annual Meetings at Washington D.C., and the seminar participants at Australian National University and Korea Institute for International Economic Policy for their valuable comments and fruitful discussions.
1 Introduction

Perfect Bayesian Equilibrium (PBE) is a standard solution concept for dynamic games with incomplete information, such as signaling or cheap talk games. However, this equilibrium has a limitation. It allows for arbitrary out-of-equilibrium beliefs, since Bayes’ rule does not specify how beliefs are derived at information sets with zero probability. In this paper, we develop a novel equilibrium for signaling games that admits updating of beliefs for all messages, including out-of-equilibrium messages. In our equilibrium, receiver’s beliefs are derived by selecting and updating hypotheses about a sender’s behavior. Hypotheses provide a language for reasoning about beliefs. First, we show that our equilibrium is equivalent to PBE, providing an explanation for the origin of PBE beliefs. Second, we introduce a stronger equilibrium as a refinement criterion for PBE. Finally, we show that our refinement approach is consistent with experimental findings.

In signaling games, there are two players: an informed sender and an uninformed receiver. The sender sends a message to convey information about his type. The receiver observes the message, forms her posterior belief about sender’s types, and responds to the message by taking an action.

Our first goal is to develop an equilibrium in which posterior beliefs are well-defined for each message and to compare it with PBE. For this purpose, we need a theory of how the receiver forms her posterior beliefs. We assume that the receiver constructs hypotheses about a sender’s behavior. For each message, the receiver selects a hypothesis about a sender’s strategy that generates the message. Formally, a hypothesis is a joint probability measure on the set of all pairs of messages and types. Two components define a hypothesis: a receiver’s belief about sender’s (pure) strategies and the prior information about types available in a signaling game. A hypothesis is said to be consistent with a message if it assigns a strictly positive probability to it. If a selected hypothesis is consistent with the observed message, the receiver may use it to derive her posterior belief via Bayes’ rule.

To illustrate our hypothesis notion, consider a labor-market game in the spirit of Spence (1973). A worker applying for a job has private information about his skill, being low (L) or high (H). An employer does not know the worker’s skill. However, she has a prior information about workers’ skills in the form of probabilities \( p(H) \) and \( p(L) \). To convey further information about his skill, the worker signals whether he has acquired education or not. Given the observed signal, the employer makes inferences about skills, and assigns the worker to either an executive job or a manual job.

Suppose that the employer observes education. The employer may believe that workers

---

1 This game is depicted in Figure 1 and analyzed throughout the paper.
follow a pooling strategy (i.e., education is acquired regardless of skills). This belief together with the prior information defines a pooling hypothesis according to which the low-skilled worker signals education with probability \( p(L) \) and the high-skilled worker does the same with probability \( p(H) \). This hypothesis is consistent with the observed signal. By updating the hypothesis via Bayes’ rule, the employer concludes that education conveys the same information about skills as the prior one.

However, if no education is observed, the pooling hypothesis is inconsistent with the signal. Thus, the employer concludes that her initial hypothesis is wrong. She selects a new hypothesis that can explain the observed signal. For instance, the employer may believe that only the high-skilled worker has education while the low-skilled worker does not. This belief induces a separating hypothesis. After updating it, the employer infers that the signal is sent by the low-skilled worker.

The idea of deriving beliefs from hypotheses about players’ behavior is not new. It was informally used by Kreps and Wilson (1982) to justify beliefs in sequential equilibrium. In particular, they argued that beliefs should be structurally consistent. That is, at each information set, each belief can be derived from a single strategy that governed the previous moves via Bayes’ rule. In a later paper, Kreps and Ramey (1987, p. 1332) interpreted structurally consistency as follows:

“This means that, at any information set, 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 some single “alternative hypothesis” as to the strategy governing prior play, such that under the hypothesis the information set is reached with positive probability.”

We take this idea further. In particular, we introduce an equilibrium notion, called the Hypothesis Testing Equilibrium (henceforth, HTE), that specifies how hypotheses are selected and updated.

Our equilibrium concept incorporates the updating procedure axiomatized by Ortoleva (2012). In HTE, the receiver has a set of hypotheses about a sender’s behavior. Before any message is observed, the receiver chooses an initial hypothesis. It is the most likely
hypothesis with respect to her prior over hypotheses (i.e., her second-order prior). The initial hypothesis is a receiver’s belief that the sender plays his equilibrium strategy. If an equilibrium message is observed, the receiver updates the initial hypothesis via Bayes’ rule to determine her posterior belief. However, if an out-of-equilibrium message is observed, the receiver rejects the initial hypothesis. Then, she updates her prior over hypotheses via Bayes’ rule, and selects a new hypothesis that is the most likely one according to the updated prior.\(^3\) The new hypothesis is a receiver’s belief about a sender’s strategy that generates the out-of-equilibrium message. The receiver updates the new hypothesis via Bayes’ rule to determine her out-of-equilibrium belief. This updating procedure delivers a system of posterior beliefs that are well-defined for all messages, including out-of-equilibrium messages.

In HTE, posterior beliefs satisfy three desirable properties. First, they are consistent with a theory of choice that admits belief updating on all events. That is, posterior beliefs are well-defined for equilibrium and out-of-equilibrium messages. Second, beliefs are structurally consistent as required by Kreps and Wilson (1982). That is, for each message, a posterior belief is derived from a hypothesis about a sender’s strategy that generates the message. Finally, posterior beliefs are consistent with the prior information about sender’s types that is a primitive of any signaling game.

Our first main result establishes the relationship between the standard PBE and HTE. The primary motivation for the comparison is to study how many PBE beliefs can be explained by hypotheses. We show that PBE and HTE are equivalent solution concepts. In particular, each PBE can be justified by an HTE. This result provides an explanation for the origin of all PBE beliefs. For each PBE belief, there exists a hypothesis about a sender’s strategy – consistent with the prior information about types – that induces the belief. Thus, for signaling games, we show that PBE beliefs are structurally consistent in the sense of Kreps and Wilson (1982).

Our second goal is to strengthen our equilibrium so that we can use it as a refinement for PBE. We impose two behavioral restrictions on hypotheses. First, we require that hypotheses are about a sender’s rational behavior, called strong hypotheses. More precisely, a strong hypothesis is a receiver’s belief about a sender’s strategy that best responds to some of receiver’s (pure) strategies.\(^4\) The corresponding equilibrium is called Strong HTE. Second, we require that strong hypotheses are behaviorally consistent with the initial hypothesis. A

\(^3\)In Ortoleva’s (2012) theory, called the Hypothesis Testing 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 theory, the threshold is assumed to be \(\epsilon = 0\). That is, the receiver rejects her initial hypothesis and selects a new hypothesis if the observed message is a zero-probability event according to the initial hypothesis.

\(^4\)Notice that in HTE, only the initial hypothesis is strong, however, new hypotheses do not need to be strong.
strong hypothesis is behaviorally consistent if, after updating it along the equilibrium path, it rationalizes the same behavior as the initial hypothesis. The corresponding equilibrium is called Behaviorally Consistent HTE.

In Strong HTE, receiver’s beliefs are justified by a sender’s rational behavior. Therefore, we use the Strong HTE that supports a PBE as an argument in favor of the PBE and its out-of-equilibrium beliefs. In other words, a PBE is said to pass the Strong Hypothesis Testing (HT) refinement if there exists a Strong HTE that supports the PBE. Otherwise, the equilibrium fails the refinement.

We compare the Strong HT refinement with the Intuitive Criterion of Cho and Kreps (1987). The Intuitive Criterion does not build on a theory of belief updating. Instead, the idea is to eliminate beliefs that assign strictly positive probability to types that cannot benefit by sending an out-of-equilibrium message. We show that the Intuitive Criterion and our refinement are not nested.\(^5\)

Nonetheless, there is a class of intuitive PBEs that can always be supported by Strong HTE.\(^6\) If for each out-of-equilibrium message of an intuitive PBE there is a single type that could benefit from sending the message, we show that the PBE also passes the Strong HT refinement. This single-type condition guarantees that for each out-of-equilibrium message there is a strong hypothesis that induces the Intuitive Criterion belief according to which the receiver learns the type.

Behaviorally Consistent HTE is more stringent than Strong HTE. Out-of-equilibrium beliefs are justified by hypotheses that are behaviorally consistent with the initial hypothesis (i.e., such hypotheses rationalize a receiver’s behaviors on the equilibrium path – as the initial hypothesis does – and off the equilibrium path). An important property of Behaviorally Consistent HTE is that it selects beliefs that are immune to 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 may use beliefs that lead to inconsistencies in reasoning about behaviors on the equilibrium path and off the equilibrium paths. Behaviorally consistent hypotheses rule out such inconsistencies. We derive a condition under which an intuitive PBE can be justified by 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 predictions of the Intuitive Criterion.\(^7\) In one of

\(^5\)That is, there is a PBE that passes the Intuitive Criterion but fails the Strong HT refinement and vice versa; there is a PBE that passes the Strong HT refinement but fails the Intuitive Criterion.

\(^6\)We use the term *intuitive* PBE as a reference to a PBE that passes the Intuitive Criterion.

\(^7\)In this paper, the games implemented by Brandts and Holt (1992) are presented in Figure 1 and Figure 2.
their experiments, a majority of subjects behaved in line with a pooling PBE that fails the Intuitive Criterion. We show that Behaviorally Consistent HTE can explain the results of Brandts and Holt (1992).

This paper is organized as follows. In Section 2, we recall the standard PBE for signaling games. In Section 3, we formalize the HTE notion and prove our equivalence result. In Section 4, we introduce Strong HTE and define our refinement criterion. In Section 5, we compare the Strong HT refinement with the Intuitive Criterion. In Section 6, we recall the Stiglitz-Mailath critique of the Intuitive Criterion and introduce Behaviorally Consistent HTE. In Section 7, we explain the experimental findings of Brandts and Holt (1992). In Section 8, we provide final remarks.

2 Perfect Bayesian Equilibrium

In this section, we recall the standard Perfect Bayesian Equilibrium (PBE) for signaling games.

In signaling games, there are two players: an informed sender and an uninformed receiver. Nature draws a type from a set of sender’s types $\Theta$ according to a prior probability distribution $p$ on $\Theta$. The sender learns his type and chooses a message $m$ from $M$, a set of messages. A sender’s strategy is denoted by $s : \Theta \rightarrow M$. The receiver observes the message, but not the type, chooses an action $a \in A$, and the game ends. A receiver’s strategy is denoted by $r : M \rightarrow A$. Players’ payoffs are given by $u_S, u_R : \Theta \times M \times A \rightarrow \mathbb{R}$. All sets $\Theta, M$ and $A$ are finite. It is assumed that $p$, called the prior information about types, is known by the players and it has a full support (i.e., $\text{supp}(p) = \Theta$). The class of such signaling games is denoted by $\mathcal{G}$.

Given a message $m \in M$, a receiver’s posterior belief is represented by $\mu(\cdot|m)$, a probability distribution over $\Theta$. We denote by $\mu := \{\mu(\cdot|m)\}_{m \in M}$ a family of posterior beliefs. A strategy profile together with a family of receiver’s posterior beliefs is summarized by a tuple $(s, r, \mu)$.

In this paper, we consider PBE in pure strategies. Formally, a pure PBE is defined as follows.\footnote{In Appendix B, we show that our analysis can be extended to PBE in behavioral strategies.}

**Definition 1 (Perfect Bayesian Equilibrium)** $(s^*, r^*, \mu^*)$ is a pure Perfect Bayesian Equi-
librium for a signaling game in $G$ if:

\begin{equation}
(i) \quad s^*(\theta) \in \arg \max_{m \in \mathcal{M}} u_S(\theta, m, r^*(m)) \text{ for each } \theta \in \Theta,
\end{equation}

\begin{equation}
(ii) \quad r^*(m) \in \arg \max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \mu^*(\theta|m)u_R(\theta, m, a) \text{ for each } m \in \mathcal{M},
\end{equation}

\begin{equation}
(iii) \quad \mu^*(\theta|m) = \frac{\pi(\theta, m)}{\pi(\Theta, m)} \text{ for each } \theta \in \Theta \text{ if } \pi(\Theta, m) > 0, \text{ and }
\end{equation}

\begin{equation}
\mu^*(\cdot|m) \text{ is an arbitrary probability distribution over } \Theta \text{ if } \pi(\Theta, m) = 0, \text{ where }
\end{equation}

\begin{equation}
\pi(\theta, m) = \begin{cases} 
p(\theta), & \text{if } s^*(\theta) = m, \\
0, & \text{otherwise.} \end{cases}
\end{equation}

Conditions (i) and (ii) ensure sequential rationality. That is, each sender’s type best responds to the receiver’s (optimal) strategy, and the receiver best responds to each message by taking into account her posterior belief. Condition (iii) specifies how receiver’s posterior beliefs are determined. For each equilibrium message (i.e., $m \in \mathcal{M}$ such that $\pi(\Theta, m) > 0$), Bayes’ rule is applied. However, for an out-of-equilibrium message (i.e., $m \in \mathcal{M}$ such that $\pi(\Theta, m) = 0$), Bayes’ rule cannot be applied and posterior beliefs are determined arbitrarily.

Given a PBE, $(s^*, r^*, \mu^*)$, we denote by $\mathcal{M}^\circ$ the set of out-of-equilibrium messages. For each $m^\circ \in \mathcal{M}^\circ$, we can specify $\mathcal{O}(m^\circ)$, the set of all out-of-equilibrium beliefs associated with the PBE. That is, each $\mu^*(\cdot|m^\circ) \in \mathcal{O}(m^\circ)$ supports the receiver’s (equilibrium) best response to $m^\circ$. Notice there may be multiple out-of-equilibrium beliefs. Therefore, there may be multiple PBEs supporting the same behavior $(s^*, r^*)$. We will sometime refer to $(s^*, r^*, \mu^*)$ together with the family $\{\mathcal{O}(m^\circ)\}_{m^\circ \in \mathcal{M}^\circ}$ of sets of out-of-equilibrium beliefs as a PBE (instead of multiple PBEs).

**Example 1A.** To illustrate PBE, consider a discrete version of the labor-market game in the spirit of Spence (1973). The game was alluded to in the Introduction and it is depicted in Figure 1.

A worker applying for a job has either low (type $\theta_L$) or high skill (type $\theta_H$). Knowing his type, the worker decides whether to make an investment in education (signal $E$) or not (signal $N$). An employer observes the signal, and matches the worker to either an executive job ($e$) or a manual job ($m$). The prior information $p$ about types is $p(\theta_L) = 1/3$ and $p(\theta_H) = 2/3$. Notice that both worker types prefer the executive job regardless of the 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 the high-skilled worker with the executive job and to match the low-skilled worker...
There are two (pure) PBEs in this game.

**PBE-1**: In the first equilibrium, both worker types signal education. That is, the strategy profile

\[ s^*(\theta_L) = s^*(\theta_H) = E, \quad r^*(E) = e, \quad r^*(N) = m, \]

and posterior beliefs \( \mu^*(\theta_L|E) = 1/3 \) and \( \mu^*(\theta_L|N) \geq 1/2 \) form the first pooling PBE.

**PBE-2**: In the second equilibrium, both worker types do not signal education. That is, the strategy profile

\[ s^*(\theta_L) = E^*(\theta_H) = N, \quad r^*(E) = m, \quad r^*(N) = e, \]

and posterior belief \( \mu^*(\theta_L|N) = 1/3 \) and \( \mu^*(\theta_L|E) \geq 1/2 \) form the second pooling PBE. □

Both equilibria suffer from the multiplicity of out-of-equilibrium beliefs. Therefore, our first goal is to derive a theory that can explain all PBE beliefs. Then, building on our theory, we suggest refinement criteria to reduce the number of out-of-equilibrium beliefs.

In the next section, we present a novel solution concept for signaling games that admits updating of beliefs for all messages, including out-of-equilibrium messages.

# 3 Hypothesis Testing Equilibrium

In this section, we introduce the Hypothesis Testing Equilibrium (henceforth, HTE). We then show that the pure PBE and our HTE are equivalent solution concepts for signaling...
games in $G$.

The main component of HTE is the hypothesis notion. A hypothesis is a probability distribution $\pi$ on $\Theta \times M$. More precisely, a hypothesis is a receiver’s belief about a sender’s choice of (pure) strategies combined with the prior information about types, $p$.

Denote by $S := \{s : \Theta \to M\}$ the set of sender’s (pure) strategies. Let $\beta$ be a probability measure on $S$ that represents a receiver’s belief about a sender’s choice of strategies. A belief $\beta$ on $S$ combined with the prior probability distribution $p$ on $\Theta$ defines a hypothesis as follows.

**Definition 2 (Hypothesis)** A probability distribution $\pi$ on $\Theta \times M$ is called a hypothesis if there exists a belief $\beta$ on $S$ such that, for every $(\theta, m) \in \Theta \times M$:

$$
\pi(\theta, m) = \sum_{s \in S \ s.t. \ s(\theta) = m} \beta(s)p(\theta).
$$

A hypothesis $\pi$ ascribes probability $\pi(\theta, m)$ to the event “type $\theta$ sends message $m$.” Hypotheses are consistent with the prior information $p$ (i.e., $\pi(\theta, M) = p(\theta)$). Let $m$ be a message. A strategy $s$ is said to generate $m$, if $s(\theta) = m$ for some $\theta \in \Theta$. A hypothesis $\pi$ is said to be consistent with $m$, if $\pi(\theta, m) > 0$ (i.e., $\beta(s) > 0$ for some $s \in S$ that generates $m$).

