Testing Behavioral Hypotheses in Signaling Games*

Adam Dominiak†1 and Dongwoo Lee‡2

1,2 Department of Economics, Virginia Tech

April 11, 2017

Abstract
In this paper, we apply Ortoleva’s (2012) idea of Hypothesis Testing Equilibrium (HTE) as a refinement tool for Perfect Bayesian Equilibrium (PBE) in signaling games. HTE is a solution concept that admits updating of beliefs on out-of-equilibrium paths by selecting most likely hypotheses (i.e., receiver’s beliefs about sender’s strategic behavior). When hypotheses are about mixed strategies, it shows that each PBE can be supported by an HTE, yet without refining the out-of-equilibrium beliefs. However, if hypotheses are about pure strategies, HTE might not exist; but if it does, it refines the posterior beliefs of a given PBE. It is demonstrated that the Hypothesis Testing refinement is unrelated to the well-known Intuitive Criterion, unless the set of types who could benefit from sending an out-of-equilibrium message is a singleton. Moreover, we suggest a strengthening of the HTE notion under which the posterior beliefs are immune against Mailath’s (1988) critic and show that HTE can much better explain the experimental findings of Brandts and Holt (1992) than the Intuitive Criterion.

Keywords: Signaling games, Perfect Bayesian Equilibrium, Hypothesis Testing, updating rule, out-of-equilibrium beliefs, refinements.

JEL Classification: C72, D81, D83

---

*The authors are deeply indebted to Hans Haller and Gerelt Tserenjimid for their valuable comments and fruitful discussions.
†Email: dominiak@vt.edu
‡Email: dwlee05@vt.edu
1 Introduction

Perfect Bayesian Equilibrium (PBE) is a standard solution concept for dynamic games with incomplete information such as signaling games, cheap talk games, etc. However, the equilibrium notion has a limitation. It allows for arbitrary out-of-equilibrium beliefs since Bayes’ rule does not specify how beliefs are derived on events with a zero probability. To narrow down the plethora of equilibrium beliefs, many refinements have been suggested, including the predominant Intuitive Criterion of Cho and Kreps (1987). Yet, the Intuitive Criterion has been criticized. For instance, Mailath (1988) pointed out that the Intuitive Criterion might lead to “implausible” beliefs due to inconsistencies between on-equilibrium and out-of-equilibrium beliefs. In an experimental study on the Intuitive Criterion, Brandts and Holt (1992) found that individuals behave in line with an “unituitive” equilibrium.

In this paper, we elaborate a refinement tool which eliminates Mailath’s type of inconsistencies and which can perfectly accommodate the experimental findings of Brandts and Holt (1992).

Recently, Ortoleva (2012) introduced a novel theory of updating beliefs on zero-measure events, the so-called Hypothesis Testing Model. In the context of signaling games, the receiver updates her beliefs by testing hypotheses (i.e., first-order priors) about sender’s strategic behavior. Before any information is revealed, the receiver chooses the most likely “hypothesis” according to her prior over hypotheses (i.e., second-order prior) and updates it in the Bayesian manner whenever possible. If, however, an unexpected message is observed, the initial hypothesis is rejected. The receiver updates her second-order prior by Bayes’ rule given the out-of-equilibrium message and then she chooses an alternative hypothesis in the maximum likelihood fashion. The chosen hypothesis is updated by Bayes’ rule again. This updating procedure provides posterior beliefs which are well-defined on out-of-equilibrium paths. For this reason, the “hypotheses passing the test” can serve as a refinement criterion for the posterior beliefs of a given PBE.

This paper applies the Hypothesis Testing (HT) updating rule as a formal language to reason about out-of-equilibrium beliefs. If an unexpected message is observed, the receiver discards the most likely hypothesis that describes the sender’s (pure) equilibrium behavior (i.e., the sender’s best response against the receiver’s equilibrium strategy). While searching for an alternative hypothesis, there are a few conceivable ways the receiver might think about sender’s behavior on out-of-equilibrium paths. The unexpected message might be regarded as realization of sender’s (non-equilibrium) pure or mixed strategy which in turn might be (or not) a best response against some of the receiver’s strategy. If, for every out-of-equilibrium

\[\text{That is, an equilibrium that fails the Intuitive Criterion.}\]
message, there exists a hypothesis about the sender’s strategic behavior inducing a posterior belief of a given PBE, then the equilibrium is said to pass the HT refinement.

Our analysis unfolds in steps: First, we will assume that hypotheses are about sender’s pure strategies that best respond to some strategy of the receiver and compare the HT refinement concept with the Intuitive Criterion. Second, we will introduce the notion of \textit{behaviorally consistent} hypothesis motivated by Mailath’s (1988) criticism of the Intuitive Criterion and explain the experimental results of Brandts and Holt (1992). Finally, we will show that depending on whether hypotheses are about pure or mixed strategies of the sender, it substantially affects the refinement outcome.

The first step begins with illustrating how the HT refinement concept works. Consider a pooling PBE in a labor-market game à la Spence (1973) in which a job applicant with private information about his skills - being either high or low - decides to obtain education independent of his ability level. An employer, knowing (ex-ante) that low-ability workers are less likely than high-ability workers and that education is more costly for low-ability workers, assigns the job applicant with education to an executive job. An applicant signaling no education receives a manual task as long as the employer’s posterior belief stipulates that no education is more likely to be a signal of low ability. If one assume that hypotheses (i.e. employer’s first-order beliefs) are about pure strategies, there exists only one hypothesis that passes the test. The hypothesis conjectures that workers “separate”, i.e., the high-ability workers obtain education whereas the low-ability workers do not. After updating it, the employer believes that no education is solely chosen by low-ability workers, thus justifying the unique posterior belief of the pooling PBE.

Suppose now that low-ability workers could benefit from sending the out-of-equilibrium message (no education) by assuming that an executive job is always assigned. Then, the before mentioned hypothesis is even “stronger” in the sense that it describes the worker’s best response against the employer’s strategy that matches all workers with an executive task. Moreover, in this case, the pooling equilibrium behavior survives the Intuitive Criterion, yielding the same posterior belief as the HT refinement. Then, the former criterion asserts that no education shall be interpreted as a signal of low ability, since only the low-ability workers.

\footnote{This game is depicted in Figure 1.}

\footnote{There are two other hypotheses describing worker’s pure strategy behavior. This first one asserts that all worker’s types do not obtain education. The second one asserts a “reverse” type-dependence (i.e., the low-ability workers obtain education whereas the high-ability workers do not). Yet, both hypotheses induce, after updating, that no education is more likely to be sent by high-ability workers, conflicting with the posteriors of the pooling PBE.}

\footnote{A similar observation was made by Ortoleva (2012) in an example of the Beer and Quiche game.}
workers could benefit from sending the signal.

Given this observation, it is naturally to ask whether the outcomes of the HT refinement and the Intuitive Criterion are somehow related? In general, it is not the case. It is demonstrated that the two refinement concepts are not “nested” in any sense. However, there are signaling games in which the Intuitive Criterion outcome can always be justified by an alternative hypothesis. If a given PBE passes the Intuitive Criterion and there is a single type who could benefit from sending an out-of-equilibrium message, then the PBE also passes the HT refinement. Yet, while the refinement outcome of the Intuitive Criterion is unique in this case, there might be more than one alternative hypothesis justifying a finite number of posterior beliefs of the “intuitive” equilibrium. If, in addition, an out-of-equilibrium message is a never best response for all types who can never benefit from sending the message than their equilibrium payoff, then the HT refinement outcome is unique as well.

In the next step, let us now motivate the notion of behaviorally consistent hypothesis. One may wonder how “reasonable” are the alternative hypotheses? Suppose there is a version of the above pooling PBE in which the employer matches workers signaling no education with manual jobs \textit{whatever} her posterior belief is. In this case, any alternative hypothesis asserting that no education is sent by either high-ability workers, low-ability workers, or both workers’ types, will support an out-of-equilibrium belief. Consider the hypothesis conjecturing that the low-ability workers obtain education whereas the high-ability workers signal no education. This hypothesis justifies the posterior belief that no education is a signal of high-ability. Suppose that the employer also updates the hypothesis on-the-equilibrium message (i.e., education). Then, she will conclude that education is sent only by low-ability workers, and thus the employer will assign the manual job. As a reaction to this, a low-ability worker - for whom education is costly - will signal no education, thus contradicting the separating behavior conjectured by the alternative hypothesis. In the wake of this reasoning, the hypothesis appears “implausible” and it shall be refuted. A similar argument was provided by Mailath (1988) against the posterior beliefs underlying the reasoning behind the Intuitive Criterion.

Notice that the argument against the above hypothesis is triggered by the employer’s preference reversal on-the-equilibrium path. For this reason, we suggest a strengthening of the HT refinement to exclude such “implausible” and “refutable” hypotheses. More precisely, we require that alternative hypotheses are behaviorally consistent. A hypothesis is said to be \textit{behaviorally consistent} if, after updating it on-the-equilibrium path, it rationalizes the receiver’s equilibrium behavior. It is then shown that if the refinement outcome of the Intuitive Criterion is unique, the uniquely selected posterior belief can always be justified by
a behaviorally consistent hypothesis, thus making the belief immune against Mailath's critic.

The refinement approach based on behaviorally consistent hypotheses is an empirically relevant concept. We elucidate that the novel refinement can very well accommodate the observed equilibrium behavior in the experimental study of Brandts and Holt (1992). In particular, in one of their treatments, a majority of subjects behaved in line with a pooling PBE that is ruled out by the Intuitive Criterion. Yet, the “unintuitive” equilibrium passes the HT refinement yielding a unique posterior belief justified by a behaviorally consistent hypothesis. It is also noteworthy that subjects who acted in the role of senders behaved according to the alternative hypothesis underlying the unique posterior belief of the “unintuitive” PBE.