Furthermore, $\pi$ is called a simple hypothesis, if $\beta$ is a degenerate belief on $S$ (i.e., $\beta(s) = 1$ for $s \in S$ and $\beta(s') = 0$ for any $s' \neq s$). According to a simple hypothesis, the receiver believes that the sender plays a single (pure) strategy. That is, $\pi$ ascribes probability $\pi(\theta, m) = p(\theta)$ to the event “type $\theta$ sends message $m$” if $\beta(s) = 1$ for $s \in S$ such that $s(\theta) = m$; otherwise $\pi(\theta, m) = 0$.

**Example 2A.** In the labor-market game depicted in Figure 1, there are four strategies of the sender, i.e., $S = \{s_1, s_2, s_3, s_4\}$. Therefore there are four simple hypotheses:

1) $\pi_1 := \{\pi_1(\theta_L, N) = 1/3, \pi_1(\theta_H, E) = 2/3\}$ if $\beta(s_1) = 1$ and $s_1 := \{s(\theta_L) = N, s(\theta_H) = E\}$,
2) $\pi_2 := \{\pi_2(\theta_L, E) = 1/3, \pi_2(\theta_H, N) = 2/3\}$ if $\beta(s_2) = 1$ and $s_2 := \{s(\theta_L) = E, s(\theta_H) = N\}$,
3) $\pi_3 := \{\pi_3(\theta_L, E) = 1/3, \pi_3(\theta_H, E) = 2/3\}$ if $\beta(s_3) = 1$ and $s_3 := \{s(\theta_L) = E, s(\theta_H) = E\}$,
4) $\pi_4 := \{\pi_4(\theta_L, N) = 1/3, \pi_4(\theta_H, N) = 2/3\}$ if $\beta(s_4) = 1$ and $s_4 := \{s(\theta_L) = N, s(\theta_H) = N\}$.

According to $\pi_1$, the employer believes that workers separate; the high-skilled worker signals $E$ while the low-skilled worker signals $N$. According to $\pi_2$, the employer believes that workers “reversely” separate; the high-skilled worker signals $N$ while the low-skilled worker signals...
According to $\pi_3$ (resp., $\pi_4$), the employer believes that both workers pool on $E$ (resp., on $N$).

The employer might believe that workers pool on $N$ with probability $\lambda \in [0, 1]$ and that they separate with probability $1 - \lambda$. This belief together with $p$ induces a non-simple hypothesis:

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

We assume that the receiver uses hypotheses to derive her posterior beliefs. For each message, the receiver selects a hypothesis that is consistent with the message and updates it via Baye’s rule. To specify how hypotheses are selected, we apply Ortoleva’s (2012) Hypothesis Testing model. Below, we elucidate how his model works in our setup.

Denote by $\Delta(\Theta \times M)$ the set of all probability distributions on $\Theta \times M$. Let $\Pi \subset \Delta(\Theta \times M)$ be the set of all hypotheses associated with a signaling game in $\mathcal{G}$. The receiver is assumed to hold a second-order prior, denoted by $\rho$. The second-order prior $\rho$ is a probability distribution on $\Pi$ with a finite support (i.e., $|\text{supp}(\rho)| \in \mathbb{N}$). Before any information is revealed, the receiver selects an initial hypothesis $\pi^*$. It is the most likely hypothesis with respect to her second-order prior $\rho$. That is,

$$\{ \pi^* \} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi). $$ \hspace{1cm} (2)

After a message $m \in M$ is observed, the receiver conducts a test. If $\pi^*$ is consistent with the message (i.e., $\pi^*(\Theta, m) > 0$), the receiver accepts $\pi^*$ and updates it via Bayes’ rule. However, if $\pi^*$ assigns zero probability to the message (i.e., $\pi^*(\Theta, m) = 0$), then the receiver concludes that the initial hypothesis $\pi^*$ was wrong and rejects it. In this case, the receiver follows Bayes’ rule to update her second-order prior $\rho$ given the unexpected message $m$. Then, she chooses a new hypothesis $\pi_{m^*}$ that is the most likely hypothesis according to her updated second-order prior, i.e.,

$$\{ \pi_{m^*} \} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi|m), $$ \hspace{1cm} (3)

where

$$\rho(\pi|m) = \frac{\pi(\Theta, m)\rho(\pi)}{\sum_{\pi' \in \text{supp}(\rho)} \pi'(\Theta, m)\rho(\pi')} $$ \hspace{1cm} (4)

The new hypothesis $\pi_{m^*}$ is used to derive a receiver’s posterior belief over $\Theta$ via Bayes’ rule.\(^9\) We call this updating procedure the Hypothesis Testing updating rule. It is formally

\(^9\)As mentioned in the Introduction, we consider a special case of the Hypothesis Testing model. In
Definition 3 (Hypothesis Testing Updating Rule) Let $\rho$ be a probability distribution over $\Pi$ (i.e., a second-order prior) such that $|\text{supp}(\rho)| \in \mathbb{N}$. The family of posterior beliefs $\mu_\pi = \{\mu_\pi(\cdot|m)\}_{m \in \mathcal{M}}$ over $\Theta$ is derived via the Hypothesis Testing updating rule if $\mu_\pi$ satisfies:

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

(ii) $\mu_\pi(\theta|m) = \frac{\pi_{m}^*(\theta,m)}{\pi_{m}^*(\Theta,m)}$ if $\pi^*(\Theta,m) = 0$, where $\{\pi_{m}^*\} \coloneqq \arg\max_{\pi \in \text{supp}(\rho)} \rho(\pi|m)$.

According to the Hypothesis Testing updating rule, posterior beliefs are well-defined if for each message $m \in \mathcal{M}$, there is a hypothesis $\pi \in \text{supp}(\rho)$ that is consistent with $m$ (i.e., $\pi(\Theta,m) > 0$).

Given the updating procedure, we can introduce the Hypothesis Testing Equilibrium (henceforth, HTE). An HTE consists of a strategy profile $(s^*, r^*)$, a second-order prior $\rho$, and a family of posterior beliefs $\mu_\pi = \{\mu_\pi(\cdot|m)\}_{m \in \mathcal{M}}$ derived via the Hypothesis Testing updating rule. Formally,

Definition 4 (Hypothesis Testing Equilibrium) $(s^*, r^*, \rho, \mu_\pi^*)$ is a Hypothesis Testing Equilibrium (HTE) for a signaling game in $\mathcal{G}$ if:

(i) $s^*(\theta) \in \arg\max_{m \in \mathcal{M}} u(\theta,m,r^*(m))$ for each $\theta \in \Theta$,

(ii) $r^*(m) \in \arg\max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \mu_{\pi}^*(\theta|m)u_R(\theta,m,a)$ for each $m \in \mathcal{M}$,

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

$$\pi^*(\theta,m) = \begin{cases} p(\theta), & \text{if } s^*(\theta) = m, \\ 0, & \text{otherwise}, \end{cases}$$

(iv) $\mu_{\pi}^*(\theta|m) = \frac{\pi_{m}^{**}(\theta,m)}{\pi_{m}^{**}(\Theta,m)}$ if $\pi^*(\Theta,m) = 0$, where $\{\pi_{m}^{**}\} \coloneqq \arg\max_{\pi \in \text{supp}(\rho)} \rho(\pi|m)$.

Ortoleva (2012), an agent rejects her initial hypothesis if the observed event has a probability equal or smaller than a threshold $\epsilon \geq 0$. In our setup, the receiver rejects her initial hypothesis after a zero probability event is observed (i.e., $\epsilon = 0$).
Conditions (i) and (ii) ensure sequential rationality (as in PBE). Conditions (iii) and (iv) ensure that posterior beliefs are well-defined for all messages, including out-of-equilibrium messages. In HTE, the receiver best responds to each message with respect to a posterior belief derived from a hypothesis about a sender’s behavior that, in her view, explains the message.

The initial hypothesis \( \pi^* \) is a receiver’s belief that the sender plays her equilibrium strategy \( s^* \). This hypothesis is used to derive her posterior beliefs along the equilibrium path (i.e., for each \( m \in M \) with \( \pi^*(\Theta, m) > 0 \)). If, however, an out-of-equilibrium message is observed (i.e., \( m^o \in M^o \) such that \( \pi^*(\Theta, m^o) = 0 \)), the initial hypothesis is rejected. Then, a new hypothesis \( \pi_{m^o}^{**} \) that is consistent with \( m^o \) is selected. The new hypothesis is another receiver’s belief that assigns a strictly positive probability to a (non-equilibrium) strategy \( s \) that generates \( m^o \) (i.e., \( \beta(s) > 0 \) for \( s \in S \) such that \( s(\theta) = m \) for some \( \theta \), implying that \( \pi^{**}(\Theta, m^o) > 0 \)). The new hypothesis \( \pi_{m^o}^{**} \) is used to derive a receiver’s out-of-equilibrium belief given \( m^o \). In the example below, we illustrate the HTE notion.

**Example 2B.** Consider the set of simple hypotheses \( \{\pi_1, \pi_2, \pi_3, \pi_4\} \) presented in Example 2A. There are two pooling HTEs in the labor-market game of Figure 1.

**HTE-1:** In the first HTE, both worker types signal \( E \). That is, the strategy profile

\[
s^*(\theta_L) = s^*(\theta_H) = E, \quad r^*(E) = e, \quad r^*(N) = m,
\]

together with \( \text{supp}(\rho) = \{\pi_1, \pi_3\} \) such that \( \rho(\pi_1) < \rho(\pi_3) \) forms the first pooling HTE. Initially, the employer selects the pooling hypothesis \( \pi_3 \) according to which the employer believes that both worker types signal \( E \) (i.e., \( \pi^* = \pi_3 \)). Updating of \( \pi_3 \) given \( E \) yields the prior information \( p \). However, when \( N \) is observed, the pooling hypothesis \( \pi_3 \) is rejected. The employer chooses \( \pi_1 \) according to which she believes that workers separate. (i.e., \( \pi_{N}^{**} = \pi_1 \)). By updating \( \pi_1 \) given \( N \), the employer concludes that \( N \) is the signal of low skills (i.e., \( \mu^*_p(\theta_L|N) = 1 \)).

**HTE-2:** In the second HTE, both worker types signal \( N \). That is, the strategy profile

\[
s^*(\theta_L) = s^*(\theta_H) = N, \quad r^*(E) = m, \quad r^*(N) = e,
\]

together with \( \text{supp}(\rho) = \{\pi_2, \pi_4\} \) such that \( \rho(\pi_2) < \rho(\pi_4) \) forms the second pooling HTE. The initial hypothesis is the pooling hypothesis \( \pi_4 \). However, when \( E \) is observed, \( \pi_4 \) is discarded and \( \pi_2 \) is selected (i.e., \( \pi_{E}^{**} = \pi_2 \)). That is, the employer believes that workers “reversely” separate; i.e., the high-skilled worker signals \( N \) while the low-skilled worker signals \( E \). By updating \( \pi_2 \) given \( E \), the employer thus concludes that education must be the signal of high

12
skills (i.e., $\mu_\pi^*(\theta_L|E) = 1$).

Notice that HTE-1 and HTE-2 explain the pooling behaviors of PBE-1 and PBE-2. □

Our first main result shows that, in fact, each PBE can be explained by an HTE.

**Theorem 1** Let $(s^*, r^*, \mu^*)$ be a PBE and $\mathcal{M}^o$ be the set of out-of-equilibrium messages. Then, there exists an HTE, $(s^*, r^*, \rho, \mu_\pi^*)$, that supports the PBE, i.e.,

1. $\mu_\pi^*(\cdot|m) = \mu^*(\cdot|m)$ for each equilibrium message $m$,
2. $\mu_\pi^*(\cdot|m^o) = \mu^*(\cdot|m^o)$ for each out-of-equilibrium message $m^o \in \mathcal{M}^o$.

Recall that Bayes' rule does not specify how out-of-equilibrium beliefs in PBE are determined. Therefore, there may be multiple beliefs that rationalize the receiver’s (equilibrium) best response to $m^o$. If we consider a PBE, $(s^*, r^*, \mu^*)$, together with the family $\{\mathcal{O}(m^o)\}_{m^o \in \mathcal{M}^o}$ of sets of out-of-equilibrium beliefs, Theorem 1 proves that there exists an HTE supporting the PBE for each belief in $\mathcal{O}(m^o)$. That is, $(s^*, r^*, \mu^*)$ for each $\mu^*(\cdot|m^o) \in \mathcal{O}(m^o)$ can be explained by an HTE.

Since each HTE is PBE by Definition 4, we conclude that both concepts are equivalent.

**Corollary 1** For signaling games, PBE and HTE are equivalent solution concepts.

The above results have an important implication. In PBE, out-of-equilibrium beliefs are arbitrarily determined. The HTE notion provides an explanation for the origin of PBE beliefs. In PBE, out-of-equilibrium beliefs might be seen as being derived from Ortoleva’s (2012) Hypothesis Testing model. That is, for each out-of-equilibrium belief of a PBE, there always exists a hypothesis (i.e., a receiver’s belief about a sender’s choice of (non-equilibrium) strategies) that justifies the belief.

Where do PBE beliefs come from? On the equilibrium path, beliefs are derived from the initial hypothesis according to which the sender plays her equilibrium strategy. Since the equilibrium strategy is pure, the initial hypothesis is simple. However, off-the-equilibrium-path beliefs are derived from hypotheses that do not need to be simple. That is, if an out-of-equilibrium message is observed, the receiver believes that the sender chooses from a set of pure strategies according to $\beta$. Since $\beta$ might be interpreted as a sender’s mixed strategy, the receiver believes that the out-of-equilibrium message is an outcome of mixed behavior, even though pure strategies are played in equilibrium. Theorem 1 shows that counter-factual reasoning about mixed behavior suffices to explain all PBE beliefs.\(^\text{10}\)

\(^\text{10}\)However, if we constrain HTE to simple hypotheses then there are games for which the constrained HTE does not exist even though a pure PBE exists. Therefore, counter-factual reasoning about pure strategies is not sufficient to explain each PBE. In Appendix D, we provide an example for a PBE that cannot be supported by a constrained HTE.
As mentioned before, the idea to justify beliefs by means of hypotheses about opponents’ behavior goes back to Kreps and Wilson (1982). In particular, Kreps and Wilson asserted – without proving it – that sequential-equilibrium beliefs are structurally consistent in extensive-form games. A belief at an information set is structurally consistent if there exists a single (behavioral) strategy under which the information set can be reached and from which the belief can be derived via Bayes’ rule.\footnote{However, Kreps and Ramey (1987) showed that structural consistency is too restrictive. They provided examples for sequential equilibria whose out-of-equilibrium beliefs are not structurally consistent. Kreps and Ramey (1987) showed that sequential-equilibrium beliefs can be justified by a weaker consistency notion, called convex structural consistency. That is, they showed that each out-of-equilibrium belief in a sequential equilibrium can be justified by updating a convex combination of probabilities derived from a finite set of (behavioral) strategies.} In our setup, structural consistency means that a posterior belief given a message can be derived from a hypothesis that is consistent with the message by updating it via Bayes’ rule. Since each hypothesis can be associated with a sender’s mixed strategy, and by the Kuhn Theorem, each mixed strategy has a payoff-equivalent behavioral strategy, Theorem 1 proves for signaling games that PBE beliefs are structurally consistent in the sense of Kreps and Wilson (1982).\footnote{In Appendix B, we show that Theorem 1 can be extended to PBE in behavioral strategies. That is, we prove that beliefs of each (behavioral) PBE are structurally consistent in the sense of Kreps and Wilson (1982). This extension shows that structural consistency is sufficient to justify sequential-equilibrium beliefs in signaling games.}

There are two crucial differences to Kreps and Wilson (1982). First of all, besides proving that beliefs are structurally consistent, we require each PBE belief to be consistent with the prior information about types that is a primitive in any signaling game. Second, we require that hypotheses are selected via the Hypothesis Testing model of Ortoleva (2012). That is, the receiver chooses most likely hypotheses according to her prior (resp., updated prior) over hypotheses. In contrast, Kreps and Wilson (1982) informally used a sequence of hypotheses with a lexicographic belief system.

In the next section, we apply hypotheses as a formal language to reason about out-of-equilibrium beliefs. That is, our goal is to strengthen the hypothesis notion in order to reduce the number of PBE beliefs (or equivalently, the number of HTEs).

## 4 Strong Hypothesis Testing Equilibrium

In this section, we introduce a stronger notion of HTE, and suggest a refinement criterion for PBE.

In HTE, beliefs are derived from hypotheses about a sender’s behavior. The initial hypothesis is about a sender’s rational behavior. This hypothesis states that the sender plays his (equilibrium) strategy that best responds to the receiver’s (equilibrium) strategy. How-
ever, new hypotheses do not need to be about sender’s best-response strategies. Evidently, strategies that generate an out-of-equilibrium message must differ from the sender’s equilibrium strategy and thus some of them may be irrational.

From our previous result, we know that there are too many hypotheses. As a consequence, each PBE belief can be backed up by a belief about a sender’s choice of (non-equilibrium) strategies.

To narrow down the plethora of PBE beliefs, we require out-of-equilibrium beliefs to be justified by a sender’s rational behavior. For this purpose, we introduce a more stringent hypothesis notion. Form now on, we require that hypotheses are about sender’s best-response strategies, called strong hypotheses. More precisely, a strong hypothesis is a receiver’s belief about a sender’s choice of (pure) strategies, each best responds to some receiver’s strategy.

Denote by $B \subseteq S$ the set of sender’s best-response strategies for a signaling game in $G$. That is, for each $s \in B$, there exists a receiver’s strategy $r: M \to A$ such that, for each $\theta \in \Theta$:

$$s(\theta) = m \in \arg \max_{m' \in M} u_S(\theta, m', r(m')).$$

A belief $\beta$ on $B$ together with the prior information $p$ about types defines a strong hypothesis.

**Definition 5 (Strong Hypothesis)** A probability distribution $\pi$ on $\Theta \times M$ represents a strong hypothesis if there exists a probability measure $\beta$ on $B$ such that, for every $(\theta, m) \in \Theta \times M$,

$$\pi(\theta, m) = \sum_{s \in B, s.t. s(\theta) = m} \beta(s) p(\theta).$$

A strong hypothesis is called simple, if $\beta$ is a degenerate belief on $B$ (i.e., $\beta(s) = 1$ for $s \in B$ and $\beta(s') = 0$ for any $s' \neq s$). According to a simple-strong hypothesis $\pi$, the receiver believes that her opponent chooses a single strategy $s$ that best responds to some of her strategy $r: M \to A$.

**Example 3A.** In the labor-market game depicted in Figure 1, the worker has three best responses. That is, $B = \{s_1, s_2, s_3\}$ where

1) $s_1 := \{s(\theta_L) = N, s(\theta_H) = E\}$ is the best response against $r(E) = m$ and $r(N) = m$,

2) $s_2 := \{s(\theta_L) = N, s(\theta_H) = N\}$ is the best response against $r(E) = m$ and $r(N) = e$,

3) $s_3 := \{s(\theta_L) = E, s(\theta_H) = E\}$ is the best response against $r(E) = e$ and $r(N) = m$.

A degenerate belief $\beta$ on $B$ combined with $p$ induces the following simple-strong hypotheses:
1) \( \pi_1 := \{ \pi_1(\theta_L, N) = 1/3, \pi_1(\theta_H, E) = 2/3 \} \) if \( \beta(s_1) = 1 \),

2) \( \pi_2 := \{ \pi_2(\theta_L, N) = 1/3, \pi_2(\theta_H, N) = 2/3 \} \) if \( \beta(s_2) = 1 \),

3) \( \pi_3 := \{ \pi_3(\theta_L, E) = 1/3, \pi_3(\theta_H, E) = 2/3 \} \) if \( \beta(s_3) = 1 \).

An HTE with \( \text{supp}(\rho) \) that contains strong hypotheses is called Strong HTE.

**Definition 6 (Strong Hypothesis Testing Equilibrium)** \((s^*, r^*, \rho, \mu^*_\pi)\) is a Strong Hypothesis Testing Equilibrium for a signaling game in \( G \) if \((s^*, r^*, \rho, \mu^*_\pi)\) satisfies conditions (i), (ii) and (iii) in Definition 4 and \( \text{supp}(\rho) \) only contains strong hypotheses.

It is immediate that Strong HTE is a more stringent solution than PBE (or equivalently, HTE). Therefore, we use the Strong HTE that supports a given PBE as a refinement criterion for the PBE.

At this stage, we should remark that Strong HTE is more general than the equilibrium originally suggested by Ortoleva (2012). However, Ortoleva’s equilibrium coincides with Strong HTE if each in \( \text{supp}(\rho) \) is assumed to be a simple-strong hypothesis.\(^{13}\)

Our first refinement is defined as follows. For a signaling game in \( G \), a given PBE is said to pass the Strong Hypothesis Testing (HT) refinement if there exists a Strong HTE supporting the PBE. If, however, there is no Strong HTE that can explain the PBE, the equilibrium fails the refinement.

**Definition 7 (Strong Hypothesis Testing Refinement)** A PBE, \((s^*, r^*, \mu^*)\), is said to pass the Strong Hypothesis Testing refinement if there exists a Strong HTE, \((s^*, r^*, \rho, \mu^*_\pi)\), that supports the PBE. That is, there exist \( \rho \) and a family of posterior beliefs \( \mu^*_\pi := \{ \mu^*_\pi(\cdot | m) \}_{m \in M} \) such that

(i) \( \mu^*_\pi(\cdot | m) = \mu^*(\cdot | m) \) for each equilibrium message \( m \), and

(ii) \( \mu^*_\pi(\cdot | m^o) = \mu^*(\cdot | m^o) \), for each out-of-equilibrium message \( m^o \in M^o \).

For a PBE, \((s^*, r^*, \mu^*)\), the algorithm that verifies our refinement criterion operates in two steps. In the first step, for each \( m^o \in M^o \), we verify if there exists a strong hypothesis that is consistent with the out-of-equilibrium message (i.e., \( \pi \) such that \( \pi(\Theta, m^o) > 0 \)). Let \( \Pi_{m^o} \) be the set of such hypotheses. If \( \Pi_{m^o} \) is a non-empty set then the second step is applied. In the

\(^{13}\)Sun (2016) proves the existence of Strong HTE under simple-strong hypotheses (i.e., for Ortoleva’s (2012) equilibrium) by imposing restrictions on players’ payoff structure. Our approach is different. In our setup, we enlarge the set of conceivable hypotheses (see Definition 2) and show that a (pure) HTE exists for any signaling game for which a (pure) PBE exists (see Theorem 1). Therefore, our approach does not restrict players’ payoff structure.
second step, we verify if there is a strong hypothesis $\Pi_{m^o}$ that justifies the out-of-equilibrium belief of the PBE; i.e.,

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

Accordingly, a PBE might fail the Strong HT refinement for two reasons. First, there is an out-of-equilibrium message $m^o \in M^o$ for which $\Pi_{m^o}$ is an empty set. This means that $m^o$ cannot be generated by a sender’s best-response behavior.\footnote{For instance, an out-of-equilibrium message that is a dominated action for every type $\theta \in \Theta$ cannot be generated by a best-response strategy.} Second, $\Pi_{m^o}$ is a non-empty set. However, none of the strong hypotheses induces the out-of-equilibrium belief of the PBE. That is, for all $\pi \in \Pi_{m^o}$, $\mu_\pi(\cdot|m^o) \neq \mu^*(\cdot|m^o)$. This means that the out-of-equilibrium belief $\mu^*(\cdot|m^o)$ is inconsistent with the prior information about types under a sender’s best-response behavior that generates $m^o$.\footnote{As Example 3B shows, PBE-1 fails the Strong HT refinement for this reason.}

Consider a PBE, $(s^*, r^*, \mu^*)$, together with the family $O(m^o)$ of all out-of-equilibrium beliefs. The Strong HT refinement is said to reduce the number out-of-equilibrium beliefs of the PBE, if $(s^*, r^*, \mu^*)$ passes the Strong HT refinement for some beliefs in $O(m^o)$ but not all of them.

The following examples illustrate how the Strong HT refinement works.

**Example 3B.** Recall the two pooling equilibria for the game of Figure 1 (i.e., PBE-1 and PBE-2). Consider the set of best responses $B = \{s_1, s_2, s_3\}$ and the simple-strong hypotheses $\pi_1, \pi_2$ and $\pi_3$ presented in Example 3A. We show that PBE-1 but not PBE-2 passes the Strong HT refinement.

Consider the PBE with pooling on $E$ (i.e., PBE-1). Let $\lambda \in [0, 1]$ be a parameter and $\beta$ be the employer’s belief on $B$ such that $\beta(s_1) = \lambda$ and $\beta(s_2) = (1 - \lambda)$. Each $\beta$ together with the prior information $p$ induces the following strong hypothesis:

$$\pi(\lambda) := \{\pi(\theta_L, N) = 1/3, \pi(\theta_L, E) = 0, \pi(\theta_H, N) = (1 - \lambda)2/3, \pi(\theta_H, E) = \lambda2/3\}.$$ 

For each $\lambda \in [1/2, 1]$, by updating $\pi(\lambda)$ given $N$ via Bayes’ rule, we get

$$\mu_\pi(\theta_L|N) = \frac{1/3}{1/3 + (1 - \lambda)2/3} \geq 1/2.$$ 

Therefore, for each $\lambda \in [1/2, 1]$, the strategy profile $s^*(\theta_L) = s^*(\theta_H) = E$, $r^*(E) = e$, $r^*(N) = m$ and $\text{supp}(\rho) = \{\pi(\lambda), \pi_3\}$ such that $\rho(\pi(\lambda)) < \rho(\pi_3)$ form a Strong HTE.
that supports PBE-1. Therefore, PBE-1 passes the Strong HT refinement.\footnote{Notice that each out-of-equilibrium belief of PBE-1 can be justified by a strong hypothesis. Thus, the refinement does not reduce the number of out-of-equilibrium beliefs in PBE-1. However, if we assume simple-strong hypotheses as Ortoleva (2012) does (i.e., $\lambda = 1$), then the Strong HT refinement yields a unique belief $\mu^*_\pi(\theta_L|N) = 1$.}

Now, consider the PBE with pooling on $N$ (i.e., PBE-2). The pooling hypothesis $\pi_2$ rationalizes the employer’s on-the-equilibrium-path behavior. However, given the out-of-equilibrium message $E$, there is no strong hypothesis that can rationalize the employer’s decision to match the worker with the manual job (i.e., $r^*(N) = m$). To see this, for $\gamma \in [0,1]$, let $\beta$ be the employer’s belief on $B = \{s_1, s_2, s_3\}$ such that $\beta(s_1) = \gamma$ and $\beta(s_3) = (1 - \gamma)$. Then, $\beta$ together with $p$ induces

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

For each $\gamma \in [0,1]$, updating of $\pi(\gamma)$ given $E$ delivers

$$\mu^*_{\pi}(\theta_L|E) = \frac{(1 - \gamma) 1/3}{1 - \gamma 1/3} \leq 1/3.$$ 

Thus, the receiver concludes that education is more likely to be chosen by the high-skilled worker, and the worker is matched with the executive job $e$. Hence, PBE-2 fails the Strong HT refinement.

In the next example, we show that our refinement reduces the number of PBE beliefs.

**Example 4** Consider a modified version of the previous labor-market game depicted in Figure 1. The employer gets $115^*$ by matching the high-skilled worker signaling $N$ with the manual job $m$.

There is one pooling PBE in which both worker types acquire education. The employer assigns the executive job $e$ when $E$ is observed and the manual job $m$ when $N$ is observed. That is,

$$s^*(\theta_L) = s^*(\theta_H) = E \text{ and } r^*(E) = e, r^*(N) = m,$$

and posterior beliefs $\mu^*(\theta_L|E) = 1/3$ and $\mu^*(\theta_L|N) \geq 1/6$ form the pooling PBE. This equilibrium passes the Strong HT refinement. Notice that the set of strong hypotheses is the same as in Example 3A. For each $\lambda \in [0,1]$, let $\beta$ be the employer’s belief on $B = \{s_1, s_2, s_3\}$ such that $\beta(s_1) = \lambda$ and $\beta(s_2) = (1 - \lambda)$. Then, $\beta$ together with $p$ induces the following strong hypothesis:

$$\pi(\lambda) := \{\pi(\theta_L, N) = 1/3, \pi(\theta_L, E) = 0, \pi(\theta_H, N) = (1 - \lambda) 2/3, \pi(\theta_H, E) = \lambda 2/3\}.$$
By updating $\pi(\lambda)$ given $N$, we have

$$
\mu_\pi^*(\theta_L|N) = \frac{1/3}{1/3 + (1 - \lambda)2/3} \geq 1/3.
$$

A Strong HTE with $\text{supp}(\rho) = \{\pi(\lambda), \pi_3\}$, such that $\rho(\pi(\lambda)) < \rho(\pi_3)$ supports the pooling PBE. Moreover, since $\mu_\pi^*(\theta_L|N) \geq 1/3$, the Strong HT refinement reduces the number of PBE beliefs. If $\lambda = 1$ (i.e., $\pi(\lambda)$ is a simple-strong hypothesis), the Strong HT refinement yields the out-of-equilibrium belief according to which the employer learns the worker’s type, i.e., $\mu_\pi^*(\theta_L|N) = 1$. □

One remark is in order. As in Ortoleva (2012), we could constrain our refinement by requiring strong hypotheses to be simple (i.e., the receiver believes that the sender plays a singly pure strategy). In this case, the Strong HT refinement will always reduce out-of-equilibrium beliefs. Since there are finite types and messages, there is a finite number of simple-strong hypotheses. Therefore, the Strong HT refinement will select a finite collection of out-of-equilibrium beliefs.

Summing up, we have shown that Strong HTE can be used to refine the multiple PBE beliefs. In the next section, we compare the Strong HT refinement with the well-known Intuitive Criterion.

## 5 Strong HT Refinement versus Intuitive Criterion

In this section, we take for granted that posterior beliefs are derived from strong hypotheses. Our goal is to compare the Strong HT refinement with the Intuitive Criterion of Cho and Kreps (1987).

Contrary to our refinement approach, the Intuitive Criterion does not build on a theory of belief updating. Instead, the Intuitive Criterion is a payoff-based refinement. That is, the idea is to select out-of-equilibrium beliefs that are “plausible” in the sense of assigning zero probability to sender’s types that cannot be better off by deviating from the equilibrium strategy in a PBE.

Let us briefly recall the Intuitive Criterion. For a PBE, $(s^*, r^*, \mu^*)$, consider an out-of-equilibrium message $m^\circ \in \mathcal{M}^\circ$. Denote by $u^*_S(\theta)$ the sender’s equilibrium payoff when his type is $\theta \in \Theta$ (i.e., $u^*_S(\theta) = u_S(\theta, s^*(\theta), r^*(m))$). Let $T(m^\circ) \subseteq \Theta$ be the set of types that cannot improve upon their equilibrium payoff by sending message $m^\circ$. That is, for all $\theta \in T(m^\circ)$:

$$
 u^*_S(\theta) > \max_{a \in \mathcal{R}(\theta, m^\circ)} u_S(\theta, m^\circ, a), \tag{7}
$$
where
\[ BR(\Theta, m^o) := \bigcup_{\mu : \mu(\Theta|m^o) = 1} BR(\mu, m^o) \] (8)
is a set of receiver's best replies against \( m^o \) with respect to beliefs concentrated on types in \( \Theta \).

Denote by \( I(m^o) = \Theta \setminus T(m^o) \) the set of types who could be better off than their equilibrium payoff by choosing the out-of-equilibrium message \( m^o \). Let
\[ BR(I(m^o), m^o) := \bigcup_{\mu : \mu(I(m^o)|m^o) = 1} BR(\mu, m^o) \] (9)
be the set of receiver's best replies against \( m^o \) with respect to posterior beliefs defined over \( I(m^o) \). If there exists a type \( \theta' \in \Theta \) such that
\[ u^*_S(\theta') < \min_{a \in BR(I(m^o), m^o)} u_S(\theta', m^o, a), \] (10)
then the PBE fails the Intuitive Criterion. If, however, the PBE passes the Intuitive Criterion, the out-of-equilibrium beliefs are refined by admitting only probability distributions over \( \Theta \) that assign a zero probability to each type in \( T(m^o) \) (i.e., \( \mu(\theta|m^o) = 0 \) for all \( \theta \in T(m^o) \)).

If \( I(m^o) \) is a singleton (i.e., \( I(m^o) = \{\theta_{m^o}\} \) for some \( \theta_{m^o} \in \Theta \)), the receiver learns the type that could benefit from sending the out-of-equilibrium message \( m^o \) (i.e., \( \mu(\theta_{m^o}|m^o) = 1 \)). A class of signaling games in which \( I(m^o) \) is a singleton set for every out-of-equilibrium message \( m^o \) is important for economic applications. In such games, the Intuitive Criterion outcome is unique.

The example below illustrates the Intuitive Criterion outcome for the game of Figure 1.

**Example 1B.** Recall the first pooling PBE with pooling on \( E \) (i.e., PBE-1). Given the equilibrium payoff, the low-skilled type could be better off by sending the out-of-equilibrium message \( N \) of he believes that he will be matched with the executive job. That is, \( I(N) = \{\theta_L\} \) and \( T(N) = \{\theta_H\} \). As long as the employer believes that \( N \) is sent by the low-skilled worker (i.e., \( \mu^*(\theta_L|N) = 1 \)), there is no type that has an incentive to signal \( N \). Therefore, PBE-1 passes the Intuitive Criterion yielding the unique out-of-equilibrium belief \( \mu^*(\theta_L|N) = 1 \).

\(^{17}\)More specifically, \( BR(\mu, m^o) := \arg \max_{a \in A} \sum_{\theta \in \Theta} \mu(\theta|m^o) u_R(\theta, m^o, a) \).

\(^{18}\)PBE-2 fails the Intuitive Criterion. Only the high-skilled type can benefit from sending the out-of-equilibrium message \( E \). Therefore, the employer learns that \( E \) is sent by the high-skilled worker (i.e., \( \mu^*(\theta_H|E) = 1 \)) and best responds with \( e \). This, however, causes the high-skilled worker to signal \( E \) instead of \( N \).
Notice that the Intuitive Criterion outcome can be justified by a Strong HTE (see Example 3B). Suppose that the employer believes that workers separate after the out-of-equilibrium message $N$ is observed. That is, she selects the simple-strong hypothesis $\pi_1$ (i.e., $\lambda = 1$ in Equation (4)). By updating $\pi_1$ given $N$, the employer infers that $N$ is sent by the low-skilled worker, yielding the same out-of-equilibrium belief as the Intuitive Criterion (i.e., $\mu^*_\pi(\theta_L|N) = 1$).