Finally, we address the question of how the different notions of hypothesis testing affect the refinement outcome. This question is closely relate to the existence of Hypothesis Testing Equilibrium (HTE), a solution concept on which the novel refinement builds. When hypotheses are about sender’s pure strategies, it is shown that the existence of an HTE is not guaranteed regardless of whether the strategies are best responses or not. However, if an alternative hypothesis is about sender’s mixed behavior, then an HTE always exists whenever a pure PBE exists. But, the existing HTEs will not refine the posterior beliefs of a given PBE. Then, there are too many “mixed” hypothesis, and thus all posterior beliefs can be justified.

The reminder of the paper is organized as follows: In Section 2, we recall the HT updating rule and formalize the idea of Hypothesis Testing Equilibrium. In Section 3, we compare the HT refinement with the Intuitive Criterion. In Section 4, we introduce the notion of Behaviorally Consistent Hypothesis Testing Equilibrium. In Section 5, we address the existence question of Hypothesis Testing Equilibrium. In Section 6, we conclude.

2 Hypothesis Testing Equilibrium.

In signaling games, there are two players: an informed sender and an uninformed receiver. Nature draws a type for the sender from a set of types $\Theta$ according to a 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 a 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 mappings $u_S, u_R : \Theta \times M \times A \rightarrow \mathbb{R}$. It is assumed that $p$ is known by the players and that all sets $\Theta, M$ and $A$ are finite.

Given a message $m$, the receiver’s posterior beliefs are represented by $\mu(\cdot \mid m)$, a prob-
ability distribution over $\Theta$. We denote by $\mu = \{\mu(\cdot | m)\}_{m \in M}$ a family of posterior beliefs. In this paper, we consider signaling games for which a Perfect Bayesian Equilibrium (PBE) in pure strategies exists.

**Definition 1** $(s^*, r^*, \mu^*)$ constitute a pure Perfect Bayesian Equilibrium for a signaling game if:

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

(ii) $r^*(m) \in \arg\max_{a \in A} \sum_{\theta \in \Theta} \mu^*(\theta | m) u_R(\theta, m, a) \quad \forall m \in M$,

(iii) $\mu^*(\theta | m) := \frac{\pi(\theta, m)}{\pi(\Theta, m)}$ if $\pi(\Theta, m) > 0$, and

$\mu^*(\cdot | m)$ is an arbitrary probability distribution over $\Theta$ if $\pi(\Theta, m) = 0$, where

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

Conditions (i) and (ii) ensure sequential rationality. That is, all sender’s types best respond to the receiver’s (optimal) strategy, and the receiver best responds to every message by taking into account her posterior beliefs. Condition (iii) specifies how the receiver’s posterior beliefs are derived. If a message is sent with a positive probability, Bayes’ rule is applied. However, for out-of-equilibrium messages, Bayes’ rule does not work and posterior beliefs are determined arbitrarily.

In the Hypothesis Testing Equilibrium of Ortoleva (2012), the receiver derives her posterior beliefs by “testing hypotheses”. Let $\Delta(\Theta \times M)$ denote the set of all probability distributions over $\Theta \times M$ and $\rho$ be a prior over $\Delta(\Theta \times M)$, called the second-order prior. It is assumed that $\rho$ has a finite support $\text{supp}(\rho)$. The elements of $\text{supp}(\rho)$ (i.e., the first-order priors) are called hypotheses.

How are the posterior beliefs derived? Before any information is revealed, the receiver adopts a hypothesis $\pi^*$ with the highest likelihood according to the second-order prior $\rho$. That is,

$$\{\pi^*\} = \arg\max_{\pi \in \Delta(\Theta \times M)} \rho(\pi).$$

(1)

After a message $m$ is observed, the receiver tests this hypothesis. If $\pi^*$ assigns a strictly
positive probability to the message observed, i.e.,
\[ \pi^*(\Theta, m) = \sum_{\theta \in \Theta} \pi^*(\theta, m) > 0, \quad (2) \]
then the receiver accepts her initial hypothesis and updates her posterior belief using \( \pi^* \) in the standard Bayesian manner. However, if the hypothesis \( \pi^* \) assigns a zero probability to the observed message, i.e., \( \pi^*(\Theta, m) = 0 \), then the receiver concludes that \( \pi^* \) is a wrong hypothesis and discards \( \pi^* \). In this case, the receiver follows Bayes’ rule to update her second-order prior \( \rho \) in the face of the unexpected message \( m \), and then she chooses a new hypothesis \( \pi^{**}_m \) with the highest likelihood with respect to the updated second-order prior, i.e.,
\[ \{\pi^{**}_m\} = \arg \max_{\pi \in \Delta(\Theta \times M)} \rho(m \mid \pi), \quad (3) \]
where
\[ \rho(m \mid \pi) = \frac{\pi(\Theta, m)\rho(\pi)}{\sum_{\pi' \in \text{supp}(\rho)} \pi'(\Theta, m)\rho(\pi')} \quad (4) \]
The new hypothesis \( \pi^{**}_m \) is used to derive her posterior beliefs over \( \Theta \) by using Bayes’ rule again. The Hypothesis Testing (HT) updating procedure is formally summarized below.

**Definition 2** The receiver having a second-order prior \( \rho \) over \( \Delta(\Theta \times M) \) with a finite support, \( \text{supp}(\rho) = \{\pi_1, \ldots, \pi_N\} \), derives her family of posterior beliefs \( \mu_\rho = \{\mu_\rho(\cdot \mid m)\}_{m \in \mathcal{M}} \) as follows:

(i) \( \mu_\rho(\theta \mid m) := \frac{\pi^*(\theta, m)}{\pi^*(\Theta, m)} \) if \( \pi^*(\Theta, m) > 0 \), where \( \{\pi^*\} = \arg \max_{\pi \in \Delta(\Theta \times M)} \rho(\pi) \), and

(ii) \( \mu_\rho(\theta \mid m) := \frac{\pi^{**}_m(\theta, m)}{\pi^{**}_m(\Theta, m)} \) if \( \pi^*(\Theta, m) = 0 \), where \( \{\pi^{**}_m\} = \arg \max_{\pi \in \Delta(\Theta \times M)} \rho(m \mid \pi) \).

Notice that posterior beliefs are well-defined if, for every message \( m \in \mathcal{M} \), there is a hypothesis \( \pi \in \text{supp}(\rho) \) with \( \pi(\Theta, m) > 0 \).

Following [Ortoleva, 2012], a hypothesis refers to receiver’s joint beliefs about the sender’s pure strategies that best respond to some of her pure strategy. For a given signaling game, a prior \( \pi \) is called a (strong) hypothesis if there exist strategies \( s : \Theta \to \mathcal{M} \) and \( r : \mathcal{M} \to \mathcal{A} \)

\[ ^6 \text{In the following, whenever we use term “hypothesis” we refer to the strong notion of hypothesis à la Ortoleva. When other notions of hypothesis are meant, it will be stated explicitly.} \]
such that, for every \((\theta, m) \in \Theta \times M\):

\[
\pi(\theta, m) = \begin{cases} 
p(\theta), & \text{if } s(\theta) = m, \text{ and } m \in \arg \max_{m' \in M} u_S(\theta, m', r(m')), \\
0, & \text{otherwise.}
\end{cases}
\]

In words, a hypothesis \(\pi\) ascribes probability \(p(\theta)\) to the joint event “type \(\theta\) sends message \(m\)” if \(s(\theta) = m\) is a best response for type \(\theta\) against some strategy of the receiver. Notice that any hypothesis is consistent with \(p\), the initial information about the sender’s types.

By assuming that \(\text{supp}(\rho)\) consists of (strong) hypotheses, we can formalize the Hypothesis Testing Equilibrium.

**Definition 3** \((s^*, r^*, \mu^*, \rho)\) constitute a Hypothesis Testing Equilibrium for a signaling game if:

\[
\begin{align*}
(i) & \quad s^*(\theta) \in \arg \max_{m \in M} u(\theta, m, r^*(m)) \quad \forall \theta \in \Theta, \\
(ii) & \quad r^*(m) \in \arg \max_{a \in A} \sum_{\theta \in \Theta} \mu^*_\rho(\theta | m) u_R(\theta, m, a) \quad \forall m \in M, \text{ and} \\
(iii) & \quad \mu^*_\rho(\cdot | m) \text{ is derived by applying the HT updating rule.}
\end{align*}
\]

Condition (i) requires that the sender best responds against a given receiver’s strategy as in PBE. However, condition (ii) is stronger than under PBE in the sense that the receiver best responds to all messages with respect to unique posterior beliefs. Notice also that each HTE is a PBE and therefore the HTE concept can be used as a refinement criterion for the multiple PBEs.

A given PBE is said to pass the Hypothesis Testing (HT) refinement if there exists an HTE that supports the given PBE behavior. Since there is a finite number of types and messages, there can only be a finite number of hypotheses and thus the HT refinement will select a finite collection of posteriors beliefs, thus refining the out-of-equilibrium posterior beliefs of a given PBE. The following example illustrates how the refinement approach works.

Consider a discrete version of the labor-market signaling game in the spirit of Spence (1973) mentioned in the Introduction. The game is depicted in Figure 1. It was implemented in an experimental study by Brandts and Holt (1992).\(^7\) A worker applying for a job has either low skills \((t_1)\) or high skills \((t_2)\). Knowing his type, the worker decides whether to

\(^7\)It shall be emphasized that our results are not restricted to games with \(|\Theta| = |M| = |A| = 2\), but apply to any (finite) signaling games.
make an investment in schooling ($S$) or not ($N$). An employer observes the signal but not worker’s skills and assigns the job applicant to either an executive job ($e$) or a manual job ($m$). Notice that both worker’s types prefer the executive job regardless of the education status and that schooling is more costly for the unskilled worker. For the employer, schooling is not productive since her payoff is unaffected by the signal. Thus, the employer prefers to match a skilled worker with the executive job and the match an unskilled worker with the manual job.