This observation raises the question of whether our refinement is nested with the Intuitive Criterion. As we show below, both refinement criteria are not nested in any sense. We can find a PBE that passes the Strong HT refinement but fails the Intuitive Criterion, and vice versa; i.e., there exists a PBE that passes the Intuitive Criterion but fails the Strong HT refinement.

**Proposition 1** The Strong HT refinement and the Intuitive Criterion are not nested.

Now, we consider equilibria that survive against the Intuitive Criterion. We refer to a PBE that passes the Intuitive Criterion as the intuitive PBE.\(^{19}\) Since a Strong HTE supporting a PBE might not exist, we derive a condition under which intuitive PBEs pass the Strong HT refinement. In particular, we prove that a Strong HTE supporting an intuitive PBE always exists if, for each out-of-equilibrium message $m^o \in M^o$, there exists a single type that could benefit from sending $m^o$. In this case, the Strong HT refinement justifies the same out-of-equilibrium belief as the Intuitive Criterion. According to the belief, the receiver learns the type. This result is stated below.

**Theorem 2** Consider a PBE, $(s^*, r^*, \mu^*)$. Suppose that the following conditions are satisfied:

(i) for each out-of-equilibrium message $m^o \in M^o$, $I(m^o)$ is a singleton, i.e., $I(m^o) = \{\theta_{m^o}\}$ for some $\theta_{m^o} \in \Theta$, and

(ii) the PBE passes the Intuitive Criterion.

Then, there exists a Strong HTE that supports the intuitive PBE. In particular, for each $m^o \in M^o$:

$$
\mu^*(\theta_{m^o}|m^o) = \mu^*_\pi(\theta_{m^o}|m^o) = 1. \quad (11)
$$

\(^{19}\)A PBE that fails the Intuitive Criterion is referred to as the unintuitive PBE.

In the labor-market game of Example 4, the pooling PBE passes the Intuitive Criterion. Since $T(N) = \{\theta_H\}$ and $I(N) = \{\theta_L\}$, the Intuitive Criterion justifies the out-of-equilibrium belief $\mu^*_\pi(\theta_L|N) = 1$.\(^{19}\)
There are two remarks in order. First, Theorem 2 derives a class of PBEs that can be supported by a Strong HTE with respect to simple-strong hypotheses. The single-type condition, \((i)\), guarantees that there is a best-response strategy \(s \in B\) that generates the out-of-equilibrium message \(m^o\) (i.e., \(s(\theta_{m^o}) = m^o\)). If the receiver believes that the sender follows such strategy, i.e. \(\beta(s) = 1\), her belief together with the prior \(p\) induces a simple-strong hypothesis \(\pi_{m^o}\) that is consistent with \(m^o\). After updating \(\pi_{m^o}\) given \(m^o\), the receiver infers that the out-of-equilibrium message is sent by \(\theta_{m^o}\). Thus, \(\pi_{m^o}\) justifies the out-of-equilibrium belief \(\mu^*(\theta_{m^o}|m^o) = 1\) of the intuitive PBE.

Second, the single-type condition implies that the Intuitive Criterion outcome is unique. However, the Strong HT refinement does not need to be unique. To guarantee uniqueness, an auxiliary condition is needed. If, in addition to the single-type condition, each out-of-equilibrium message \(m^o \in \mathcal{M}^o\) is a never-best response for each type that cannot be better off by sending \(m^o\) than the equilibrium payoff (i.e., for all \(\theta \in T(m^o)\)), then also the Strong HT refinement outcome is unique. Formally, the uniqueness condition is stated below.

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

\[
m^o \neq s(\theta) \in \arg \max_{m \in \mathcal{M}} u_S(\theta, m, r(m)),
\]

for any \(r : \mathcal{M} \rightarrow A\) and each \(\theta \in T(m^o)\), then the Strong HT refinement outcome is unique.

In a PBE in which sending an out-of-equilibrium message \(m^o\) is dominated by another message for each type that cannot be better off than his equilibrium payoff (i.e., for each \(\theta \in T(m^o)\)), the uniqueness condition (12) is naturally satisfied.

Finally, we should remark that the Intuitive Criterion has the same limitation as PBE if for some out-of-equilibrium message \(m^o\), \(I(m^o)\) contains more than one type. In this case, the Intuitive Criterion admits arbitrary beliefs over \(I(m^o)\). In the extreme case where \(I(m^o) = \Theta\), the Intuitive Criterion does not reduce the set of posterior beliefs at all. For this reason, a variety of stronger refinements have been suggested in the economics literature (see Cho and Kreps (1987), Banks and Sobel (1987), Mailath, Okuno-Fujiwara, and Postlewaite (1993),

---

\(^{20}\)For this reason, Theorem 2 extends Ortoleva’s (2012, Section IV) refinement approach for the Beer and Quiche game of Cho and Kreps (1987) to any signaling game in \(\mathcal{G}\).

\(^{21}\)In Online Appendix, we show that for each out-of-equilibrium message \(m^o\) of an intuitive PBE with \(|I(m^o)| \geq 1\) there exists a strong hypothesis that is consistent with \(m^o\) (see Lemma A1).

\(^{22}\)Notice that the converse of Theorem 2 does not work. For instance, consider PBE-II for the game in Figure 2. This equilibrium will be discussed in Section 6. PBE-II satisfies the single-type condition \((i)\) of Theorem 2. However, while PBE-II passes the Strong HT refinement, it fails the Intuitive Criterion.

\(^{23}\)However, a Strong HTE with a non-singleton set \(I(m^o)\) refines PBE beliefs (see an example in Appendix C).
Fudenberg and He (2017, 2018)). In Online Appendix, we connect the Strong HTE with other popular refinements for intuitive PBE.

6 Behaviorally Consistent Hypotheses

In this section, we recapitulate the Stiglitz-Mailath critique of the Intuitive Criterion. To justify posterior beliefs that are immune to their critique, we introduce behaviorally consistent hypotheses.

In Strong HTE, the initial hypothesis rationalizes the receiver’s behavior on the equilibrium path. However, this hypothesis is rejected if an out-of-equilibrium message is observed. Then, a new strong hypothesis that is consistent with the out-of-equilibrium message is selected. The new hypothesis rationalizes the receiver’s (equilibrium) best response to the out-of-equilibrium message.

So far, we have not required that new hypotheses also maintain the receiver’s behavior on the path. However, the receiver may select a new hypothesis which – after updating it along the equilibrium path – will rationalize a different action than her equilibrium best response. Such hypotheses are said to be behaviorally inconsistent with the initial hypothesis. Behavioral inconsistencies might be problematic. As we show below, a behaviorally inconsistent hypothesis might provide an argument against the hypothesis along similar lines as the argument of Stiglitz (see Cho and Kreps, 1987, p.203) and Mailath (1988) against the belief that underlies the Intuitive Criterion reasoning.

Let us briefly recapitulate the Stiglitz-Mailath critique.

Example 5A Consider another version of the labor-market game depicted in Figure 2.

There are two pooling PBEs.

PBE-I: In the first PBE, both worker types signal $E$. That is, the strategy profile

$$s^*(\theta_L) = s^*(\theta_H) = E, \quad r^*(E) = e, \quad r^*(N) = m$$

and posterior beliefs $\mu^*(\theta_L|E) = 1/3$, and $\mu^*(\theta_L|N) \geq 1/2$ form PBE-I.

PBE-II: In the second PBE, both worker types signal $N$. That is, the strategy profile

$$s^*(\theta_L) = s^*(\theta_H) = N, \quad r^*(E) = m, \quad r^*(N) = e,$$

and posterior beliefs $\mu^*(\theta_L|N) = 1/3$ and $\mu^*(\theta_L|E) \geq 1/2$ form PBE-II. □

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

However, the employer might reason further. Since the worker who signals \( E \) receives the executive job, the high-skilled type is strictly better off by signaling \( E \) instead of \( N \). If the employer reasons consistently, then she should infer that only the low-skilled worker signals \( N \). Therefore, she will best respond by matching the worker signaling \( N \) with the manual job \( m \). This will in turn induce the low-skilled worker to signal \( E \). This chain of reasoning provides an argument against the belief that the high-skilled worker signals \( E \) with certainty. (i.e., \( \mu(\theta_H | E) = 1 \)). As a consequence, we might discard the unintuitive PBE even though the belief used against the equilibrium is “implausible” itself. This is the essence of the Stiglitz-Mailath critique of the Intuitive Criterion.

To eliminate “implausible” beliefs, we require that new hypotheses are behaviorally consistent. Let \( \pi^* \) be a receiver’s initial hypothesis. A strong hypothesis is said to be behaviorally consistent with the initial hypothesis \( \pi^* \) if it rationalizes the same behavior as \( \pi^* \). In other words, behavioral consistency requires that new hypotheses also maintain the receiver’s behavior on the equilibrium path.

**Definition 8 (Behaviorally Consistent Hypothesis)** Let \( \pi^* \) be an initial hypothesis. A strong hypothesis \( \pi \) is behaviorally consistent with the initial hypothesis \( \pi^* \) if for each mes-

\[ 24 \]

PBE-I passes the Intuitive Criterion yielding the posterior belief \( \mu(\theta_L | N) = 1 \).
sage \( m \in \mathcal{M} \) such that \( \pi^*(\Theta, m) > 0 \) and for each action \( a^* \in A \) such that

\[
a^* \in \arg\max_{a \in A} \sum_{\theta \in \Theta} \frac{\pi^*(\theta, m)}{\pi^*(\Theta, m)} u_R(\theta, m, a),
\]

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

\[
a^* \in \arg\max_{a \in A} \sum_{\theta \in \Theta} \frac{\pi(\theta, m)}{\pi(\Theta, m)} u_R(\theta, m, a).
\]

In Strong HTE, new hypotheses do not need to be behaviorally consistent. To emphasize the more stringent condition for hypotheses, we introduce the notion of Behaviorally Consistent HTE.

**Definition 9 (Behaviorally Consistent Hypothesis Testing Equilibrium)** \((s^*, r^*, \rho, \mu^*_\pi)\) is a Behaviorally Consistent Hypothesis Testing Equilibrium for a signaling game in \( G \) if \((s^*, r^*, \rho, \mu^*_\pi)\) satisfies conditions (i), (ii) and (iii) in Definition 4, and \( \text{supp}(\rho) \) only contains strong hypotheses that are behaviorally consistent with the most likely hypothesis \( \pi^* \) according to \( \rho \).

Since behaviorally consistent hypotheses are included in the set of strong hypotheses, we can apply Behaviorally Consistent HTE as an additional refinement criterion for PBE beliefs. We define the Behaviorally Consistent HT refinement in the same way the Strong HT refinement is defined (see Definition 7).

A PBE that passes the Behaviorally Consistent HT refinement satisfies an important property. Its out-of-equilibrium beliefs are immune to the Stiglitz-Mailath critique. Therefore, the Behaviorally Consistent HTE supporting a given PBE that fails the Intuitive Criterion can be seen as an argument in favor of the PBE.\(^{25}\) The Behaviorally Consistent HT refinement is illustrated below.

**Example 5B.** PBE-II presented in Example 5A passes the Behaviorally Consistent HT refinement. Consider the following two best responses of the employer:\(^{26}\)

1) \( s'_1 := \{s(\theta_L) = N, s(\theta_H) = N\} \) is the best response against \( r(E) = m \) and \( r(N) = e \),

2) \( s'_2 := \{s(\theta_L) = E, s(\theta_H) = N\} \) is the best response against \( r(E) = m \) and \( r(N) = m \).

\(^{25}\)However, deriving conditions under which an unintuitive PBE can be supported by a Behaviorally Consistent HTE goes beyond the scope of this paper.

\(^{26}\)Altogether, there are four best responses, \( B = \{s'_1, s'_2, s'_3, s'_4\} \) and four simple-strong hypotheses in this game. The other best responses, \( s'_3 \) and \( s'_4 \) and the strong hypotheses, \( \pi'_3 \) and \( \pi'_4 \), are depicted in Example 5C.
A degenerate belief \( \beta \) on \( \{s'_1, s'_2\} \) induces the following simple-strong hypotheses:

1) \( \pi'_1 := \{\pi_1(\theta_L, N) = 1/3, \pi_1(\theta_H, N) = 2/3\} \) if \( \beta(s_1) = 1 \),
2) \( \pi'_2 := \{\pi_2(\theta_L, E) = 1/3, \pi_2(\theta_H, N) = 2/3\} \) if \( \beta(s_2) = 1 \).

The strategy profile \( s^*(\theta_L) = s^*(\theta_H) = N, \ r^*(E) = m, \ r^*(N) = e \) together with \( \rho \) such that \( \text{supp}(\rho) = \{\pi'_1, \pi'_2\} \) and \( \rho(\pi'_2) < \rho(\pi'_1) \) forms Behaviorally Consistent HTE-II with pooling on \( N \). Therefore, PBE-II passes the Behaviorally Consistent HT refinement.

By updating the initial hypothesis \( \pi'_1 \), the employer concludes that \( N \) is more likely to be chosen by the high-skilled type, and therefore the worker is matched with the executive job. If the out-of-equilibrium message \( E \) is observed, the employer selects \( \pi'_2 \). According to \( \pi'_2 \), the employer believes that workers “reversely” separate; i.e., the high-skilled worker has education while the low-skilled worker does not. The new hypothesis \( \pi'_2 \) induces the out-of-equilibrium belief \( \mu(\theta_L | E) = 1 \). By updating \( \pi'_2 \) given the equilibrium message \( N \), the employer infers that \( N \) is sent by the high-skilled worker. Therefore, her behavior on the path remains optimal under \( \pi'_2 \), showing that \( \pi'_2 \) is behaviorally consistent with the initial hypothesis \( \pi'_1 \). Since \( \pi'_2 \) is behaviorally consistent, the employer has no reason to deviate from her equilibrium strategy. At this stage, the inconsistency in reasoning about behaviors on the equilibrium path and off the equilibrium path is prevented. The belief \( \mu(\theta_L | E) = 1 \) justified by \( \pi'_2 \) is immune to the Stiglitz-Mailath critique. Thus, we argue that there is no reason to refute PBE-II with the out-of-equilibrium belief \( \mu(\theta_L | E) = 1 \).

Below, we derive a class of Strong HTEs that pass the Behaviorally Consistent HT refinement.

**Proposition 2** Consider a Strong HTE, \((s^*, r^*, \rho, \mu^*_\pi)\). If a receiver’s best response is a single-valued correspondence for each equilibrium message, i.e.,

\[
\{r^*(m)\} = \arg \max_{a \in A} \sum_{\theta \in \Theta} \mu^*_\pi(\theta|m)u_R(\theta,m,a) \quad \text{for each equilibrium message } m, \tag{15}
\]

then there exists a Behaviorally Consistent HTE that supports the Strong HTE.

One question remains open. Can we justify the Intuitive Criterion outcome by a Behaviorally Consistent HTE? As we have shown before, if a PBE passes the Intuitive Criterion and \( I(m^\circ) \) satisfies the single-type condition, then the PBE passes the Strong HT refinement yielding the same out-of-equilibrium 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.
Let $\pi$ be a probability distribution on $\Theta \times M$. Let $m$ be a message and $\theta$ be a type such that $\pi(\theta, m) > 0$. Another type $\theta^d \in \Theta$ signaling $m$ (i.e., $\pi(\theta^d, m) > 0$) is called a dummy for message $m$, if the receiver’s best response given $m$ with respect to $\pi$ remains unchanged under any probability distribution $\bar{\pi}$ on $\Theta \times M$ such that $\bar{\pi}(\theta, m) = \pi(\theta, m)$ for all $\theta \in \Theta \setminus \{\theta^d\}$ and $\bar{\pi}(\theta^d, m) = 0$ (i.e., under any $\bar{\pi}$ that ascribes zero probability to the event “type $\theta^d$ sends $m$”).

**Definition 10 (Dummy Types)** Let $(s^*, r^*, \mu^*)$ be a PBE and $\pi$ be a probability distribution on $\Theta \times M$ induced by the equilibrium strategy $s^*$, i.e., $\pi(\theta, m) = p(\theta)$ if, $s^*(\theta) = m$ and $\pi(\theta, m) = 0$, otherwise. Consider a type $\theta^d$ that sends an equilibrium message $m \in M$, i.e., $s^*(\theta^d) = m$. Type $\theta^d$ is called a dummy for $m$ if for any $\bar{a} \in A$ such that $\bar{a} \in \arg \max_{a \in A} \sum_{\theta \in \Theta} \frac{\pi(\theta, m)}{\pi(\Theta, m)} u_R(\theta, m, a)$, it is true that $\bar{a} \in \arg \max_{a \in A} \sum_{\theta \in \Theta} \frac{\bar{\pi}(\theta, m)}{\bar{\pi}(\Theta, m)} u_R(\theta, m, a)$, for any probability distribution $\bar{\pi}$ on $\Theta \times M$ such that

$$\bar{\pi}(\theta, m) = \begin{cases} \pi(\theta, m), & \text{if } \theta \in \Theta \setminus \{\theta^d\}, \\ 0, & \text{if } \theta = \theta^d. \end{cases}$$

Consider a PBE in which type $\theta^d$ sends an equilibrium message $m$ (i.e., $\pi(\theta^d, m) > 0$). If $\theta^d$ is a dummy for $m$, then the equilibrium message does not convey any relevant information about type $\theta^d$. That is, if the receiver considers a probability distribution $\bar{\pi}$ that assigns zero probability to the event “type $\theta^d$ sends $m$,” then her best response to $m$ with respect to $\bar{\pi}$ is the same as her equilibrium best response with respect to $\pi$. The existence of dummies leads to the following result.