There are two (pure) PBEs: In the first PBE, both types of the worker obtain education. In the second pooling PBE, both types of the worker decide against schooling.

Consider the first pooling PBE; that is, the strategy profile

$$s^*(t_1) = s^*(t_2) = S, \quad r^*(S) = e, \quad r^*(N) = m,$$  \hspace{1cm} (5)

together with a family of posterior beliefs such that $\mu^*(t_1 \mid S) = 1/3$ and $\mu^*(t_1 \mid N) \geq 1/2$. This equilibrium can be supported by an HTE. There are three possible candidates for hypotheses:

- $\pi_1 := \{\pi_1(t_1, N) = 1/3, \pi_1(t_2, S) = 2/3\}$ since $s(t_1) = N$ and $s(t_2) = S$ is the sender’s best-response strategy against $r(S) = m$ and $r(N) = m$,

- $\pi_2 := \{\pi_2(t_1, N) = 1/3, \pi_2(t_2, N) = 2/3\}$ since $s(t_1) = N$ and $s(t_2) = N$ is the sender’s best-response against $r(S) = e$ and $r(N) = m$,,
• $\pi_3 := \{\pi_3(t_1, S) = 1/3, \pi_3(t_2, S) = 2/3\}$ since $s(t_1) = S$ and $s(t_2) = S$ is the sender’s best-response against $r(S) = e$ and $r(N) = e$.

The pooling HTE is where $\text{supp}(\rho) = \{\pi_1, \pi_3\}$ and $\rho(\pi_1) < \rho(\pi_3)$ (i.e., $\pi^* = \pi_3$ and $\pi^{**} = \pi_1$). When observing no schooling, the receiver adopts a new hypothesis $\pi_1$. According to this hypothesis, the employer reasons that there is a “type-dependence”; i.e., the low-skilled worker decides against education and the high-skilled worker chooses schooling. Hence, the employer infers that the unskilled workers do not obtain education. Therefore, the pooling PBE passes the HT refinement yielding the unique out-of-equilibrium belief, $\mu^*_\rho(t_1 \mid N) = 1$.

However, the second pooling PBE where

$$s^*(t_1) = s^*(t_2) = N, \quad r^*(S) = m, \quad r^*(N) = e,$$

and $\mu^*(t_1 \mid S) \geq 1/2$ and $\mu^*(t_1 \mid N) = 1/3$, does not survive the HT refinement. In this case, hypothesis $\pi_2$ rationalizes the employer’s behavior when no schooling is observed. However, when schooling is observed, there is no hypothesis that can rationalize the employer’s decision in favor of a manual job. By updating hypothesis $\pi_1$ or $\pi_3$, the receiver concludes that education is more likely to be chosen by a high-ability worker and thus the worker would be matched with an executive job. Therefore, there does not exist an HTE supporting the second pooling PBE behavior.

Bradt and Holt (1992) conducted a series of experiments on equilibrium behavior in the context of signaling games just described. There are a few interesting observations from this study. First of all, the reported behavior can be explained very well by the HTE notion and its extension that will be introduced in Section 3. A vast majority of subjects behaved in line with the first pooling PBE (135 out of 166 decisions matched the equilibrium behavior with pooling on $S$). This result may be explained by the fact that pooling on $S$ but not pooling on $N$ is consistent with an HTE. Second, and more important, the sender’s behavior supports the “type-dependence” underlying the receiver’s alternative hypothesis $\pi_2$. All subjects who acted in the role high-ability senders sent message $S$. However, 79% of the low skill type subjects signaled $N$ in the first games played (part a), followed by 42% and 25% in the consecutive games (parts b and c of Treatment 1; see Brandts and Holt, 1992, p. 1357).

Although the percentage diminishes with more experienced gained in consecutive games,

---

8The existence question is further explored in Section 5.

9Brandts and Holt (1992) argue that the low skill worker’s decision to send $N$ might be motivated by the belief that the receiver always assigns an executive job. This belief is, however, contradicted by the decisions of the receivers who always responded with $e$ to signal $S$ and almost always (i.e., 22 out of 29) replied with $m$ to signal $N$. 

---
This data provides sound evidence that the “type-dependence” is a reasonable explanation for the sender’s out-of-equilibrium behavior.

3 HTE and Intuitive Criterion: A Comparison

In this section, we compare the HT refinement outcome with the Intuitive Criterion.

The Intuitive Criterion does not build on any theory of updating beliefs. Instead, the refinement idea is to select out-of-equilibrium beliefs which are “plausible” in the sense of ascribing a zero probability to all sender’s types who can never benefit by deviating from a fixed equilibrium.

For a given PBE, let \( m^o \) be an out-of-equilibrium message. Denote by \( u^*_S(\theta) \) the sender’s equilibrium payoff when his type is \( \theta \in \Theta \). Let \( T(m^o) \subseteq \Theta \) be the set of types who cannot improve upon their equilibrium payoff by choosing \( m^o \). That is, for all \( \theta \in T(m^o) \):

\[
  u^*_S(\theta) > \max_{a \in BR(\Theta, m^o)} u_S(\theta, m^o, a),
\]

(7)

where

\[
  BR(\Theta, m^o) := \bigcup_{\{\mu : \mu(\Theta|m^o)=1\}} BR(\mu, m^o)
\]

is a set of receiver’s best replies against \( m^o \) with respect to posterior 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 \( m^o \). If there exists a type \( \theta' \in \Theta \) such that

\[
  u^*_S(\theta') < \min_{a \in BR(I(m^o), m^o)} u_S(\theta', m^o, a),
\]

(8)

where

\[
  BR(I(m^o), m^o) := \bigcup_{\{\mu : \mu(I(m^o)|m^o)=1\}} BR(\mu, m^o)
\]

is a set of receiver’s best replies against \( m^o \) with respect to posterior beliefs defined only on types in \( I(m^o) \), then the PBE is said to fail the Intuitive Criterion. If, however, the PBE passes the Intuitive Criterion, the out-of-equilibrium posterior beliefs are refined by admitting probability distributions that assign a zero measure to every type in \( T(m^o) \) (i.e., \( \mu(\theta | m^o) = 0 \) for all \( \theta \in T(m^o) \)). In the special case where \( I(m^o) = \{\theta\} \) is a singleton set,

\[
  \text{More precisely, if } (s^*, r^*, \mu^*) \text{ is the given PBE, then } u^*_S(\theta) = u_S(\theta, s^*(\theta), r^*(m)).
\]

\[
\text{Specifically, } BR(\mu, m^o) := \arg \max_{a \in A} \mu(\theta | m^o) u_R(\theta, m^o, a).
\]
the receiver learns the type who could benefit from sending the out-of-equilibrium message \(m^\circ\) (i.e., \(\mu(\theta | m^\circ) = 1\)). A class of signaling games in which \(I(m^\circ)\) is a singleton set for every out-of-equilibrium message \(m\) is important for economic applications since in such games, the refinement outcome of the Intuitive Criterion is always unique.

Consider again the first pooling PBE of the labor-market game in Figure 1. Given the worker’s equilibrium payoff, only the low skill type could be better off by sending the out-of-equilibrium message \(N\) (given the worker’s belief that the executive job follows signal \(N\)). Thus, \(I(N) = \{t_1\}\) and \(T(N) = \{t_2\}\). As long as the employer concludes that no education is solely chosen by an unskilled worker (i.e., \(\mu^* (t_1 | N) = 1\)), there is no type who has an incentive to signal \(N\). Thus, the pooling PBE passes the Intuitive Criterion yielding the same posterior belief as the HT refinement.\(^{12}\)

In the labor-market game of Figure 1, the outcomes of the two refinement approaches coincide. However, as Proposition 1 asserts, the two refinement criteria are in general not “nested” in any sense. That is to say, (1) there exists a PBE which passes the HT refinement but fails the Intuitive Criterion; and (2) there exists a PBE which passes the Intuitive Criterion but fails the HT refinement.\(^{13}\)

**Proposition 1** The refinement outcomes based on the HTE and Intuitive Criterion are not nested.

As a next step, we focus on equilibria that survive against the Intuitive Criterion and examine under which condition such an equilibrium will also pass the HT refinement. It turns out that if the refinement outcome of the Intuitive Criterion is unique, it can always be justified by an HTE. That is, if there is only a single type who could benefit from sending an out-of-equilibrium message \(m^\circ\) in a given PBE (i.e., \(I(m^\circ) = \{\theta^\circ\}\)), then the PBE also passes the HT refinement. However, in such games, the posterior beliefs selected by the HT refinement might be non-unique.

To illustrate this point, consider a modified version of the labor-market game in Figure 1 in which the employer receives \(115^*\) (instead of \(75\)) when a high-skilled worker with no education is matched with the manual job \(m\). There is one pooling PBE where both worker’s

---

\(^{12}\)The second pooling PBE fails the Intuitive Criterion. Only the high skill type can be better off by sending the out-of-equilibrium message \(S\). Thus, the employer concludes that a high-ability worker obtains education (i.e., \(\mu^* (t_2 | S) = 1\)) and best responds with \(e\). This in turn leads the high-ability worker to signal \(S\) instead of \(N\).

\(^{13}\)Sun (2016) compares the Hypothesis Testing refinement outcome with the Intuitive Criterion under stronger behavioral assumptions than ours. In her framework, whenever a PBE fails the Intuitive Criterion, it cannot be explained by an HTE.
types decide for schooling and the employer chooses \( e \) after observing \( S \) and \( m \) after observing signal \( N \). This behavior is rationalized by any family of posterior beliefs such that \( \mu^* (s_1 \mid S) = 1/3 \) and \( \mu^* (s_1 \mid N) \geq 1/6 \). There exist two HTEs supporting the pooling PBE. There are three candidates for hypotheses:

- \( \pi_1 := \{ \pi_1(t_1, N) = 1/3, \pi_1(t_2, S) = 2/3 \} \) since \( s(t_1) = N \) and \( s(t_2) = S \) is the sender’s best-response strategy against \( r(S) = m \) and \( r(N) = m \),

- \( \pi_2 := \{ \pi_2(t_1, N) = 1/3, \pi_2(t_2, N) = 2/3 \} \) since \( s(t_1) = N \) and \( s(t_2) = N \) is the sender’s best-response against \( r(S) = m \) and \( r(N) = e \),

- \( \pi_3 := \{ \pi_3(t_1, S) = 1/3, \pi_3(t_3, S) = 2/3 \} \) since \( s(t_1) = S \) and \( s(t_2) = S \) is the sender’s best-response against \( r(S) = e \) and \( r(N) = m \).

The first pooling HTE is where \( \text{supp}(\rho) = \{ \pi_1, \pi_3 \} \) and \( \rho(\pi_1) < \rho(\pi_3) \) (i.e., \( \pi^* = \pi_3 \) and \( \pi^{**} = \pi_1 \)). In this HTE, the employer adopts the alternative hypothesis \( \pi_1 \) according to which she infers that the out-of-equilibrium signal \( N \) is chosen by a low-skilled worker (i.e., \( \mu^*_p (t_1 \mid N) = 1 \)). Notice that this posterior belief is also selected by the Intuitive Criterion (since \( T(N) = \{ t_2 \} \) and \( I(N) = \{ t_1 \} \)). The second HTE is where \( \text{supp}(\rho) = \{ \pi_2, \pi_3 \} \) and \( \rho(\pi_2) < \rho(\pi_3) \) (i.e., \( \pi^* = \pi_3 \) and \( \pi^{**} = \pi_2 \)). In this HTE, the employer switches to the alternative hypothesis \( \pi_2 \). According to this hypothesis observing no education conveys the same information about the sender’s types as the initial prior \( p \) (i.e., \( \mu^*_p (t_1 \mid N) = 1/3 \)).

In sum, the above pooling PBE passes the HT refinement yielding two families of posterior beliefs while only the first one coincides with the Intuitive Criterion outcome.

To guarantee uniqueness, a further condition is required. If, in addition to \( I(m^o) \) being a singleton set, every out-of-equilibrium message \( m^o \) is a never-best response for all types who cannot be better off by sending \( m^o \) than their equilibrium payoff (i.e., for all \( \theta \in T(m^o) \)), then the HTE refinement outcome is unique, and it coincides with the Intuitive Criterion outcome.

**Proposition 2** Consider a PBE. Let \( M^o \) be the set of all out-of-equilibrium messages. If, for every \( m^o \in M^o \), the set \( I(m^o) \) is a singleton set and the PBE passes the Intuitive Criterion, then there exists an HTE yielding the same posterior beliefs as the Intuitive Criterion.

---

14 Of course, there is another HTE where \( \text{supp}(\rho) = \{ \pi_1, \pi_2, \pi_3 \} \) and \( \pi^* = \pi_3 \). Depending on whether the updated second-order prior \( \rho(N \mid \pi) \) ascribes the highest weight to either \( \pi_1 \) (i.e., \( \pi_1 = \pi^{**} \)) or to \( \pi_2 \) (i.e., \( \pi^{**} = \pi_2 \)), we obtain the same two HTEs as described above.
Moreover, if, for every \( m^0 \in \mathcal{M} \), it is true that
\[
m^0 \neq s(\theta) \in \arg\max_{m \in \mathcal{M}} u(\theta, m, r(m)) \text{ for any strategy } r: \mathcal{M} \to \mathcal{A} \text{ and every type } \theta \in T(m^0),
\]
then the HT refinement outcome is unique.  

In signaling games in which sending an out-of-equilibrium message \( m^0 \) is dominated by another message for any type who cannot be better off than their equilibrium payoff, i.e., \( \theta \in T(m^0) \), the uniqueness condition (9) is naturally satisfied, but not vice versa. \(^{15}\)

It should also be noted that the Intuitive Criterion admits arbitrary beliefs over \( I(m^0) \), if \( I(m^0) \) is a non-singleton set for an out-of-equilibrium message \( m^0 \). In the extreme case where \( I(m^0) = \Theta \), the Intuitive Criterion does not reduce the set of posterior beliefs at all while HTE always does, provided it exists. An example demonstrating this case is provided in Appendix B.

### 4 Behaviorally Consistent Hypothesis Testing Equilibrium

So far, alternative hypotheses were solely used to rationalize the receiver’s off-the-equilibrium behavior. Her on-the-equilibrium behavior was neglected. However, it is conceivable that the receiver adopts an alternative hypothesis which, after updating, might rationalize a different action than her best response on-the-equilibrium path. To avoid such possibilities, we propose a strengthening of the hypothesis notion.

We require the receiver’s behavior that is rationalized by her most likely hypothesis to be maintained under all alternative hypotheses. Formally, a set of hypothesis \( \text{supp}(\rho) \) is said to be behaviorally consistent with respect to the most likely hypothesis \( \pi^* \) if for every message \( m \in \mathcal{M} \) such that \( \pi^*(\Theta, m) > 0 \) and every action \( a^* \in \mathcal{A} \) such that
\[
a^* \in \arg\max_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \frac{\pi^*(\theta, m)}{\pi^*(\Theta, m)} u_R(\theta, m, a),
\]

\(^{15}\)Notice that the reverse direction of Proposition 2 does not hold. Consider a pooling behavior in the signaling game of Figure 2 in which both types choose \( N \), and the receiver chooses \( m \) when he observes \( S \) and \( e \) when he observes \( N \). Given this equilibrium behavior, we have a singleton set \( I(S) = \{t_2\} \). The pooling equilibrium passes the HT refinement, but it does not survive against the Intuitive Criterion.
it is true that $\pi(\Theta, m) > 0$ and

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

(11)

for all $\pi \in \text{supp}(\rho)$. By imposing behavioral consistency, we obtain a Behaviorally Consistent Hypothesis Testing Equilibrium.

**Definition 4** $(s^*, r^*, \mu^*, \rho)$ constitute a Behaviorally Consistent Hypothesis Testing Equilibrium (BCHTE) for a signaling game if, $(s^*, r^*, \mu^*, \rho)$ is a HTE and $\text{supp}(\rho)$ contains behaviorally consistent hypotheses with respect to $\pi^*$, the maximum likelihood hypothesis according to $\rho$.

In a BCHTE, the receiver observes an out-of-equilibrium message and adopts an alternative hypothesis which preserves her optimal behavior on-the-equilibrium path. Since the set of behaviorally consistent hypotheses is included in the set of hypotheses previously studied, the BCHTE notion narrows down HTE and thus it can be used as a further refinement of PBE.

Our primary rationale for introducing BCHTE is Mailath’s (1988) criticism of the Intuitive Criterion. To elucidate Mailath’s argument, suppose that a PBE fails the Intuitive Criterion. Typically, such an equilibrium is regarded as being “unintuitive” or “implausible”. However, as pointed out by Mailath, the reasoning based on the Intuitive Criterion
might lead to “internal inconsistencies”. More specifically, the posterior beliefs induced by
the Intuitive Criterion might lead to behavioral implications which in turn might provide ev-
idence against the induced belief. If, however, a given PBE fails the Intuitive Criterion and
such an internal inconsistency is present, then the equilibrium might be abandoned although
the posterior belief used against the equilibrium is “implausible” itself.

To illustrates Mailath’s critic, consider the signaling game depicted in Figure 2. It is
another version of the labor-market game experimentally studied by [Brandts and Holt(1992)].
In this game, there exist two pooling PBEs. The first pooling PBE is where

\[
s^*(t_1) = s^*(t_2) = N, \; r^*(S) = m, \; r^*(N) = e, \; \mu^*(t_1 \mid N) = \frac{1}{3}, \text{ and } \mu^*(t_1 \mid S) \geq \frac{1}{2}.
\]

The Intuitive Criterion asserts that only a high-ability worker could benefit from sending the
out-of-equilibrium message \(S\) (i.e., \(I(S) = \{t_2\}\)) and that the employer should infer from
this signal that the job applicant has high skills (i.e., \(\mu(t_2 \mid S) = 1\)). Knowing this, however,
the employer will match the worker with the executive job \(e\) instead of matching him with the
manual job \(m\). Since the pooling PBE fails the Intuitive Criterion, it’s behavior is regarded
as being “unintuitive”. Hence, we call it the “unintuitive” PBE.

However, the employer might reason further. Since applicants with education receive
the executive job \(e\), a skilled worker is strictly better off by signaling \(S\) instead of \(N\). If
the employer is consistent with her reasoning, then she should infer that only an unskilled
worker does not obtain education and she will best respond by offering the manual job \(m\) to
the applicant signaling \(N\). This in turn will induce the unskilled worker to signal \(S\).
Thus, the concluded behavior contradicts the posterior belief that only high-ability workers
choose education, causing the internal inconsistency. For this reason, the employer should
discard the posterior belief induced by the Intuitive Criterion reasoning rather then refuting
the equilibrium behavior itself.