**Proposition 3** Consider a pooling PBE that passes the Intuitive Criterion. Let $m^*$ be the pooling message. If $I(m^*)$ is a singleton for each $m^* \in M^*$, and $\theta^d \in I(m^*)$ is a dummy for the equilibrium message $m^*$, then the intuitive PBE passes the Behaviorally Consistent HT refinement.

In the example below, we illustrate the role of a dummy for an equilibrium message.

**Example 5C.** Consider PBE-I for the labor-market game of Figure 2 (see Example 5A). This equilibrium can be supported by a Behaviorally Consistent HTE. Consider the other two best responses:
3) $s'^3 := \{s(\theta_L) = E, s(\theta_H) = E\}$ is the best response against $r(E) = m$ and $r(N) = e$,

4) $s'^4 := \{s(\theta_L) = N, s(\theta_H) = E\}$ is the best response against $r(E) = e$ and $r(N) = e$.

A degenerate belief $\beta$ on $\{s'^3, s'^4\}$ induces the following simple-strong hypothesis:

3) $\pi'^3 := \{\pi_3(\theta_L, E) = 1/3, \pi_3(\theta_H, E) = 2/3\}$ if $\beta(s'^3) = 1$,

4) $\pi'^4 := \{\pi_4(\theta_L, N) = 1/3, \pi_4(\theta_H, E) = 2/3\}$ if $\beta(s'^4) = 1$.

The strategy profile $s^*(\theta_L) = s^*(\theta_H) = E$, $r^*(E) = e$, $r^*(N) = m$ together with $\rho$ such that $\text{supp}(\rho) = \{\pi'^3, \pi'^4\}$ and $\rho(\pi'^4) < \rho(\pi'^3)$ forms Behaviorally Consistent HTE-I with pooling on $E$. Therefore, PBE-I passes the Behaviorally Consistent HT refinement. □

Notice that the low-skilled type (i.e., $\theta_L$) is a dummy type for the equilibrium message $E$. That is, even if the employer would believe that the low-skilled worker does not acquire education (i.e. $\bar{\pi}(\theta_L, S) = 0$), it would not change her best response to the equilibrium message $E$. Therefore, the intuitive PBE-I can be supported be a Behaviorally Consistent HTE as asserted in Proposition 3.

In the next section, we show that Behaviorally Consistent HTE can much better explain the experimental findings of Brandts and Holt (1992, 1993) than the Intuitive Criterion.

7 Experimental Findings of Brandts and Holt

In this section, we show that our strongest solution concept is consistent with empirical findings. In particular, Behaviorally Consistent HTE can accommodate the experimental results of Brandts and Holt (1992, 1993).

Brandts and Holt (1992) conducted a series of experiments on behavior in the labor-market games analyzed throughout this paper (i.e., the games presented in Figure 1 and Figure 2.) Their goal was to test predictions of the Intuitive Criterion. Interestingly, their subjects behaved consistently with the predictions of our strongest solution concept, the Behaviorally Consistent HTE.

Consider the labor-market game in Figure 1. There are two equilibria: the intuitive PBE-1 with pooling on $E$ and the unintuitive PBE-2 with pooling on $N$ (see Examples 1A and 1B). Notice that PBE-1 but not PBE-2 passes the Behaviorally Consistent HT refinement.\footnote{In Example 3B, we have shown that PBE-1 passes the Strong HT refinement. Consider the Strong HTE with respect to the simple-strong hypothesis $\pi(1) = \pi_1$ (i.e., $\lambda = 1$). By updating $\pi_1$ given $N$, we have that $\mu_\pi(\theta_L | N) = 1$. Since the employer learns that $N$ is sent by the low-skilled type, the worker is matched with the manual job $m$. Therefore, $\pi_1$ rationalizes the employer’s behaviors on the equilibrium path and off the equilibrium path showing that $\pi_1$ is behaviorally consistent with $\pi_3$. Thus, PBE-1 passes the Behaviorally Consistent HT refinement.}

In the next section, we show that Behaviorally Consistent HTE can much better explain the experimental findings of Brandts and Holt (1992, 1993) than the Intuitive Criterion.
Therefore, our prediction is in line with the Intuitive Criterion. In their Treatment 1, Brandts and Holt (1992) observed that 102 out of 128 subjects’ decisions matched with the intuitive PBE-1 and only 7 decisions matched with the unintuitive PBE-2.28 This result is consistent with our prediction.

There is another interesting finding. Brandts and Holt (1992) analyzed the behaviors of senders. They found evidence supporting the new hypothesis \( \pi_2 \) of the Behaviorally Consistent HTE-1. According to the new hypothesis \( \pi_2 \), the low-skilled type signals \( N \) and the high-skilled type signals \( E \). The authors observed that 88 out of 88 high-skilled subjects signaled \( E \). However, 24 out of 44 low-skilled subjects sent 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).

Furthermore, a majority of subjects acting as receivers seemed to believe that \( N \) is sent by low-skilled types in accordance with \( \pi_2 \). Then, 17 out of 24 receivers who observed \( N \) responded with the equilibrium action \( m \). The observed type-dependence and receivers’ reply to \( N \) show that the new hypothesis \( \pi_2 \) is a reasonable explanation for the out-of-equilibrium belief and behavior.

Consider the labor-market game in Figure 2 with two equilibria; the intuitive PBE-I with pooling on \( E \) and the unintuitive PBE-II with pooling on \( N \) (see Example 5A). For this game, the predictions of our refinement and the Intuitive Criterion are different. Then, both equilibria pass the Behaviorally Consistent HT refinement (see Examples 5B, and 5C). Interestingly, a majority of subjects behaved consistently with the unintuitive PBE-II that is supported by the Behaviorally Consistent HTE-II. In Treatment 5 of Brandts and Holt (1992), 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.29 Since PBE-II passes the Behaviorally Consistent HT refinement, we conclude that our strongest solution concept can explain the subjects’ behavior better than the Intuitive Criterion.

Again, in their Treatment 5, Brandts and Holt (1992) found evidence for the new hypothesis \( \pi_2' \) that justifies the out-of-equilibrium belief of the Behaviorally Consistent HTE-II. According to \( \pi_2' \), senders “reversely” separate, i.e., the low-skilled type signals \( E \) while the high-skilled type signals \( N \). The authors found that 72 out of 99 high-skilled senders signaled \( N \) while 20 out of 45 low-skilled senders signaled \( E \). Notably, a significant number of

---

28 Brandts and Holt (1992, p.1358) summarize the results of Treatment 1 in Table 3 (see parts (a) and (c)).
29 Brandts and Holt (1992, p.1363) summarize the results of Treatment 5 in Table 4 (see parts (a) and (b)).
subjects acting as receivers believed that senders separate. Then, 24 out of 47 receivers who observed $E$ responded with $m$.

In another study, Banks, Camerer, and Porter (1994) found more evidence in favor of intuitive PBE. However, using similar games as Banks, Camerer, and Porter (1994), Brandts and Holt (1993) could not find unequivocal support for the intuitive PBE behavior. Instead, Brandts and Holt (1993) could replicate a similar pattern of equilibrium behaviors as reported by Brandts and Holt (1992). Over a series of games, a majority of subjects behaved consistently with an unintuitive PBE. Also, the experimental results of Brandts and Holt (1993) can be accommodated by our solution concept.

In sum, Behaviorally Consistent HTE can perform better than the Intuitive Criterion in experimental studies conducted by Brandts and Holt (1992, 1993). In light of these studies, we conclude that our strongest solution concept is empirically relevant.

8 Conclusion

In this paper, we have developed a theory that provides an explanation for the origin of PBE beliefs. In signaling games, PBE beliefs might be seen as being derived via Ortoleva’s (2012) updating rule. In particular, we have shown that PBE beliefs can be derived by selecting and updating hypotheses (i.e., receiver’s beliefs about a choice of sender’s strategies that generate the observed message).

Hypotheses provide a formal language to reason about out-of-equilibrium beliefs. Therefore, we imposed two behavioral restrictions on hypotheses in order to refine the number of PBE beliefs.

Our first refinement requires that out-of-equilibrium beliefs are derived from strong hypotheses. This criterion ensures that out-of-equilibrium beliefs are justified by a sender’s rational behavior. Our second refinement is even more stringent. It requires that out-of-equilibrium beliefs are derived from strong hypotheses that are behaviorally consistent with the receiver’s initial hypothesis. This criterion ensures that beliefs are immune to the Stiglitz-Mailath critique of the Intuitive Criterion.

Our strongest refinement criterion provides 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 is also consistent with experimental findings on equilibrium behavior. Therefore, we believe that our strongest refinement is worth
further exploration and application.

A Proofs

Proof of Theorem 1. Consider a (pure) PBE, \((s^*, r^*, \mu^*)\). Let \(\mathcal{M}^o\) be the set of out-of-equilibrium messages and \(\{\mu^*(\cdot|m^o)\}_{m^o \in \mathcal{M}^o}\) be the family of posterior beliefs of the PBE. The proof consists of three steps. In Step 1, we construct a hypothesis that induces the receiver’s on-the-equilibrium belief. In Step 2, for each out-of-equilibrium message \(m^o \in \mathcal{M}^o\), we construct a hypothesis that induces a receiver’s out-of-equilibrium belief. In Step 3, we construct an HTE that supports the given PBE.

Step 1: Since \(s^*\) is the sender’s equilibrium strategy, there exists a hypothesis \(\pi^*\) that justifies the receiver’s on-the-equilibrium belief. Consider a degenerate belief \(\beta\) such that \(\beta(s^*) = 1\) and \(\beta(s) = 0\) for any \(s \in \mathcal{S} \setminus \{s^*\}\) where \(\mathcal{S}\) is the set of sender’s strategies. This belief together with the prior information \(p\) about sender’s types induces the following hypothesis:

\[
\pi^*(\theta, m) = \begin{cases} 
  p(\theta), & \text{if } s^*(\theta) = m, \\
  0, & \text{otherwise.}
\end{cases} \tag{18}
\]

By condition \((iii)\) in Definition 1, hypothesis \(\pi^*\) induces the receiver’s belief on the equilibrium path. That is, for any \(m \in \mathcal{M}\) such that \(\pi^*(\Theta, m) > 0\), we have

\[
\mu^*(\theta|m) = \mu^*_\pi(\theta|m) := \frac{\pi^*(\theta, m)}{\pi^*(\Theta, m)} \text{ for each } \theta \in \Theta. \tag{19}
\]

Step 2: Fix an out-of-equilibrium message \(m^o \in \mathcal{M}^o\). We will show that there exists a hypothesis \(\pi_{m^o}\) which – after updating it given \(m^o\) – justifies the out-of-equilibrium belief \(\mu^*(\cdot|m^o)\). Without loss of generality, we assume that there are \(N\) types in \(\text{supp}(\mu^*(\cdot|m^o))\). That is, \(\text{supp}(\mu^*(\cdot|m^o)) = \{\theta_1, \ldots, \theta_N\}\). For each \(i \in \{1, \ldots, N\}\), we construct a strategy \(s_i\) as follows

\[
s_i(\theta) = \begin{cases} 
  m^o, & \text{if } \theta = \theta_i, \\
  m \in \mathcal{M} \setminus \{m^o\}, & \text{otherwise.}
\end{cases} \tag{20}
\]

Now, define a receiver’s belief \(\beta\) over \(\mathcal{S}\) such that

\[
\beta(s_i) = \left(\frac{\mu^*(\theta_i|m^o)}{p(\theta_i)}\right) / \left(\frac{\mu^*(\theta_1|m^o)}{p(\theta_1)} + \ldots + \frac{\mu^*(\theta_N|m^o)}{p(\theta_N)}\right) \text{ for each } i \in \{1, \ldots, N\}, \tag{21}
\]

and \(\beta(s) = 0\) for any \(s \in \mathcal{S} \setminus \{s_1, \ldots, s_N\}\). Then, \(\beta\) together with the prior information \(p\)
induces the following hypothesis: For every \((\theta, m) \in \Theta \times \mathcal{M}\),
\[
\pi_{m^o}(\theta, m) = \sum_{s \in S \text{ s.t. } s(\theta) = m} \beta(s)p(\theta) = \sum_{s_i \in \{s_1, \ldots, s_N\} \text{ s.t. } s_i(\theta) = m} \beta(s_i)p(\theta). \tag{22}
\]

By updating \(\pi_{m^o}\) given \(m^o\), we have
\[
\mu_\pi(\theta|m^o) = \frac{\pi_{m^o}(\theta, m^o)}{\pi_{m^o}(\Theta, m^o)} \quad \text{for every } \theta \in \Theta, \tag{23}
\]
where \(\pi_{m^o}(\theta, m^o)\) and \(\pi_{m^o}(\Theta, m^o)\) are defined as follows:
\[
\pi_{m^o}(\theta, m^o) = \mu^*(\theta|m^o) / \left( \frac{1}{\mu^*(\theta_1|m^o) p(\theta_1)} + \ldots + \frac{\mu^*(\theta_N|m^o)}{p(\theta_N)} \right), \tag{24}
\]
and
\[
\pi_{m^o}(\Theta, m^o) = \frac{1}{\mu^*(\theta_1|m^o) p(\theta_1)} + \ldots + \frac{\mu^*(\theta_N|m^o)}{p(\theta_N)}. \tag{25}
\]

By plugging \(\pi_{m^o}(\theta, m^o)\) and \(\pi_{m^o}(\Theta, m^o)\) from Equations (24) and (25) into Equation (23), we get an out-of-equilibrium belief \(\mu_\pi(\cdot|m^o)\) that coincides with the PBE belief \(\mu^*(\cdot|m^o)\). That is,
\[
\mu_\pi(\theta|m^o) = \mu^*(\theta|m^o) \quad \text{for each } \theta \in \Theta. \tag{26}
\]
This shows that there exists a hypothesis \(\pi_{m^o}\) that induces the out-of-equilibrium belief \(\mu^*(\cdot|m^o)\) of the PBE. Notice that the out-of-equilibrium message \(m^o\) was arbitrarily chosen. Therefore, for each out-of-equilibrium message \(m^o \in \mathcal{M}^o\), we can construct a hypothesis \(\pi_{m^o}\) that will induce the out-of-equilibrium belief \(\mu^*(\cdot|m^o)\). Let \(\{\pi_{m^o}\}_{m^o \in \mathcal{M}^o}\) be the family of such hypotheses.

**Step 3:** We can suitably choose a second-order prior \(\rho\) with \(\text{supp}(\rho) = \{\pi^*, \pi_{m^o}^{**}\}_{m^o \in \mathcal{M}^o}\) such that
\[
\{\pi^*\} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi), \tag{27}
\]
(i.e., \(\pi^*\) is the most likely hypothesis) and
\[
\{\pi_{m^o}^{**}\} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi|m^o) \quad \text{for each } m^o \in \mathcal{M}^o. \tag{28}
\]
(i.e., \(\pi_{m^o}^{**}\) is the most likely hypothesis after updating \(\rho\) given \(m^o \in \mathcal{M}^o\)). Therefore, there exists an HTE \((s^*, r^*, \rho, \mu^*_\pi)\) supporting the PBE \((s^*, r^*, \mu^*)\). ■
Proof of Proposition 1. We prove that the Strong HT refinement and the Intuitive Criterion are not nested in two steps. In Step 1, we derive a PBE that passes the Strong HT refinement but fails the Intuitive Criterion. In Step 2, we provide a PBE that fails the former but passes the latter criterion.

Step 1: Consider the labor-market game in Figure 2 that we analyzed throughout Section 6. Consider the PBE with pooling on $N$ (i.e., PBE-II). Recall, the strategy profile

$$s^*(\theta_L) = s^*(\theta_H) = N, \ r^*(E) = m, \ r^*(N) = e,$$

and posterior beliefs $\mu^*(\theta_L|N) = 1/3$ and $\mu^*(\theta_L|S) \geq 1/2$ form PBE-II.

First, we show that PBE-II passes the Strong HT refinement. Consider the strong hypotheses $\pi'_1$ and $\pi'_2$ that are depicted in Example 5B. Then, the strategy profile

$$s^*(\theta_L) = s^*(\theta_H) = N, \ r^*(E) = m, \ r^*(N) = e,$$

together with $\text{supp}(\rho) = \{\pi'_1, \pi'_2\}$ such that $\rho(\pi'_2) < \rho(\pi'_1)$ constitutes the Strong HTE supporting the pooling PBE. By updating $\pi'_2$ given $E$, we get the out-of-equilibrium belief $\mu^*_\pi(\theta_L|E) = 1$.

Now, we show that PBE-II fails the Intuitive Criterion. According to the Intuitive Criterion, $\theta_H$ could be better off than her equilibrium payoff if she sends 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, when the receiver learns that $E$ is sent by $\theta_H$, she will choose $e$ instead of $m$. Then, given that the receiver plays $e$ against $E$, $\theta_H$ will indeed choose $E$. Thus, PBE-II does not survive against the Intuitive Criterion.