Notice that the receiver’s behavioral inconsistency is necessary for the internal inconsis-
tency. Therefore, a BCHTE supporting a given PBE can be viewed as an argument in favor
of the equilibrium even though it fails the Intuitive Criterion. For instance, in this game,
the “unintuitive” PBE is supported by a BCHTE with \(\rho\) such that \(\text{supp}(\rho) = \{\pi_1, \pi_2\}\) and
\(\rho(\pi_1) > \rho(\pi_2)\) where

\[\text{supp}(\rho) = \{\pi_1, \pi_2\}\] and \(\rho(\pi_1) > \rho(\pi_2)\) where \(\pi_1 = \{\pi_1(t_1, S) = 1/3, \pi_1(t_1, S) = 2/3\}\) and
\(\pi_2 = \{\pi_2(t_1, N) = 1/3, \pi_2(t_2, S) = 2/3\}\).
• \( \pi_1 := \{ \pi_1(t_1, N) = 1/3, \pi_1(t_2, N) = 2/3 \} \) since \( s(t_1) = N \) and \( s(t_2) = N \) is the sender’s best-response strategy against \( r(S) = m \) and \( r(N) = e \),

• \( \pi_2 := \{ \pi_2(t_1, S) = 1/3, \pi_2(t_2, N) = 2/3 \} \) since \( s(t_1) = S \) and \( s(t_2) = N \) is the sender’s best-response against \( r(S) = m \) and \( r(N) = m \).

The most likely hypothesis \( \pi^* = \pi_1 \) describes the worker’s pooling strategy on \( N \). After updating \( \pi_1 \) in the face of \( N \), the employer concludes that no education is more likely to be chosen by skilled workers and thus the job applicant will be matched with the executive job \( e \).

This decision remains optimal under the alternative hypothesis \( \pi^{**} = \pi_2 \). According to this hypothesis, there is a “reverse” type-dependence; a low-ability worker does not obtain education while a high-ability worker does. Since the employer’s behavior on-the-equilibrium path is preserved under \( \pi_2 \), the employer has no reason to deviate from her equilibrium strategy. At this stage, the argument leading to an internal inconsistency is broken and the employer has no reason to refute her posterior belief induced by the alternative hypothesis (i.e., \( \mu(t_1 | S) = 1 \)).

The empirical results in Brandts and Holt (1992) on equilibrium behavior in the game of Figure 2 are particularly interesting. Surprisingly, a majority of subjects behaved in line with the “unintuitive” PBE. In their Treatment 5, among all the subjects’ decisions that were consistent with either of the two PBEs, 72% of the decisions in part (a) and 84% of the decisions in part (b) matched the “unintuitive” equilibrium. The empirical distributions indicate that the BCHTE notion can better explain the observed behavior than the Intuitive Criterion. While the Intuitive Criterion asserts that only the “intuitive” equilibrium behavior shall be observed, the BCHTE notion makes both equilibria equally likely. Since both PBEs pass the BCHTE refinement, there is no reason to favor one equilibrium over the other one within the theory. The slight bias towards the “unintuitive” equilibrium could be due to the observed tendency toward the “reverse” type-dependence of the senders. The tendency became more pronounced in part (b) where 85% of the low-ability types signaled no education whereas 48% of the low-ability types sent message \( S \). These frequencies provide evidence supporting the alternative hypothesis \( \pi_2 \) of the “unintuitive” BCHTE rather than the alternative hypothesis \( \pi_2' \) of the “intuitive” BCHTE.

Since posterior beliefs justified by behaviorally consistent hypotheses are immune against Mailath’s critic, it is interesting to ask whether the (unique) Intuitive Criterion outcome can

\[17\] Obviously, not all equilibria that fail the Intuitive Criterion can be “defeated” by the BCHTE notion. However, deriving conditions under which such equilibria can be supported by BCHTEs is beyond the scope of this paper.

\[18\] See Footnote 16.
also be supported by a BCHTE. As we already know, if a PBE passes the Intuitive Criterion and \( I(m^\circ) \) is a singleton set for every out-of-equilibrium message \( m^\circ \), then there always exists a hypothesis justifying the unique posterior belief (i.e., \( \mu(\theta^\circ \mid m^\circ) = 1 \), see Proposition 2). However, such hypotheses do not need to be behaviorally consistent. To underpin the unique Intuitive Criterion outcome by a behaviorally consistent hypothesis, an auxiliary condition is required.

Let \((s^*, r^*, \mu^*)\) be fixed PBE and \( \pi \) be a probability distribution on \( \Theta \times M \) associated with the sender’s equilibrium strategy \( s^* \), i.e., \( \pi(\theta, m) = p(\theta) \) if, \( s^*(\theta) = m \) and \( \pi(\theta, m) = 0 \), otherwise. A type \( \theta_d \) with \( \pi(\theta_d, m) > 0 \) is called a dummy for the on-equilibrium message \( m \) if, for every \( a \in A \):

\[
a \in \arg \max_{a \in A} \sum_{\theta \in \Theta} \pi(\theta, m) u_R(\theta, m, a) \quad \text{implies} \quad a \in \arg \max_{a \in A} \sum_{\theta \in \Theta} \pi(\theta, m) u_R(\theta, m, a),
\]

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

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

In other words, the receiver regards a type who sends an on-the-equilibrium message \( m \) as a dummy in respect of the message, if her best response behavior given \( m \) remains unchanged under a probability distribution which ascribes a zero-probability to the event “type \( \theta \) sends a message \( m \)”. If type \( \theta \) is a dummy for an on-the-equilibrium-message \( m \) in a given equilibrium, then the message \( m \) does not convey any information about the type \( \theta \) that is relevant for the receiver to derive her optimal strategy in response to message \( m \).

**Proposition 3** Consider a PBE that passes the Intuitive Criterion and for which \( I(m^\circ) = \{\theta^\circ\} \) is a singleton set for every out-of-equilibrium message \( m^\circ \). If type \( \theta^\circ \) is a dummy for some on-the-equilibrium message, then there exists a BCHTE justifying the unique Intuitive Criterion outcome.

In the “intuitive” PBE in the labor-market game in Figure 2, the unskilled worker is a dummy for education. That is to say, even if the employer considers that low-skilled workers never obtain education, it does not change her equilibrium behavior on-the-equilibrium path. Therefore, there exists a BCHTE supporting the Intuitive Criterion outcome (see Footnote 16).
5 Existence of Hypothesis Testing Equilibrium

In this section, we explore how various notions of hypothesis affect the existence of an HTE.

Consider a signaling game for which a pure PBE exists. In general, existence of an HTE supporting a given PBE is not guaranteed.

For an HTE to exist, the following two conditions must be satisfied:

(i) For every message $m \in \mathcal{M}$, there exists at least one type $\theta \in \Theta$ who chooses $m$ as a sender’s best response against some pure strategy of the receiver (i.e., an alternative hypothesis exists).

(ii) The receiver’s behavior out-of-equilibrium path in the given PBE can be rationalized by posterior beliefs obtained via updating of an alternative hypothesis derived in (i).

Therefore, even though a set of alternative hypotheses exist, it can be that none of the alternative hypotheses will rationalize the receiver’s out-of-equilibrium behavior. (To illustrate non-existence of an HTE due to failure of condition (ii) see Figure 4 in Appendix C).

One may argue that the notion of (strong) hypothesis is too restrictive to guarantee existence. Can we obtain existence of an HTE under a weaker notion of hypotheses tested by the receiver? Suppose that one assumes that hypothesis is about sender’s pure strategy behavior (which is not necessarily a best response to any strategy of the receiver). That is, a prior $\pi$ over $\Theta \times \mathcal{M}$ is called a weak hypothesis if there exist a strategy $s : \Theta \rightarrow \mathcal{M}$ such that, for every $(\theta, m) \in \Theta \times \mathcal{M}$:

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

Denote by $\text{supp}(\rho^W)$ a set consisting of weak hypotheses and call $(s^*, r^*, \mu^*, \rho^W)$ together with $\text{supp}(\rho^W)$ satisfying conditions (i) through (iii) in Definition 3, a Weak Hypothesis Testing Equilibrium (WHTE). For instance, the second PBE with pooling on $N$ in the signaling game of Figure 1 can be explained by a WHTE with $\rho^W = \{\pi_1^W, \pi_2^W\}$ such that $\rho^W(\pi_1^W) > \rho^W(\pi_2^W)$ where $\pi_1^W := \{\pi_1^W(t_1, N) = 1/3, \pi_1^W(t_1, N) = 2/3\}$ and $\pi_2^W := \{\pi_2^W(t_1, S) = 1/3, \pi_2^W(t_1, N) = 2/3\}$. Notice also that the weak hypothesis $\pi_2^W$ is behaviorally consistent.

It is not surprising, however, that existence of a WHTE is not guaranteed either. That is, if the receiver takes for granted that sending an out-of-equilibrium message is realization

---

19An example where (strong) hypothesis cannot be constructed is a PBE in which an out-of-equilibrium message is a dominated action for every type $\theta \in \Theta$.

20Recall that we could not find an HTE supporting the second pooling PBE of the labor-market game in Figure 1 exactly for this reason.
of sender’s pure strategy (compatible with the prior information about types \( p \)), then such reasoning is still too restrictive to explain each PBE. In Figure 4 of Appendix C, we present a signaling game with a pooling PBE for which none of its posterior beliefs can be justified by the weak hypothesis notion.

The concept of alternative hypotheses is about how the receiver “hypothetically” reasons about strategies of the sender that would explain her equilibrium behavior. Apparently, the “hypothetical” strategies must be different than the sender’s equilibrium behavior. Hence, when the receiver is surprised by an out-of-equilibrium message, she may consider that the observed message is an outcome of a mixed strategy (even though only pure strategies are played in a given PBE). As the following proposition demonstrates, admitting hypotheses about sender’s mixed behavior warrants the existence of a WHTE.

**Proposition 4** Consider a (pure) PBE. Let \( \mathcal{M}^o \) be the set of all out-of-equilibrium messages and \( \{ \mu(\cdot \mid m^o) \}_{m^o} \in \mathcal{M}^o \) be a family of posterior beliefs associated with the given PBE. Then, a Weak Hypothesis Testing Equilibrium supporting the PBE together with \( \{ \mu(\cdot \mid m^o) \}_{m^o} \in \mathcal{M}^o \) exists if, the receiver adopts an alternative hypothesis defined as a mixture of (weak) hypotheses in \( \text{supp}(\rho^W) \).

Although “mixing” between weak hypotheses guarantees existence of WHTE, there are too many candidates for “mixtures” that pass the hypothesis testing. Therefore, the notion of WHTE admitting mixed behavior cannot be used as a refinement rule for the posterior beliefs of a PBE.

Suppose that we allow for “mixtures” between strong hypotheses (i.e., between sender’s best responses against some strategies of the receiver). Does it guarantee existence of Ortoleva’s HTE? Needless to say, this is only interesting when the before mentioned condition \( (i) \) but not \( (ii) \) is satisfied. That is, there exist at least two hypotheses about sender’s best response behavior, but none of them rationalizes, after updating, the receiver’s out-of-equilibrium behavior in a given PBE. It is, however, possible that none of the possible “mixtures” between strong hypotheses will rationalize the receiver’s out-of-equilibrium behavior. Thus, “mixtures” between sender’s best responses are neither sufficient for the existence of an HTE. Moreover, if an HTE exists, then again, there might be too many “mixtures” passing the HT refinement leaving the posterior beliefs of a given PBE unrefined. (For an example, see Case (c) in Appendix C).

---

\(^{21}\)Sun (2016) proves existence of HTE by restricting players’ payoff structure. We use a different approach. We show that existence of HTE is guaranteed by weakening the notion of hypotheses and without constraining the players’ payoff functions.

\(^{22}\)An example illustrating non-existence of an HTE under mixed hypotheses can be provided upon request.
One open question remains: Can any HTE be supported by a behaviorally consistent HTE? Since BCHTE is a subset of HTE, it is immediate that existence of BCHTE is not guaranteed. However, if “mixtures” between (strong) hypotheses of a given HTE are permitted, then the HTE can be explained by a BCHTE, making the equilibrium beliefs immune against “internal inconsistencies” à la Mailath.

**Proposition 5** Suppose an HTE exists. Then, a Behaviorally Consistent Hypothesis Testing Equilibrium exists if, (1) the receiver adopts an alternative hypothesis being a mixture of hypotheses in \( \text{supp}(\rho) \), and (2) the receiver’s best response correspondence on-the-equilibrium path is a singleton.

As a corollary of this result and Proposition 2, the unique refinement outcome of the Intuitive Criterion can always be supported by a BCHTE. That is, if a PBE passes the Intuitive Criterion and \( I(m^o) \) is a singleton set for every out-of-equilibrium message \( m^o \) (\( \theta^o \) does not need to be a dummy), then the unique posterior belief \( \mu(\theta^o \mid m^o) = 1 \) can by justified by a behaviorally consistent “mixture” between two hypotheses of the HTE supporting the given PBE.

6 Conclusion

In this paper, we explored how different notions of Hypothesis Testing can be used to justify posterior beliefs in a given PBE. In the context of signaling games, when hypotheses admit “mixtures” between pure strategies of the sender, then each PBE can be captured by a weaker notion of Hypothesis Testing Equilibrium. Although this result guarantees existence of the novel equilibrium notion, it disqualifies the idea of testing “mixed” hypotheses as a refinement criterion for the out-of-equilibrium beliefs. However, when hypotheses are about sender’s pure (and/or best response) behavior, then the Hypothesis Testing Equilibrium, provided it exists, narrows down the set of posterior beliefs of a given PBE.

We compared the Hypothesis Testing refinement with the Intuitive Criterion. From an empirical point of view, the former refinement rule can much better accommodate the experimental findings of [Brandts and Holt (1992)] than the Intuitive Criterion. From a theoretical point of view, the Hypothesis Testing refinement performs better when there are multiple types who could benefit from sending an out-of-equilibrium message in a given PBE. If there is only one such type for every out-of-equilibrium message, then the outcome of the Intuitive Criterion is unique and it can always be supported by the Hypothesis Testing Equilibrium.
Another promising application of the Hypothesis Testing refinement is to the problem of equilibrium selection. Selecting equilibria 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 on-equilibrium and out-of-equilibrium beliefs. The suggested Behaviorally Consistent Hypothesis Testing Equilibrium eliminates such inconsistencies and it is worth further exploration.

A Proofs

Proof of Proposition 1. Case (1) shows a PBE which passes the HT refinement but fails the Intuitive Criterion. In Case (2), there is a PBE which fails the former but passes the latter criterion.

Case (1): Consider the signaling game in Figure 2. We have four candidates for hypotheses:

- \( \pi_1 = \{ \pi_1(t_1, S) = 1/3, \pi_1(t_2, N) = 2/3 \} \) since \( s(t_1) = S \) and \( s(t_2) = N \) is the sender’s best-response against \( r(S) = m \) and \( r(R) = m \),

- \( \pi_2 = \{ \pi_2(t_2, N) = 1/3, \pi_2(t_2, N) = 2/3 \} \) since \( s(t_1) = N \) and \( s(t_2) = N \) is the sender’s best-response against \( r(S) = m, r(N) = e \),

- \( \pi_3 = \{ \pi_3(t_1, S) = 1/3, \pi_3(t_2, S) = 2/3 \} \) since \( s(t_1) = S \) and \( s(t_2) = S \) is the sender’s best-response against \( r(S) = e \) and \( r(S) = m \),

- \( \pi_4 = \{ \pi_4(t_1, N) = 1/3, \pi_4(t_2, S) = 2/3 \} \) since \( s(t_1) = N \) and \( s(t_2) = S \) is the sender’s best-response against \( r(S) = e \) and \( r(N) = e \).

In this example, there exists a pooling behavior where both sender’s types choose \( N \). The receiver chooses \( m \) when she observes \( S \) and \( e \) when she observes \( N \). That is, \( s^*(t_1) = s^*(t_2) = N \) and \( r^*(S) = m, r^*(N) = e \) together with a family of posterior beliefs where \( \mu^*(t_1 \mid S) \geq 1/2 \) and \( \mu^*(t_1 \mid N) = 1/3 \) constitutes the pooling PBE. This pooling behavior can be supported by an HTE with \( \rho \) such that \( supp(\rho) = \{ \pi_1, \pi_2 \} \) and \( \rho(\pi_1) < \rho(\pi_2) \) (i.e., \( \pi^* = \pi_2 \) and \( \pi^{**} = \pi_1 \)). When message \( S \) is observed, the receiver adopts a new hypothesis \( \pi_1 \) which induces her posterior belief \( \mu^*_\rho(t_1 \mid S) = 1 \).

By the Intuitive Criterion, type \( t_2 \) could be better off than her equilibrium payoff by sending \( S \) (if the receiver chooses \( e \)). Thus, \( I(S) = \{ t_2 \} \) induces the posterior belief \( \mu(t_2 \mid S) = 1 \). However, when the receiver infers that \( S \) is sent by type \( t_2 \), she will choose \( e \) instead of \( m \). But, given that the receiver plays \( e \), type \( t_2 \) will indeed choose \( S \), violating
Thus, the pooling PBE does not survive against the Intuitive Criterion but it passes the HT refinement.

Figure 3: PBE supported by the Intuitive Criterion

Case (2): Consider the signaling game depicted in Figure 3. Assume that $x = 4$. There is a PBE where both types of the sender pool on $R$. That is, $s^*(t_1) = s^*(t_2) = R$, $r^*(L) = u$ and $r^*(R) = d$ together with a family of posterior beliefs $\mu^*(t_1 \mid R) = 0.4$ and $\mu^*(t_1 \mid L) \leq 0.25$ constitute the pooling PBE. This pooling behavior cannot be supported by any HTE. There are only two possible candidates for hypotheses:

- $\pi_1 = \{\pi_1(t_1, R) = 0.4, \pi_1(t_2, R) = 0.6\}$ since $s(t_1) = R$ and $s(t_2) = R$ is the sender’s best-response against $r(L) = u$ and $r(R) = u$,

- $\pi_2 = \{\pi_2(t_1, L) = 0.4, \pi_2(t_2, L) = 0.6\}$ since $s(t_1) = L$ and $s(t_2) = L$ is the sender’s best-response against $r(L) = d$ and $r(R) = u$.

In this case, hypothesis $\pi_1$ supports the receiver’s behavior on-the-equilibrium path. However, when she observes $L$, there is no hypothesis that rationalizes playing $u$. After updating the only alternative hypothesis $\pi_2$ on $L$, the receiver prefers playing $d$. Therefore, the pooling PBE does not survive against the HT refinement.

The pooling PBE passes the Intuitive Criterion. However, since $I(L) = \Theta$, the Intuitive Criterion allows for arbitrary posterior beliefs on $\mu(\cdot \mid L)$ over $\Theta$.
Thus, Cases (a) and (b) demonstrate that the two refinement criteria are not nested.

Proof of Proposition 2. Let \((s^*, r^*, \mu^*)\) be a PBE that passes the Intuitive Criterion. Consider an out-of-equilibrium message \(m^o\) and let \(r^*(m^o) = a^*\) be a receiver’s action chosen in response to message \(m^o\). Since the set of sender’s types who have an incentive to deviate from the equilibrium strategy \(s^*\) by sending \(m^o\) contains only one type \(\theta^o\), i.e., \(I(m^o) = \{\theta^o\}\), the Intuitive Criterion implies that action \(a^*\) is a best response with respect to the posterior belief, \(\mu^*(\theta^o \mid m^o) = 1\).