Step 2: Consider the labor-market game depicted in Figure 3. Assume that $x = 4$. There exists a pooling equilibrium in which both types choose $N$. That is, the strategy profile

$$s^*(\theta_L) = s^*(\theta_H) = N, \ r^*(E) = m, \ r^*(N) = e,$$

and posterior beliefs $\mu^*(\theta_L|N) = 2/5$ and $\mu^*(\theta_L|S) \leq 1/4$ form the PBE with pooling on $N$.

First, we show that there does not exist a Strong HTE that can support this equilibrium behavior. Notice that the sender has only two best-response strategies. That is, $\mathcal{B} = \{s_1, s_2\}$ where

1) $s_1 = \{s_1(\theta_L) = E, \ s_1(\theta_H) = E\}$ is the best response against $r(E) = e$ and $r(N) = m$,
2) \( s_2 = \{ s_2(\theta_L) = N, \ s_2(\theta_H) = N \} \) is the best response against \( r(E) = m \) and \( r(N) = m \).

Consider a receiver’s belief \( \beta \) on \( B = \{ s_1, s_2 \} \) such that \( \beta(s_1) = \lambda \) and \( \beta(s_2) = (1 - \lambda) \) where \( \lambda \in [0, 1] \). Each \( \beta \) (resp., \( \lambda \)) induces the following strong hypothesis

\[
\pi(\lambda) = \{ \pi(\theta_L, N) = (1 - \lambda)(2/5, \ \pi(\theta_L, E) = \lambda(2/5, \ \pi(\theta_H, N) = (1 - \lambda)(3/5, \ \pi(\theta_H, E) = \lambda(3/5) \}.
\]

The family \( \{ \pi(\lambda) \}_{\lambda \in [0, 1]} \) depicts all strong hypotheses for this signaling game. However, none of the strong hypothesis rationalizes the receiver’s equilibrium best response to the out-of-equilibrium message \( E \). Then, for each \( \lambda \in [0, 1] \), the Bayesian update of \( \pi(\lambda) \) given \( E \) is

\[
\mu(\theta_L | E) = 2/5.
\]

However, given this posterior belief, the receiver will choose \( e \) instead of \( m \). Hence, there does not exist a Strong HTE supporting the pooling PBE. Therefore, the PBE fails the Strong HT refinement.

Now, we show that the pooling PBE passes the Intuitive Criterion. According to the Intuitive Criterion, both types \( \theta_L \) and \( \theta_H \) 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 \( \mu(\theta_L | E) \) such that \( \mu(\theta_L | E) \leq 1/4 \). We know that each player does not have an incentive to deviate from the given equilibrium strategy as long as \( \mu(\theta_L | E) \leq 1/4 \). Therefore, the pooling PBE passes the Intuitive Criterion.

\[\square\]

**Proof of Theorem 2.** Let \((s^*, r^*, \mu^*)\) be a (pure) PBE and \( \mathcal{M}^o \) be the set of out-of-
equilibrium messages of the PBE. Take an out-of-equilibrium message $m^o \in \mathcal{M}^o$. Let $r^o(m^o) = a^*$ be the receiver’s equilibrium action chosen in response to message $m^o$. By assumption (i), the set of sender’s types that have an incentive to deviate from the equilibrium strategy $s^*$ by sending $m^o$ contains only one type. That is, $I(m^o) = \{\theta_{m^o}\}$ for some $\theta_{m^o} \in \Theta$. Moreover, by assumption (ii), the PBE $(s^*, r^*, \mu^*)$ passes the Intuitive Criterion. This means that $a^*$ is the receiver’s best response to $m^o$ with respect to the posterior belief induced by the Intuitive Criterion, i.e., $\mu^*(\theta_{m^o}|m^o) = 1$.

Let $\mathcal{B}$ be the set of sender’s best-response strategies. In Step 1, we show that there exists $s^o \in \mathcal{B}$ that generates message $m^o$. Given $s^o$, we construct a strong hypothesis $\pi_{m^o}$ that is consistent with the out-of-equilibrium message $m^o$ (i.e., $\pi_{m^o}(\theta, m^o) > 0$ for some $\theta \in \Theta$), and that justifies the out-of-equilibrium belief induced by the Intuitive Criterion, i.e., $\mu^*(\theta_{m^o}|m^o) = 1$. In Step 2, we construct a Strong HTE that explains the intuitive PBE.

**Step 1:** By the single-type condition (i), there exists an action $a^o \in \mathcal{A}$ in response to $m^o$ such that
\[
u_S^*(\theta) \leq u_s(\theta, m^o, a^o) \quad \text{for} \quad \theta = \theta_{m^o}, \tag{29}\]
and
\[
u_S^*(\theta) > u_s(\theta, m^o, a^o) \quad \text{for} \quad \theta \in \Theta \setminus \{\theta_{m^o}\}, \tag{30}\]
where $u_S^*(\theta)$ is the sender’s equilibrium payoff when his type is $\theta \in \Theta$. That is, when the receiver chooses $a^o$ in response to $m^o$, only type $\theta_{m^o}$ might deviate from the equilibrium strategy $s^*$.

Now, we construct a receiver’s strategy $r^o : \mathcal{M} \rightarrow \mathcal{A}$ as follows:
\[
 r^o(m) = \begin{cases} 
 r^o(m), & \text{if} \ m \neq m^o, \\
 a^o, & \text{if} \ m = m^o. 
\end{cases} \tag{31}
\]
Then, by construction of $r^o$, only type $\theta_{m^o}$ best responds with message $m^o$ against $r^o$, i.e.,
\[
 m^o \in \arg\max_{m \in \mathcal{M}} u_s(\theta, m, r^o(m)) \quad \text{for} \quad \theta = \theta_{m^o}, \tag{32}\]
and
\[
 m^o \notin \arg\max_{m \in \mathcal{M}} u_s(\theta, m, r^o(m)) \quad \text{for} \quad \theta \in \Theta \setminus \{\theta_{m^o}\}. \tag{33}\]
Notice that the equilibrium strategy $s^*(\theta)$ is still a best response against $r^o$ for any $\theta \in \Theta \setminus \{\theta_{m^o}\}$. Therefore, there exists a best-response strategy $s^o$ against $r^o$ that generates
message \( m^o \). That is,

\[
s^o := \{ s^o(\theta_{m^o}) = m^o, \ s^o(\theta) = s^*(\theta) \text{ for any } \theta \in \Theta \setminus \{ \theta_{m^o} \} \}
\]
is a best response against \( r^o \).

Now, define a receiver’s belief \( \beta \) over \( B \) such that

\[
\beta(s^o) = 1 \quad \text{and} \quad \beta(s) = 0 \quad \text{for any } s \in B \setminus \{ s^o \}.
\]

This belief together with the prior information \( p \) induces the simple-strong hypothesis

\[
\pi_{m^o} := \{ \pi(\theta_{m^o}, m^o) = p(\theta_{m^o}), \ \pi(\theta, s^*(\theta)) = p(\theta) \text{ for any } \theta \in \Theta \setminus \{ \theta_{m^o} \} \}.
\]

Since \( p(\theta_{m^o}) > 0 \), \( \pi_{m^o} \) is consistent with \( m^o \). By updating \( \pi_{m^o} \) given \( m^o \), we then have

\[
\mu_{\pi}(\theta_{m^o}|m^o) = \frac{\pi_{m^o}(\theta_{m^o}, m^o)}{\pi_{m^o}(\Theta, m^o)} = 1,
\]

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

Since \( m^o \) was chosen arbitrary, we can construct in the same way a simple-strong hypothesis for any out-of-equilibrium message in \( \mathcal{M}^o \). That is, for any \( m^o \in \mathcal{M}^o \) there exists a simple-strong hypothesis \( \pi_{m^o} \) that justifies \( \mu^*(\theta_{m^o}|m^o) = 1 \), and thus rationalizes the receiver’s behavior off the path in the given PBE. Let \( \{ \pi_{m^o} \}_{m^o \in \mathcal{M}^o} \) be the collection of such simple-strong hypotheses.

**Step 2**: Now, we construct a simple-strong hypothesis \( \pi^* \) that rationalizes the receiver’s on-the-equilibrium-path behavior. Since \( s^* \) best responds to \( r^* \), a simple-strong hypothesis \( \pi^* \) trivially exists. More precisely, the degenerate belief \( \beta \) such that \( \beta(s^*) = 1 \) and \( \beta(s) = 0 \) for any \( s \in B \setminus \{ s^* \} \) together with the prior information \( p \) induces the simple-strong hypothesis \( \pi^* \) defined as

\[
\pi^*(\theta, m) = \begin{cases} 
p(\theta), & \text{if } s^*(\theta) = m, \\
0, & \text{otherwise.}
\end{cases}
\]

We can suitably choose a second-order prior \( \rho \) with \( \text{supp}(\rho) = \{ \pi^*, \pi^*_{m^o} \}_{m^o \in \mathcal{M}^o} \) such that

\[
\{ \pi^* \} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi),
\]

and

\[
\{ \pi^*_{m^o} \} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi|m^o) \quad \text{for each } m^o \in \mathcal{M}^o.
\]

Therefore, there exists a Strong HTE \( (s^*, r^*, \rho, \mu^*) \) supporting the intuitive PBE, \( (s^*, r^*, \mu^*) \).
\textbf{Proof of Corollary 2.} Let \((s^*, r^*, \mu^*)\) be a PBE and \(\mathcal{M}^o\) be the set of out-of-equilibrium messages of the PBE. The single-type condition states that for each \(m^o \in \mathcal{M}^o\), we have \(I(m^o) = \{\theta_m^o\}\) for some \(\theta_m^o \in \Theta\). From the proof of Theorem 2 (Step 1), we know that we can construct a family of simple-strong hypotheses \(\{\pi_{m^o}\}_{m^o \in \mathcal{M}^o}\); each \(\pi_{m^o}\) being consistent with \(m^o\) (i.e., \(\pi_{m^o}(\theta_m^o, m^o) > 0\) where \(I(m^o) = \{\theta_m^o\}\)). It remains to be shown that each \(\pi_{m^o}\) is the only simple-strong hypothesis consistent with \(m^o\). By Condition (12), there does not exist a receiver’s strategy \(r\) against which sending an out-of-equilibrium message \(m^o \in \mathcal{M}^o\) is a best response for some \(\theta \in T(m^o)\). 

Recall, \(T(m^o)\) is the set of types that cannot improve upon their equilibrium payoff by choosing \(m^o\). This implies that there does not exist a best-response strategy that can generate an out-of-equilibrium message \(m^o \in \mathcal{M}^o\) for some type in \(T(m^o)\). Equivalently, there does not exist a strong hypothesis \(\pi\) such that \(\pi(\theta, m^o) > 0\) for each \(\theta \in T(m^o)\) and \(m^o \in \mathcal{M}^o\). Therefore, for any strong hypothesis \(\pi_{m^o}\) that is consistent with \(m^o \in \mathcal{M}^o\), we must have that \(\pi_{m^o}(\theta_m^o, m^o) = \pi_{m^o}(\Theta, m^o) = p(\theta_m^o)\) where \(\{\theta_m^o\} = I(m^o)\).

\begin{equation}
\pi_{m^o}(\theta_m^o, m^o) = \pi_{m^o}(\Theta, m^o) = p(\theta_m^o) \quad \text{where} \quad \{\theta_m^o\} = I(m^o).
\end{equation}

In other words, by updating any such strong hypothesis \(\pi_{m^o}\) given \(m^o\), including \(\pi_{m^o}\), we have

\begin{equation}
\mu_\pi(\theta_m^o | m^o) = \frac{\pi_{m^o}(\theta_m^o, m^o)}{\pi_{m^o}(\Theta, m^o)} = 1.
\end{equation}

Therefore, the family \(\text{supp}(\rho) = \{\pi^*, \pi_{m^o}^{**}\}_{m^o \in \mathcal{M}^o}\) of the Strong HTE \((s^*, r^*, \rho, \mu_\rho^*)\) that we constructed in the proof of Theorem 2 and that justifies the out-of-equilibrium beliefs is unique. Hence, the Strong HT refinement outcome (i.e., \(\mu_\pi^*(\theta_m^o | m^o) = 1\) for any \(m^o \in \mathcal{M}^o\)) is unique.

\textbf{Proof of Proposition 2.} Let \((s^*, r^*, \rho, \mu_\rho^*)\) be a Strong HTE. Let \(\mathcal{M}^o\) be the set of out-of-equilibrium messages, and \(\text{supp}(\rho) = \{\pi^*, \pi_{m^o}^{**}\}_{m^o \in \mathcal{M}^o}\) be the set of hypotheses of the Strong HTE; \(\pi^*\) is the initial hypothesis, and \(\pi_{m^o}^{**}\) is the new hypothesis selected for an out-of-equilibrium message \(m^o\). In Step 1, we show that there exists another strong hypothesis \(\pi_{m^o}\) that consistent with \(m^o\) and that is behaviorally consistent with respect to \(\pi^*\), In Step 2, we show that \(\pi_{m^o}\) induces the same out-of-equilibrium belief as \(\pi_{m^o}^{**}\) does. In Step 3, we construct a Behaviorally Consistent HTE.

\textbf{Step 1:} Fix an out-of-equilibrium message \(m^o \in \mathcal{M}^o\). Let \(\mathcal{B} = \{s_1, \ldots, s_N\}\) be the set of sender’s best responses. Without loss of generality, assume that \(s_1\) is the equilibrium strategy \(s^*\) (i.e., \(s^* = s_1\)). Let \(\beta\) be a receiver’s belief over \(\mathcal{B}\) that induces the strong hypothesis \(\pi_{m^o}^{**}\).
Recall $\pi_{m^\circ}^{**}$ is consistent with $m^\circ$ if $\beta(s) > 0$ for a best-response strategy $s \in B$ that generates $m^\circ$. That is,

$$\pi_{m^\circ}^{**}(\theta, m^\circ) = \sum_{s_i \in B \text{ s.t. } s_i(\theta) = m^\circ} \beta(s_i)p(\theta) > 0. \quad (41)$$

For the sake of simplicity, we set

$$\beta(s_i) = \lambda_i \in [0, 1] \text{ for each } s_i \in \{s_1, \ldots, s_N\}. \quad (42)$$

Since $\beta$ is additive on $B$, we have that $\lambda_1 + \ldots + \lambda_N = 1$.

Now, we construct another belief $\bar{\beta}$ on $B$. For a parameter $\varepsilon \in (0, 1)$, define $\bar{\beta}$ such that

$$\bar{\beta}(s_1) = 1 - \varepsilon(1 - \lambda_1) \text{ and } \bar{\beta}(s_i) = \varepsilon \lambda_i \text{ for each } s_i \in \{s_2, \ldots, s_N\}, \quad (43)$$

Notice that $\sum_{s_i \in B} \bar{\beta}(s_i) = 1 - \varepsilon + (\lambda_1 + \ldots + \lambda_N)\varepsilon = 1$. Then, for each $\varepsilon \in (0, 1)$, $\bar{\beta}$ induces another strong hypothesis, denoted by $\bar{\pi}_{m^\circ}(\varepsilon)$, that is consistent with $m^\circ$.

Since $A$ is finite and the receiver’s best-response correspondence is single-valued on the equilibrium path, there is a sufficiently small $\varepsilon^\ast \in (0, 1)$ such that $\bar{\pi}_{m^\circ}(\varepsilon^\ast)$ satisfies the following condition: For any message $m$ such that $\pi^*(\Theta, m) > 0$, we have

$$\arg \max_{a \in A} \sum_{\theta^* \in \Theta} \frac{\pi^*(\theta, m)}{\pi^*(\Theta, m)} u_R(\theta, m, a) = \{r^*(m)\} = \arg \max_{a \in A} \sum_{\theta^* \in \Theta} \bar{\pi}_{m^\circ}(\varepsilon^\ast)(\theta, m) u_R(\theta, m, a).$$

That is, $\bar{\pi}_{m^\circ}(\varepsilon^\ast)$ rationalizes the same behavior on the equilibrium path as $\pi^*$ does. This proves that $\bar{\pi}_{m^\circ}(\varepsilon^\ast)$ is behaviorally consistent with the initial hypothesis $\pi^*$.

**Step 2**: We show that $\bar{\pi}_{m^\circ}(\varepsilon^\ast)$ induces the same out-of-equilibrium belief given $m^\circ$ as $\pi_{m^\circ}^{**}$ does. By updating $\bar{\pi}_{m^\circ}(\varepsilon^\ast)$ given $m^\circ$, we get

$$\bar{\mu}_{\pi}(\theta|m^\circ) = \frac{\bar{\pi}_{m^\circ}(\varepsilon^\ast)(\theta, m^\circ)}{\bar{\pi}_{m^\circ}(\varepsilon^\ast)(\Theta, m^\circ)} = \frac{\sum_{s_i \in B \setminus \{s_1\} \text{ s.t. } s_i(\theta) = m^\circ} \varepsilon \lambda_i p(\theta)}{\sum_{\theta^* \in \Theta} \sum_{s_i \in B \setminus \{s_1\} \text{ s.t. } s_i(\theta) = m^\circ} \varepsilon \lambda_i p(\theta)} = \frac{\sum_{s_i \in B \setminus \{s_1\} \text{ s.t. } s_i(\theta) = m^\circ} \lambda_i p(\theta)}{\sum_{\theta^* \in \Theta} \sum_{s_i \in B \setminus \{s_1\} \text{ s.t. } s_i(\theta) = m^\circ} \lambda_i p(\theta)}, \quad (44)$$

38
for every $\theta \in \Theta$. By updating $\pi^{**}_{m^*}$ given $m^*$, we get

$$\mu^*(\theta|m^*) = \frac{\pi^{**}_{m^*}(\theta,m^*)}{\pi^{**}_{m^*}(\Theta,m^*)} = \frac{\sum_{s_i \in B \setminus \{s_1\}} \lambda_i p(\theta)}{\sum_{\theta \in \Theta} \sum_{s_i \in B \setminus \{s_1\}} \lambda_i p(\theta)}, $$

(45)

for every $\theta \in \Theta$. Thus, we have

$$\overline{\mu}(\cdot|m^*) = \mu^*(\cdot|m^*)$$

(46)

showing that $\pi_{m^*}(\varepsilon^*)$ justifies the out-of-equilibrium belief of the Strong HTE.