We show that there exists a hypothesis \(\pi^o\) which rationalizes playing \(a^*\) given \(m^o\). From the Intuitive Criterion, we know that if type \(\theta^o\) sends \(m^o\), there exists an action \(a = r(m^o)\) such that
\[
 u^*_S(\theta^o) < u_S(\theta^o, m^o, a).
\]
Thus, we can construct a receiver’s strategy \(r : \mathcal{M} \rightarrow \mathcal{A}\) such that \(r(m) = r^*\) for any \(m \neq m^o\) and \(r(m^o) = a\), otherwise. Notice that the equilibrium payoff \(u^*_S(\theta^o)\) is the highest payoff that type \(\theta^o\) can achieve by sending any message \(m \neq m^o\) against \(r\), and message \(m^o\) yields a strictly higher payoff than \(u^*_S(\theta^o)\) when the receiver plays her strategy \(r\). Thus, it follows that message \(m^o\) is the sender’s best response against \(r\), i.e.,
\[
 m^o = \arg\max_{m \in \mathcal{M}} u_S(\theta^o, m, r(m)), \quad (12)
\]
Since \(I(m^o) = \{\theta^o\}\), sending \(m^o\) is the sender’s best response against \(r\) only for type \(\theta^o\), one can define a sender’s strategy \(s : \Theta \rightarrow \mathcal{M}\) such that \(s(\theta^o) = m^o\) and \(s(\theta) = s^*\) for any \(\theta \neq \theta^o\). From the given PBE and by Equation 12, strategy \(s\) is a sender’s best response against \(r\). Thus, a prior \(\pi^o\) over \(\Theta \times \mathcal{M}\) defined by
\[
 \pi^o(\theta, m) = \begin{cases} p(\theta), & \text{if } s(\theta) = m \\ 0, & \text{otherwise.} \end{cases}
\]
for every \((\theta, m) \in \Theta \times \mathcal{M}\), constitutes a hypothesis. Since \(\mu(\theta^o \mid m^o) = \pi^o(\theta^o, m^o) = 1\), the posterior belief \(\mu(\theta^o \mid m^o)\) rationalizes playing \(a^*\) given \(m^o\). Since \(m^o\) was chosen arbitrarily, we can construct a behavioral hypothesis for any out-of-equilibrium message. Let \(\{\pi^o_1, \ldots, \pi^o_K\}\) be a collection of such hypotheses supporting the receiver’s PBE behavior on out-of-equilibrium messages \(m^o_1, \ldots, m^o_K\). Also, by definition of PBE, there exists a hypothesis \(\pi\) supporting receiver’s PBE behavior on messages sent on-the-equilibrium path. Thus, \((s^*, r^*)\) and a suitable chosen second-order prior \(\rho\) with \(\text{supp}(\rho) = \{\pi, \pi^o_1, \ldots, \pi^o_K\}\)
such that
\[ \pi^* = \pi \text{ and } \pi^{**}_{mk} = \pi^0_k \text{ for all } k \in \{1, \ldots, K\}, \]

together with a family of posterior beliefs derived by HT updating rule constitutes an HTE supporting the given PBE behavior. Hence, the PBE passes the HT refinement yielding the same posterior beliefs as the Intuitive Criterion does.

Now, suppose that Condition (9) is satisfied. That is, for any out-of-equilibrium message \( m^0 \in \mathcal{M}^0 \) there is no type \( \theta \in T(m^0) \) for whom sending \( m^0 \) is a sender’s best response. This condition implies that there does not exist a behavioral hypothesis \( \pi \) such that \( \pi(\theta, m^0) = p(\theta) > 0 \) for any \( \theta \in T(m^0) \). Hence, in any HTE supporting the given PBE behavior, the posterior belief induced by the maximum likelihood hypothesis \( \pi^{**} \) given an out-of-equilibrium message \( m^0 \in \mathcal{M}^0 \) is \( \mu_\rho(\theta^0 \mid m^0) = 1 \). Therefore, the HT refinement outcome is always unique. ■

**Proof of Proposition 3.** Let \((s^*, r^*, \mu^*)\) be a PBE that passes the Intuitive Criterion. Since \( I(m^0_k) = \{\theta^0_k\} \) is a singleton for every out-of-equilibrium message \( m^0_k \) where \( k \in \{1, \ldots, K\} \), by Proposition 2, there exists an HTE supporting the PBE refined by the Intuitive Criterion. Let \((s^*, r^*, \mu^*, \rho)\) with \( \text{supp}(\rho) = \{\pi^*, \pi^0_1, \ldots, \pi^0_K\} \) where \( \pi^* \) is the most likely hypothesis and \( \pi^{**}_{mk} = \pi^0_k \) is the alternative hypothesis used for updating on \( m^0_k \), be such an HTE. By construction of \( \pi^0_k \) (see proof of Proposition 2), for every \((\theta, m) \in \Theta \times \mathcal{M} \), we have
\[ \pi^0_k(\theta, m) = \pi^*(\theta, m) \text{ if } \theta \in \Theta \setminus \{\theta^0_k\}, \text{ and } \pi^0_k(\theta^0_k, m^0_k) = p(\theta^0_m). \] (13)

(Notice that \( \pi^*(\theta^0_k, m^0_k) = 0 \) and \( \mu(\theta^0_k \mid m^0_k) = \frac{\pi^0_k(\theta^0_k, m^0_k)}{\pi^0_k(\Theta, m^0)} = 1 \)). Fix an out-of-equilibrium message \( m^0_k \). Since \( \theta^0_k \) is dummy for an on-the-equilibrium message \( m \) at \((s^*, r^*, \mu^*)\) (i.e., a message with \( \pi^*(\Theta, m) > 0 \)), it is true that for every \( a \in \mathcal{A} \):
\[ \text{if } a \in \arg \max_{a \in \Theta} \sum_{\theta^0 \in \Theta} \frac{\pi^*(\theta^0, m)}{\pi^*(\Theta, m)} u_R(\theta^0, m, a), \text{ then } a \in \arg \max_{a \in \Theta} \sum_{\theta^0 \in \Theta} \frac{\pi(\theta, m)}{\pi(\Theta, m)} u_R(\theta, m, a), \] (14)

where \( \pi \) is a probability distribution on \( \Theta \times \mathcal{M} \) such that
\[ \pi(\theta, m) = \begin{cases} \pi^*(\theta, m), & \text{if } \theta \in \Theta \setminus \{\theta^0_k\}, \\ 0, & \text{if } \theta = \theta^0_k. \end{cases} \] (15)

Notice that we have \( \pi^0_k(\theta^0_k, m^0_k) = 0 \) and \( \pi^0_k(\theta^0_k, m^0_k) = \bar{\pi}(\theta, m) \) for any \( \theta \in \Theta \setminus \{\theta^0_k\} \). Thus,
the alternative hypothesis $\pi_k$ has to satisfy Condition (14), implying that it is behaviorally consistent with respect to $\pi^*$. Since $m_k^*$ was chosen arbitrarily, any alternative hypothesis in $\{\pi_1^*, \ldots, \pi_K^*\}$ is behaviorally consistent. Therefore, $(s^*, r^*, \mu^*, \rho)$ with $\text{supp}(\rho) = \{\pi^*, \pi_1^*, \ldots, \pi_K^*\}$ constitutes a BCHTE supporting the PBE together with the unique refinement outcome by the Intuitive Criterion.

\begin{proof}[Proof of Proposition 4] Let $(s^*, r^*, \mu^*)$ be a PBE. Consider an out-of-equilibrium message $m^\circ$. Fix a posterior belief $\mu^*(\cdot \mid m^\circ)$ of the given PBE. Without loss of generality, we assume that there are $N$ number of types in $\text{supp}(\mu^*(\cdot \mid m^\circ))$; that is, $\text{supp}(\mu^*(\cdot \mid m^\circ)) = \{\theta_1, \theta_2, \ldots, \theta_N\}$. There exist pure strategies of the sender $s_1, s_2, \ldots, s_N$ such that

\begin{align*}
s_i(\theta) &= \begin{cases} m^\circ, & \text{if } \theta = \theta_i, \\ m \neq m^\circ, & \text{if } \theta \neq \theta_i, \end{cases} \quad \text{for every } i \in \{1, 2, \ldots, N\}.
\end{align*}

From $s_1, s_2, \ldots, s_N$, we can construct $N$ weak hypotheses $\pi_1, \pi_2, \ldots, \pi_N$ as follows:

\begin{align*}
\pi_1 &= \{ \pi_1(\theta_1, m^\circ) = p(\theta_1), \ \pi(\Theta, m^\circ) = p(\theta_1) \}, \\
\pi_2 &= \{ \pi_2(\theta_2, m^\circ) = p(\theta_2), \ \pi(\Theta, m^\circ) = p(\theta_2) \}, \\
& \vdots \\
\pi_N &= \{ \pi_N(\theta_N, m^\circ) = p(\theta_N), \ \pi_N(\Theta, m^\circ) = p(\theta_N) \}.
\end{align*}

Next we define a mixed hypothesis $\tilde{\pi}$. That is, let $\tilde{\pi}$ be a mixture among the weak hypotheses $\pi_1, \pi_2, \ldots, \pi_N$ with $\alpha_i \in [0, 1]$ for any $i = 1, 2, \ldots, N$ and $\sum_{i=1}^N \alpha_i = 1$ such that

$$
\tilde{\pi}_{(\alpha)}(\theta, m) = \alpha_1 \pi_1(\theta, m) + \alpha_2 \pi_2(\theta, m) + \ldots + \alpha_N \pi_N(\theta, m) \quad \text{for every } (\theta, m) \in \Theta \times M.
$$