**Step 3:** Since $m^*$ was chosen arbitrary, we can construct a behaviorally consistent hypothesis $\pi_{m^*}(\varepsilon^*(m^*))$ for each out-of-equilibrium message $m^* \in \mathcal{M}^o$.\footnote{Notice that $\varepsilon^*(m^*)$ depends on $m^* \in \mathcal{M}^o$.} Therefore, by suitable choosing a second-order prior $\rho$ with $\text{supp}(\rho) = \{\pi^*, \pi^{**}_{m^*}(\varepsilon^*(m^*))\}_{m^* \in \mathcal{M}^o}$ such that

$$\{\pi^*\} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi),$$

and

$$\{\pi^{**}_{m^*}(\varepsilon^*(m^*))\} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi|m^*)$$

for each $m^* \in \mathcal{M}^o$,

yielding a Behaviorally Consistent HTE that supports the Strong HTE. \footnote{Notice that $\varepsilon^*(m^*)$ depends on $m^* \in \mathcal{M}^o$.}

**Proof of Proposition 3.** Let $(s^*, r^*, \mu^*)$ be a pooling PBE. Let $m^*$ be the pooling message and $\mathcal{M}^o$ be the set of out-of-equilibrium messages. Suppose that the PBE passes the Intuitive Criterion and that it satisfies the single-type condition (i.e., for each $m^* \in \mathcal{M}^o$, $I(m^*) = \{\theta_{m^*}\}$ for some $\theta_{m^*} \in \Theta$). By Theorem 2, we know that there exists a Strong HTE $(s^*, r^*, \rho, \mu^*)$ that explains the Intuitive Criterion outcome. That is, for each $m^* \in \mathcal{M}^o$, there exists a simple-strong hypothesis $\pi_{m^*}$ that justifies the out-of-equilibrium belief $\mu^*(\theta_{m^*}|m^*) = 1$. In other words, we have

$$\pi_{m^*} := \{\pi(\theta_{m^*}, m^*) = p(\theta_{m^*}), \pi(\theta, s^*(\theta)) = p(\theta) \text{ for any } \theta \in \Theta \setminus \{\theta_{m^*}\}\}$$

(47)

where $I(m^*) = \{\theta_{m^*}\}$. In the Strong HTE with pooling on $m^*$, $\pi_{m^*}$ takes the particular
form:

(i) \( \pi_m(\theta, m^*) = \pi^*(\theta, m^*) = p(\theta) \) for \( \theta \in \Theta \setminus \{\theta_m^\circ\} \),
(ii) \( \pi_m(\theta_m^\circ, m^*) = p(\theta_m^\circ) \),
(iii) \( \pi_m(\theta_m^\circ, m^*) = 0 \),

where \( \pi^* \) is the initial hypothesis of the Strong HTE supporting the pooling PBE.

Now, suppose that \( \theta_m^\circ \) is a dummy for the equilibrium message \( m^* \). Therefore, for any \( \bar{a} \in \mathcal{A} \) such that

\[
\bar{a} \in \arg \max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \pi^*(\theta, m^*) u_R(\theta, m^*, a),
\]

it is true that

\[
\bar{a} \in \arg \max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \bar{\pi}(\theta, m^*) u_R(\theta, m^*, a),
\]

(49)

where \( \bar{\pi} \) is a probability distribution on \( \Theta \times \mathcal{M} \) such that

\[
\bar{\pi}(\theta, m^*) = \begin{cases} 
\pi^*(\theta, m^*), & \text{if } \theta \in \Theta \setminus \{\theta_m^\circ\} , \\
0, & \text{if } \theta = \theta_m^\circ .
\end{cases}
\]

(50)

Notice that the simple-strong hypothesis \( \pi_m^\circ \) satisfies Condition (49) and (50). By Condition (49), \( \pi_m^\circ \) is behaviorally consistent with respect to \( \pi^* \). Notice that \( m^\circ \) was arbitrarily chosen. Therefore, for each out-of-equilibrium message \( m^\circ \), we can find a simple-strong hypothesis \( \pi_m^\circ \) that rationalizes the receiver’s behavior on the equilibrium path as well as her behavior off the equilibrium path. Let \( \{\pi_m^\circ\}_{m^\circ \in \mathcal{M}^\circ} \) be the family of such behaviorally consistent hypotheses.

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

\[
\{\pi^*\} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi),
\]

(51)

(i.e., \( \pi^* \) is the initial hypothesis) and

\[
\{\pi_m^{**}\} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi|m^\circ) \text{ for each } m^\circ \in \mathcal{M}^\circ
\]

(52)

(i.e., \( \pi_m^{**} \) is the most likely hypothesis after updating \( \rho \) given \( m^\circ \in \mathcal{M}^\circ \)). This shows that there always exists a Behaviorally Consistent HTE \( (s^*, r^*, \rho, \mu^*_s) \) that supports the pooling PBE.
B PBE and HTE in Behavioral Strategies

In this Appendix, we show that Theorem 1 can be extended to setups with behavioral strategies. That is, if we allow for behavioral strategies, then each PBE can be again supported by an HTE.

A sender’s behavioral strategy is a mapping \( b^*_S : \Theta \to \Delta(\mathcal{M}) \). We denote by \( b^*_S(m|\theta) \) the probability that the sender’s type \( \theta \) sends message \( m \). A receiver’s behavioral strategy is a mapping \( b^*_R : \mathcal{M} \to \Delta(\mathcal{A}) \). We denote \( b^*_R(a|m) \) the probability that the receiver replies \( a \) after \( m \) is observed.

A PBE in behavioral strategies is defined as follows.

**Definition 11** \((b^*_S, b^*_R, \mu^*)\) is a PBE (in behavioral strategies) for a signaling game in \( \mathcal{G} \) if:

(i) \( b^*_S(\cdot|\theta) \in \arg \max_{b^*_S(\cdot;\theta) \in \Delta(\mathcal{M})} u_S(\theta, b^*_S(\cdot;\theta), b^*_R) \) for each \( \theta \in \Theta \),

(ii) \( b^*_R(\cdot;m) \in \arg \max_{b^*_R(\cdot;m) \in \Delta(\mathcal{A})} \sum_{\theta \in \Theta} \mu^*(\theta|m)u_R(\theta, m, b^*_R(\cdot;m)) \) for each \( m \in \mathcal{M} \),

(iii) \( \mu^*(\theta|m) = \frac{\pi(\theta,m)}{\pi(\Theta,m)} \) for each \( \theta \in \Theta \) if \( \pi(\Theta,m) > 0 \), and \( \mu^*(\cdot|m) \) is an arbitrary probability distribution over \( \Theta \) if \( \pi(\Theta,m) = 0 \), where

\[
\pi(\theta,m) = \begin{cases} 
  b^*_S(m;\theta)p(\theta), & \text{if } b^*_S(m;\theta) > 0, \\
  0, & \text{otherwise}.
\end{cases}
\]

Notice that a PBE is a pure PBE if \( b^*_S \) and \( b^*_R \) are degenerate behavioral strategies. That is, for each \( \theta \in \Theta \), \( b^*_S(m|\theta) = 1 \) for some \( m \in \mathcal{M} \); and for each \( m \in \mathcal{M} \), \( b^*_R(a|m) = 1 \) for some \( a \in \mathcal{A} \).

The next result shows that each PBE, \((b^*_S, b^*_R, \mu^*)\), can be supported by an HTE. That is, we can always find a set of hypotheses that induces the family of posterior beliefs of the PBE.

**Theorem 3** Let \((b^*_S, b^*_R, \mu^*)\) be a PBE for a signaling game. Then, there exists an HTE, \((b^*_S, b^*_R, \rho, \mu^*_\rho)\), that supports the PBE. That is, there are \( \rho \), a set of hypotheses \( \text{supp}(\rho) \) and a family of posterior beliefs \( \mu^*_\pi := \{\mu^*_\pi(\cdot|m)\}_{m \in \mathcal{M}} \) derived via the Hypothesis Testing updating
rule such that

\[(i) \quad \mu^*(\cdot|m) = \frac{\pi^*(\cdot,m)}{\pi^*(\Theta,m)} = \mu^*_\pi(\cdot|m) \quad \text{for each equilibrium message } m \text{ where} \]

\[\{\pi^*\} := \arg \max_{\pi \in \text{supp}(\rho)} \pi(\cdot)m \quad \text{and} \quad \frac{\pi^*(\theta,m)}{\pi^*(\Theta,m)} = \frac{b^*_S(m;\theta)}{\sum_{\theta' \in \Theta} b^*_S(m;\theta')p(\theta')},\]

\[(ii) \quad \mu^*_\pi(\cdot|m) = \frac{\pi^*_m(\cdot,m)}{\pi^*_m(\Theta,m)} = \mu^*_\pi(\cdot|m^o) \quad \text{for each out-of-equilibrium message } m^o \in \mathcal{M}^o \text{ where} \]

\[\{\pi^*_m\} := \arg \max_{\pi \in \text{supp}(\rho)} \pi(m|m^o).\]

**Proof:** Consider a PBE, \((b^*_S, b^*_R, \mu^*)\), in behavioral strategies. Let \(\mathcal{M}^o\) be the set of out-of-equilibrium messages. We can apply the argument presented in Step 2 in the proof of Theorem 1 to show that for each \(m^o \in \mathcal{M}^o\), we can contract a hypothesis \(\pi^*_m\) that induces the out-of-equilibrium belief \(\mu^*(\cdot|m^o)\) of the PBE. Let \(\{\pi^*_m\}_{m^o \in \mathcal{M}^o}\) the set of such hypothesis.

It remains to construct a hypothesis \(\pi^*\) that induces the PBE belief on the equilibrium path. That is, we need to find a belief \(\beta\) on \(\mathcal{S}\) such that for each \(\theta \in \Theta\),

\[\mu^*(\theta|m) = \frac{b^*_S(\theta|m)p(\theta)}{\sum_{\theta' \in \Theta} b^*_S(\theta'|m)p(\theta')} = \frac{\sum_{s \in \mathcal{S}} \beta(s)p(\theta)}{\sum_{\theta' \in \Theta \in \mathcal{S}} \beta(s)p(\theta')} = \frac{\pi^*(\theta,m)}{\pi^*(\Theta,m)}, \tag{53}\]

Notice \(\beta\) can be seen as a mixed strategy. Since signaling games satisfy perfect recall, by the Kuhn Theorem, there exists \(\beta^*\) that is payoff-equivalent with the equilibrium strategy \(b^*\) (Kuhn, 1953). behavioral strategy in games with perfect recall (Kuhn, 1953). Thus, \(\beta^*\) together with the prior information \(p\) induces the hypothesis \(\pi^*\) defined in Equation (53).

Thus, we can suitably choose a second-order prior \(\rho\) with \(\text{supp}(\rho) = \{\pi^*, \pi^*_m\}_{m^o \in \mathcal{M}^o}\) such that

\[\{\pi^*\} := \arg \max_{\pi \in \text{supp}(\pi)} \rho(\pi), \quad \text{and} \quad \{\pi^*_m\} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(m|m^o) \quad \text{for each } m^o \in \mathcal{M}^o \tag{54}\]

(i.e., \(\pi^*\) is the initial hypothesis and \(\pi^*_m\) is the most likely hypothesis after updating \(\rho\) given \(m^o \in \mathcal{M}^o\)). Therefore, there exists an HTE \((b^*_S, b^*_R, \rho, \mu^*_\pi)\) supporting the PBE \((b^*_S, b^*_R, \mu^*)\) in behavioral strategies. \(\blacksquare\)
C Intuitive Criterion and Arbitrary Beliefs.

In this Appendix, we present a PBE with $I(m^\circ) = \Theta$ that passes the Strong HT refinement and the Intuitive Criterion. However, the Intuitive Criterion does not reduce the number of PBE beliefs while our refinement does.

Consider the signaling game in Figure 3 with $x = 2$. In this game, the set of types is $\Theta = \{\theta_L, \theta_H\}$. Consider the PBE with pooling on $N$, i.e., the strategy profile

$$s^*(\theta_L) = s^*(\theta_H) = N, \ r^*(E) = m, \ r^*(N) = e,$$

and posterior beliefs $\mu^*(\theta_L|N) = 2/5$ and $\mu^*(\theta_L|L) \leq 1/2$.

This PBE passes the Intuitive Criterion. The Intuitive Criterion asserts that $I(E) = \{\theta_L, \theta_H\}$ and $T(E) = \emptyset$, and therefore any probability distribution over $\Theta = \{\theta_L, \theta_H\}$ is admitted. Hence, all the out-of-equilibrium beliefs are consistent with the Intuitive Criterion.

Notice that this game has the same best responses and strong hypotheses as defined in Step 2 of the proof of Proposition 1. For each $\lambda \in (0,1]$, we consider the strong hypothesis $\pi(\lambda)$:

$$\pi(\lambda) := \{\pi_\lambda(\theta_L, E) = \lambda 2/5, \ \pi_\lambda(\theta_H, E) = \lambda 3/5, \ \pi_\lambda(\theta_L, N) = (1-\lambda)2/5, \ \pi_\lambda(\theta_H, N) = (1-\lambda)3/5\}.$$  

Updating $\pi(\lambda)$ given $E$ yields the out-of-equilibrium belief $\mu_\pi(\theta_L|E) = 2/5$ for each $\lambda \in (0,1]$. Hence, the PBE can be supported by a Strong HTE with $\rho$ such that $supp(\rho) = \{\pi(\lambda), \pi_2\}$ and $\rho(\pi(\lambda)) < \rho(\pi_2)$. Moreover, the Strong HT refinement selects the unique belief $\mu_\pi(\theta_L|E) = 2/5$.

D Non-Existence of HTE under Simple Hypotheses

In this Appendix, we constrain HTE to simple hypotheses and show that such HTE may not exist. In particular, there is a signaling game for which a (pure) PBE exists and yet it cannot be supported by an HTE constrained to simple hypotheses about a sender’s behavior.

Consider the signaling game depicted in Figure 4. There exists a (unique) pooling PBE in which both worker types acquire education. That is,

$$s^*(\theta_L) = s^*(\theta_H) = E, \ r^*(E) = e, \ r^*(N) = m,$$

and posterior belief $\mu^*(\theta_L|E) = 0.9$ and $\mu^*(\theta_L) \in [0.475, 0.525]$ form the pooling PBE.

Recall that a simple hypothesis is induced by a degenerate belief on the set of sender’s
pure strategies. Therefore, there are four simple hypotheses in this game:

1) \( \pi_1 := \{ \pi_1(\theta_L, N) = 0.9, \pi_1(\theta_H, E) = 0.1 \} \) if \( \beta(s_1) = 1 \) and \( s_1 := \{ s(\theta_L) = N, s(\theta_H) = E \} \),

2) \( \pi_2 := \{ \pi_2(\theta_L, E) = 0.9, \pi_2(\theta_H, N) = 0.1 \} \) if \( \beta(s_2) = 1 \) and \( s_2 := \{ s(\theta_L) = E, s(\theta_H) = N \} \),

3) \( \pi_3 := \{ \pi_3(\theta_L, E) = 0.9, \pi_3(\theta_H, E) = 0.1 \} \) if \( \beta(s_3) = 1 \) and \( s_3 := \{ s(\theta_L) = E, s(\theta_H) = E \} \),

4) \( \pi_4 := \{ \pi_4(\theta_L, N) = 0.9, \pi_4(\theta_H, N) = 0.1 \} \) if \( \beta(s_4) = 1 \) and \( s_4 := \{ s(\theta_L) = N, s(\theta_H) = N \} \).

The simple hypothesis \( \pi_3 \) supports the receiver’s best response on the equilibrium path. However, none of the simple hypotheses in \( \{ \pi_1, \pi_2, \pi_4 \} \) can rationalize – after updating it – the receiver’s response \( m \) to the out-of-equilibrium message \( N \). Therefore, there does not exist an HTE with simple hypotheses that can explain the pooling PBE.

### E Online Appendix

In this online appendix, we compare the Strong HT refinement with other refinement criteria.

For an intuitive PBE, let \( \mathcal{M}^o \) be the set of out-of-equilibrium messages and \( I(m^o) \) be the set of types who could benefit from sending \( m^o \in \mathcal{M}^o \). The Intuitive Criterion has a limitation (similar to PBE) if either \( |I(m^o)| = 0 \) or \( |I(m^o)| > 1 \). In the former case, the Intuitive Criterion admits any probability distribution on \( \Theta \). In the latter case, it admits any probability distribution on \( I(m^o) \). Because of this limitation, a variety of stronger refinement concepts have been suggested in the economics literature including the Universally

Below, we establish a relationship between Strong HTE and the uRCE of Fudenberg and He (2017, 2018). The uRCE is the strongest solution concept among the ones mentioned above. Fudenberg and He (2017) showed that the above equilibria are nested under the notion of path-equivalence. Two (equilibrium) strategy profiles are said to be path-equivalent if the sender’s behavior and the receiver’s on-the-equilibrium-path behavior, respectively, is the same under each strategy profile. This means that path-equivalent strategy profiles may only differ in receiver’s behaviors off the equilibrium path. Fudenberg and He (2017, p.12) showed that

\[ \text{uRCE} \subseteq \text{UDE} \subseteq \text{RCE} \subset \text{Intuitive PBE}, \]

where \( \subseteq \) indicates a (proper) set inclusion of path-equivalent (equilibrium) strategy profiles. That is, for each uRCE there exists a path-equivalent UDE but not vice versa, and so on.

By Theorem 2, we know that an intuitive PBE that satisfies the single-type condition can be supported by a Strong HTE. However, beyond the single-type condition, it is not clear how the Strong HT refinement and the Intuitive Criterion are related. Below, we establish a relationship between Strong HTE and intuitive PBEs with \( |I(m^o)| \geq 1 \) for each \( m^o \in \mathcal{M}^o \).

Recall that a PBE may fail the Strong HT refinement for two reasons: (1) for some \( m^o \in \mathcal{M}^o \), there does not exist a strong hypothesis that is consistent with \( m^o \), and (2) none of the strong hypotheses justifies an out-of-equilibrium belief of the PBE. The next lemma shows that an intuitive PBE might fail the Strong HT refinement only for the second reason.

**Lemma A1** Let \( (s^*, r^*, \mu^*) \) be a PBE with \( |I(m^o)| \geq 1 \) for each out-of-equilibrium message \( m^o \in \mathcal{M}^o \). For each \( m^o \in \mathcal{M}^o \), there exists a strong hypothesis \( \pi \) that is consistent with \( m^o \).

**Proof.** Fix an out-of-equilibrium message \( m^o \in \mathcal{M}^o \). Since it is assumed that \( |I(m^o)| \geq 1 \), there is \( a^o \in BR(\Theta, m^o) \) such that

\[ u_S^o(\theta) \leq u_S(\theta, m^o, a^o) \text{ for some } \theta \in I(m^o). \]  \( (55) \)

Denote by \( I(m^o; a^o) \) the set of sender’s types satisfying Condition \( (55) \). We construct a receiver’s strategy \( r^o \) such that

\[ r^o(m) = \begin{cases} r^*(m), & \text{if } m \neq m^o, \\ a^o, & \text{if } m = m^o. \end{cases} \]  \( (56) \)
That is, \( r^o \) is identical to the equilibrium strategy \( r^* \) except for a receiver’s response to the out-the-equilibrium message \( m^o \).

Now, denote by \( s^o \) a sender’s best responding strategy to \( r^o \) that generates \( m^o \). That is,

\[
s^o(\theta) = s^*(\theta) \quad \text{for each } \theta \in \Theta \setminus I(m^o; a^o) \quad \text{and} \quad s^o(\theta) = m^o \quad \text{for each } \theta \in I(m^o; a^o).
\]

(57)

Consider a degenerate belief \( \beta \) over \( \mathcal{B} \) such that \( \beta(s^o) = 1 \) and \( \beta(s) = 0 \) for any \( s \in \mathcal{B}\setminus\{s^o\} \). This belief together with the prior information \( p \) induces a simple-strong hypothesis \( \pi_{m^o} \) such that

(i) \( \pi_{m^o}(\theta, m^o) = p(\theta) \) for each \( \theta \in I(m^o; a^o) \),

(ii) \( \pi_{m^o}(\theta, s^*(\theta)) = p(\theta) \) for each \( \theta \in \Theta \setminus I(m^o; a^o) \).

(58)

By updating \( \pi_{m^o} \) given \( m^o \), we get a posterior belief \( \mu_{\pi}(\cdot|m^o) \) that satisfies \( \mu_{\pi}(\theta|m^o) > 0 \) for each \( \theta \in I(m^o; a^o) \). In this way, for each out-of-equilibrium message \( m^o \in \mathcal{M}^o \), we can construct a strong hypothesis \( \pi_{m^o} \) that is consistent with \( m^o \), completing the proof.

By Lemma A1, an intuitive PBE might fail to pass the Strong HT refinement because some of its out-of-equilibrium beliefs cannot be justified by a strong hypothesis. Thus, the belief is inconsistent with the prior information about types under a sender’s rational behavior.

However, there is a class of intuitive PBEs that can be supported by a Strong HTE. In particular, we show that if an intuitive PBE is uRCE, it passes the Strong HT refinement.

We briefly recapitulate the uRCE notion of Fudenberg and He (2017, 2018). The equilibrium builds on the rationally-compatible order \( \succ_{m^o} \). For each out-of-equilibrium message \( m^o \in \mathcal{M}^o \), \( \succ_{m^o} \) specifies which type is more likely to send \( m^o \) against receiver’s rational strategies. The set of receiver’s rational (behavioral) strategies, \( \mathcal{B}^*_R \), is given by

\[
\mathcal{B}^*_R := \Delta(BR(\Theta, m)),
\]

where \( BR(\Theta, m) \) is

\[
BR(\Theta, m) := \bigcup_{\{\mu: \mu(\Theta|m) = 1\}} BR(\mu, m). \quad \text{(31)}
\]

That is, \( \mathcal{B}^*_R \) is the set of receiver’s strategies without conditionally dominated actions.\(^{32}\)

An out-of-equilibrium message \( m^o \) is said to be more rationally-compatible with \( \theta \) than \( \beta(s^o) = 1 \) and \( \beta(s) = 0 \) for any \( s \in \mathcal{B}\setminus\{s^o\} \).

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

\[\text{A receiver’s action } a \text{ is conditionally dominated for } m \text{ if } a \notin BR(\Theta, m). \]

\(^{31}\)
\( \theta' \), written as \( \theta \succ_{m^o} \theta' \), if, for any \( b_R \in B_R^* \),

\[
u_S(\theta', m^o, b_R(\cdot | m^o)) \geq \max_{m \neq m^o} \nu_S(\theta', m, b_R(\cdot | m)),
\]

implies that

\[
u_S(\theta, m^o, b_R(\cdot | m^o)) > \max_{m \neq m^o} \nu_S(\theta, m, b_R(\cdot | m)).
\]

That is, whenever type \( \theta' \) sends \( m^o \) as a (weak) best response against any \( b_R \in B_R^* \), type \( \theta \) sends \( m^o \) as a (strict) best response against \( b_R \).

Based on the rationally-compatible order, we define the set of uniformly rationality-compatible beliefs, \( \{ \hat{P}(m^o) \}_{m^o \in M^o} \). For each out-of-equilibrium message \( m^o \), \( \hat{P}(m^o) \) is composed of two parts: \( \Delta(\Theta_{m^o}) \) and \( P_{\theta \succ m^o} \). The former set is the set of receiver’s beliefs over \( \Theta_{m^o} \), where \( \Theta_{m^o} \) is the set of types for which \( m^o \) is not conditionally dominated. The latter set contains beliefs for which the relative ratio of \( \theta \) to \( \theta' \) exceeds the prior information \( p \). That is,

\[
P_{\theta \succ \theta'} := \left\{ \mu \in \Delta(\Theta) : \frac{\mu(\theta')}{\mu(\theta)} \leq \frac{p(\theta')}{p(\theta)} \right\}.
\]

The set of uniformly rationality-compatible beliefs \( \hat{P}(m^o) \) with respect to \( m^o \) is defined as

\[
\hat{P}(m^o) := \Delta(\Theta_{m^o}) \cap \{ P_{\theta \succ \theta'} : \theta \succ_{m^o} \theta' \}.
\]

Notice that we have \( \hat{P}(m^o) = \Delta(\Theta_{m^o}) \) whenever \( \succ_{m^o} \) is “empty” (i.e., there are no types for which the rationally-compatible relation holds true).\(^{33}\)

The uRCE introduced by Fudenberg and He (2017, 2018) is defined as follows.

**Definition 12 (uRCE)** Let \((s^*, r^*)\) be a PBE strategy profile. Then, \((s^*, r^*)\) is called a uniformly Rationality-Compatible Equilibrium if for all \( \theta \in \Theta \), it is true that

\[
u_S^*(\theta) = \nu_S(\theta, m^o, a)
\]

for each \( m^o \in M^o \) and each \( a \in BR(\hat{P}(m^o), m^o) \),

where \( \nu_S^*(\theta) \) is the equilibrium payoff for the sender’s type \( \theta \) in the PBE.

In uRCE, for each out-of-equilibrium message \( m^o \), the sender cannot be better off than her equilibrium payoff by playing \( m^o \) against any receiver’s action in \( BR(\hat{P}(m^o), m^o) \).

In the next proposition, we show that for each uRCE there exists a path-equivalent Strong HTE that passes the Intuitive Criterion.

\(^{33}\)We are grateful to Kevin He for clarifying this case.
**Proposition A1** Consider a uRCE with \(|I(m^\circ)| \geq 1\) for each \(m^\circ \in \mathcal{M}^\circ\). Then, there exists a path-equivalent Strong HTE that passes the Intuitive Criterion.

**Proof.** Let \((s^*, r^*)\) be a uRCE with \(|I(m^\circ)| \geq 1\) for each \(m^\circ \in \mathcal{M}^\circ\). Since \(|I(m^\circ)| \geq 1\), by Lemma A1, for each out-of-equilibrium message \(m^\circ \in \mathcal{M}^\circ\), there exists a simple-strong hypothesis \(\pi_{m^\circ}\) consistent with \(m^\circ\) that satisfies Equation (58). Notice that \(\pi_{m^\circ}\) is induced by a degenerate belief such that \(\beta(s^\circ) = 1\) and \(\beta(s) = 0\) for \(s \not= s^\circ\), where \(s^\circ\) is specified in Equation (57). Denote by \(\{\pi_{m^\circ}\}_{m^\circ \in \mathcal{M}^\circ}\) the family of such simple-strong hypotheses.

First, we construct a receiver’s strategy \(r^\circ\) that coincides with \(r^*\) on the equilibrium path. However, for each \(m^\circ \in \mathcal{M}^\circ\), let \(a_{m^\circ}\) be a receiver’s best response against \(m^\circ\) with respect to the posterior belief \(\mu(\cdot|m^\circ)\) derived from \(\pi_{m^\circ}\). That is, \(r^\circ\) is defined as follows:

\[
r^\circ(m) = \begin{cases} 
  r^*(m), & \text{if } m \in \{s^\circ(\theta)\}_{\theta \in \Theta} = \mathcal{M} \setminus \mathcal{M}^\circ, \\
  a_{m^\circ}, & \text{if } m^\circ \in \mathcal{M}^\circ,
\end{cases}
\]

(59)

where \(s^\circ(\theta)\) is the sender’s equilibrium message for type \(\theta\).

Next, we show that the sender’s (equilibrium) strategy \(s^*\) best responds to \(r^\circ\). To do that, we need to show that \(\mu_{\pi}(\cdot|m^\circ)\) belongs to the set of uniformly rationality-compatible beliefs, i.e.,

\[
\mu_{\pi}(\cdot|m^\circ) \in \hat{P}(m^\circ) = \Delta(\Theta_{m^\circ}) \cap \{P_{\theta \theta'} : \theta \succ_{m^\circ} \theta'\} \text{ for each } m^\circ \in \mathcal{M}^\circ.
\]

If \(m^\circ\) is a dominated action for some type \(\theta'\), then \(\theta' \not\in I(m^\circ)\) and thus we have \(\mu_{\pi}(\theta'|m^\circ) = 0\). Hence, it must be the case that

\[
\mu_{\pi}(\cdot|m^\circ) \in \Delta(\Theta_{m^\circ}) \text{ for each } m^\circ \in \mathcal{M}^\circ.
\]

Moreover, suppose that there are types \(\theta_{m^\circ}\) and \(\theta'_{m^\circ}\) such that \(\theta_{m^\circ} \succ_{m^\circ} \theta'_{m^\circ}\) for some \(m^\circ\). Since \(m^\circ\) is a best-responding message for \(\theta_{m^\circ}\) whenever \(m^\circ\) is a best response for type \(\theta'_{m^\circ}\), by Equation (58) and the construction of \(\pi_{m^\circ}\), we cannot have that

\[
\frac{\pi_{m^\circ}(\theta'_{m^\circ}, m^\circ)}{\pi_{m^\circ}(\theta_{m^\circ}, m^\circ)} > \frac{p(\theta'_{m^\circ})}{p(\theta_{m^\circ})},
\]

(60)

where \(p\) is the prior information about sender’s types. Hence, it must be the case that

\[
\mu_{\pi}(\cdot|m^\circ) \in \{P_{\theta \theta'} : \theta \succ_{m^\circ} \theta'\} \text{ for each } m^\circ \in \mathcal{M}^\circ.
\]
Therefore, we have
\[ \mu_\pi(\cdot|m^o) \in \hat{P}(m^o) \] for each \( m^o \in \mathcal{M}^o \).

By definition of uRCE, we know that \( s^*(\theta) \) best responds against any \( a \in \text{BR}(\hat{P}(m^o), m^o) \)
for each \( m^o \in \mathcal{M}^o \). Since \( \mu_\pi(\cdot|m^o) \in \hat{P}(m^o) \), it is true that \( r^o(m^o) \in \text{BR}(\hat{P}(m^o), m^o) \)
for each \( m^o \in \mathcal{M}^o \). This shows that \( s^* \) best responds to \( r^o \).

Finally, we can construct a Strong HTE with the strategy profile \((s^*, r^o)\). A degenerate belief \( \beta \) such that \( \beta(s^*) = 1 \) together with the prior information \( \rho \) induces a strong hypothesis \( \pi^* \) that rationalizes the receiver's on-the-equilibrium-path behavior. By construction of \( r^o \), for each \( m^o \in \mathcal{M}^o \), \( r^o(m^o) \) best responds to \( m^o \) with respect to \( \mu_\pi(\cdot|m^o) \) derived from \( \pi^o \). Thus, we can choose a second-order prior \( \rho \) with \( \text{supp}(\rho) = \{\pi^*, \pi^o\}_m \in \mathcal{M}^o \) such that
\[ \{\pi^o\} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi) \] and \[ \{\pi^o\} := \arg \max_{\pi \in \text{supp}(\rho)} \rho(\pi|m^o) \] for each \( m^o \in \mathcal{M}^o \).

Notice that \((s^*, r^o)\) of the Strong HTE is path-equivalent to \((s^*, r^*)\) of the uRCE.

It remains to be shown that the Strong HTE passes the Intuitive Criterion. Since each hypotheses in \( \{\pi^o\}_m \in \mathcal{M}^o \) satisfies Equation (58), it must be true that \( \mu_\pi(\cdot|m^o) \in \Delta(I(m^o)) \) for each \( m^o \in \mathcal{M}^o \). Hence, the Strong HTE passes the Intuitive Criterion. \( \blacksquare \)

We end this section by providing a condition under which each uRCE is Strong HTE.

**Corollary A1** Let \((s^*, r^*)\) be a uRCE with \( |I(m^o)| \geq 1 \) for each \( m^o \in \mathcal{M}^o \). Let \( \hat{P}(m^o) \) be the set of uniformly rationality-compatible beliefs of the uRCE and \( \mathcal{O}(m^o) \) be the set of beliefs supporting the receiver’s best response \( r^*(m^o) \). If \( \hat{P}(m^o) \subseteq \mathcal{O}(m^o) \) for each \( m^o \in \mathcal{M}^o \), then there exists a Strong HTE that supports the uRCE. Moreover, it passes the Intuitive Criterion.

**Proof.** Let \((s^*, r^*)\) be a uRCE with \( |I(m^o)| \geq 1 \) for each \( m^o \in \mathcal{M}^o \). Consider an out-of-equilibrium message \( m^o \). By Lemma A1, we know that there is a simple-strong hypothesis \( \pi^o \) that is consistent with \( m^o \) and satisfies Equation (58). In the proof of Proposition A1, we have shown that \( \mu_\pi(\cdot|m^o) \) derived from \( \pi^o \) is an element of \( \hat{P}(m^o) \).

If \( \hat{P}(m^o) \subseteq \mathcal{O}(m^o) \), then each \( \mu_\pi(\cdot|m^o) \in \hat{P}(m^o) \) rationalizes the receiver’s (equilibrium) best response \( r^*(m^o) \) to \( m^o \). Since this is true for each \( m^o \in \mathcal{M}^o \), the uRCE is supported by a Strong HTE constructed in the proof of Proposition A1. Moreover, since each \( \pi^o \) satisfies Equation (58), it is true that \( \mu_\pi(\cdot|m^o) \in \Delta(I(m^o)) \) showing that the Strong HTE passes the Intuitive Criterion. \( \blacksquare \)

Notice that uRCE under the condition that \( \hat{P}(m^o) \subseteq \mathcal{O}(m^o) \) is an intuitive PBE. Therefore, we have shown that if an intuitive PBE with \( |I(m^o)| \geq 1 \) for each \( m^o \in \mathcal{M}^o \) is uRCE
with \( \hat{P}(m^*) \subseteq O(m^*) \), then there exists a Strong HTE supporting the intuitive PBE.

References