By updating $\tilde{\pi}$ via Bayes’ rule, we get a vector of posterior beliefs $\mu_{(\alpha)}(\cdot \mid m^\circ)$:

$$
\mu_{(\alpha)}(\cdot \mid m^\circ) = \begin{pmatrix}
\frac{\alpha_1 p(\theta_1)}{\alpha_1 p(\theta_1) + \alpha_2 p(\theta_2) + \ldots + \alpha_N p(\theta_N)} \\
\frac{\alpha_2 p(\theta_2)}{\alpha_1 p(\theta_1) + \alpha_2 p(\theta_2) + \ldots + \alpha_N p(\theta_N)} \\
\vdots \\
\frac{\alpha_N p(\theta_N)}{\alpha_1 p(\theta_1) + \alpha_2 p(\theta_2) + \ldots + \alpha_N p(\theta_N)}
\end{pmatrix}.
$$

\end{proof}
By defining $\alpha$ such that

$$
\alpha = \begin{pmatrix}
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_N
\end{pmatrix} = \begin{pmatrix}
\frac{\mu^*(\theta_1 | m^o)}{p(\theta_1)} K \\
\frac{\mu^*(\theta_2 | m^o)}{p(\theta_2)} K \\
\vdots \\
\frac{\mu^*(\theta_N | m^o)}{p(\theta_N)} K
\end{pmatrix},
$$

and

$$
K = \frac{1}{\frac{\mu^*(\theta_1 | m^o)}{p(\theta_1)} + \frac{\mu^*(\theta_2 | m^o)}{p(\theta_2)} + \ldots + \frac{\mu^*(\theta_N | m^o)}{p(\theta_N)}},
$$

we obtain a mixture among weak hypotheses that induces the same posterior beliefs as $\mu^*(\cdot | m^o)$. Since $\mu^*(\cdot | m^o)$, $p(\cdot)$ and $K$ are strictly positive, so are $\alpha_1, \alpha_2, \ldots, \alpha_N$ strictly positive. Also, we have $\sum_{i=1}^N \alpha_i = 1$. Hence, $\tilde{\pi}(\alpha)$ is a well-defined mixed hypothesis.

Since $m^o$ was chosen arbitrarily, we can find a mixed hypothesis supporting receiver’s behavior at any out-of-equilibrium message. Therefore, WHTE with an alternative defined as a mixture of (weak) hypotheses exists for any PBE.

**Proof of Proposition 5.** Consider a PBE $(s^*, r^*, \mu^*)$. Assume that there exists a HTE supporting $(s^*, r^*, \mu^*)$. Hence, there exist the most likely hypothesis $\pi^*$ and alternative hypotheses $\{\pi^{**}_m\}_{m^o \in \mathcal{M}^o}$ where $\mathcal{M}^o$ is the set of out-of-equilibrium messages. Fix an out-of-equilibrium message $m^o$. We define a mixed hypothesis $\tilde{\pi}$ between $\pi^*$ and $\pi^{**}_m$. That is, for $\alpha \in [0, 1]$:

$$
\tilde{\pi}(\alpha)(\theta, m) = (1 - \alpha)\pi^*(\theta, m) + \alpha \pi^{**}_m(\theta, m) \quad \text{for every } (\theta, m) \in \Theta \times \mathcal{M}.
$$

Since the receiver’s best response correspondence with respect to $\pi^*$ is a singleton on-the-equilibrium path, by continuity, there exists small enough $\varepsilon_{m^o} > 0$ such that the mixed hypothesis:

$$
\tilde{\pi}(\varepsilon_{m^o})(\theta, m) = (1 - \varepsilon_{m^o})\pi^*(\theta, m) + \varepsilon_{m^o} \pi^{**}_m(\theta, m) \quad \text{for every } (\theta, m) \in \Theta \times \mathcal{M},
$$

maintains, after updating, the receiver’s behavior on-the-equilibrium path. Therefore, $\tilde{\pi}(\varepsilon_{m^o})$ is behaviorally consistent. Moreover, by updating $\tilde{\pi}(\varepsilon_{m^o})$ on $m^o$, we get

$$
\mu(\theta | m^o) = \frac{\varepsilon_{m^o} \pi^{**}(\theta, m^o)}{\varepsilon_{m^o} \pi^{**}(\Theta, m^o)} = \frac{\pi^{**}(\theta, m^o)}{\pi^{**}(\Theta, m^o)}, \quad \text{for every } \theta \in \Theta.
$$
since \( \pi^*(\Theta, m^\circ) = 0 \) and \( \pi^{**}(\Theta, m^\circ) > 0 \). That is, \( \bar{\pi}(m^\circ) \) induces the same posterior belief as \( \pi^{**}(m^\circ) \), thus maintaining the out-of-equilibrium behavior on \( m^\circ \). Note that the above argument is true for any mixed hypothesis \( \bar{\pi}(\alpha) \) with \( \alpha \in (0, \varepsilon_m^\circ] \). Since \( m^\circ \) was chosen arbitrary, there exists such \( \varepsilon_m^\circ \) for any out-of-equilibrium message \( m^\circ \in M^\circ \). Among all such \( \{\varepsilon_m^\circ\}_{m^\circ \in M^\circ} \), take a minimum, denoted by \( \varepsilon^* \) (i.e., \( \varepsilon^* = \min\{\varepsilon_m^\circ\}_{m^\circ \in M^\circ} \)). Then, for every \( m^\circ \in M^\circ \), a mixed hypothesis \( \bar{\pi}(\varepsilon^*) \) is behaviorally consistent. Thus, we constructed a BCHTE with \( \pi^* \) and \( \{\pi_m^{**} = \bar{\pi}(\varepsilon^*)\}_{m^\circ \in M^\circ} \) supporting the given HTE.

\[ \square \]

### B Intuitive Criterion and Arbibray Beliefs.

Consider the signaling game in Figure 3 with \( x = 2 \). There is a pooling PBE where \( s^*(t_1) = s^*(t_2) = R \), \( r^*(L) = u \), \( r^*(R) = d \), \( \mu^*(t_1 | R) = 0.4 \) and \( \mu^*(t_1 | L) \leq 0.5 \). Since \( T(L) = \{\emptyset\} \) and \( I(L) = \Theta \), the Intuitive Criterion allows for any \( \mu^*(t_1 | L) \leq 0.5 \). However, the posteriors beliefs can be refined by the unique HTE with \( supp(\rho) = \{\pi_1, \pi_2\} \) and \( \rho(\pi_1) > \rho(\pi_2) \) where

- \( \pi_1 := \{\pi_1(t_1, R) = 0.4, \pi_2(t_2, R) = 0.6\} \) since \( s(t_1) = s(t_2) = R \) is the sender’s best-response against \( r(L) = u \) and \( r(R) = d \),

- \( \pi_2 := \{\pi_2(t_1, L) = 0.4, \pi_2(t_2, L) = 0.6\} \) since \( s(t_1) = s(t_2) = L \) is the sender’s best-response against \( r(L) = d \) and \( r(R) = u \).

The alternative hypothesis \( \pi_2 \) induces the unique posterior belief \( \mu^*(t_1 | L) = 0.4 \).

### C Non-Existence of HTE and WHTE.

Consider the signaling game in Figure 4. There exist a (unique) pooling PBE for this game where

\[ s^*(t_1) = s^*(t_2) = L, \quad r^*(L) = d, \quad r^*(R) = m, \quad \text{for any} \quad \mu^*(t_1 | R) \in [0.475, 0.525]. \]

**Case (a):** There are three possible candidates for (strong) hypotheses:

- \( \pi_1 := \{\pi_1(t_1, L) = 0.9, \pi_1(t_2, L) = 0.1\} \) since \( s(t_1) = L \) and \( s(t_2) = L \) is the sender’s best-response strategy against \( r(L) = u \) and \( r(R) = u \),
Figure 4: Signaling Game without HTE and WHTE

- \( \pi_2 := \{\pi_2(t_1, L) = 0.9, \pi_2(t_2, R) = 0.1\} \) since \( s(t_1) = L \) and \( s(t_2) = R \) is the sender’s best-response against \( r(L) = d \) and \( r(R) = u \).

- \( \pi_3 := \{\pi_3(t_1, R) = 0.9, \pi_3(t_2, L) = 0.1\} \) since \( s(t_1) = R \) and \( s(t_2) = L \) is the sender’s best-response against \( r(L) = d \) and \( r(R) = d \).

While \( \pi_1 \) rationalizes the receiver’s on-the-equilibrium behavior, none of the alternative hypothesis \( \pi_2 \) and \( \pi_3 \) can rationalize her behavior when the out-of-equilibrium message \( R \) is observed. Thus, there is no HTE supporting the PBE.

**Case (b):** Besides the (strong) hypotheses \( \pi_1, \pi_2 \) and \( \pi_3 \), there is a weak hypothesis: \( \pi_4 := \{\pi_4(t_1, R) = 0.9, \pi_4(t_2, R) = 0.1\} \). However, neither \( \pi_4 \) can rationalize, after updating, the receiver’s out-of-equilibrium behavior. Thus, there does not exist a WHTE supporting the PBE.

**Case (c):** Here, we show that an HTE exists when we allow for mixing between (strong) hypothesis presented in Case (a). Consider a mixed hypothesis \( \tilde{\pi} \) between \( \pi_2 \) and \( \pi_3 \). That is, for \( \alpha \in [0, 1] \),

\[
\tilde{\pi}(\theta, m) = \alpha \pi_2(\theta, m) + (1 - \alpha) \pi_3(\theta, m) \quad \text{for every } (\theta, m) \in \Theta \times \mathcal{M}.
\]
By updating $\tilde{\pi}$ via Baye’s rule, we get a posterior belief given message $R$:

$$
\mu_\alpha(t_1 \mid R) = \frac{\tilde{\pi}(t_1, R)}{\tilde{\pi}(\Theta, R)} = \frac{(1 - \alpha)p(t_1)}{\alpha p(t_2) + (1 - \alpha)p(t_1)} = \frac{0.9(1 - \alpha)}{0.1\alpha + 0.9(1 - \alpha)}.
$$

Since $\mu_{(\alpha=0)}(t_1 \mid R) = 1$, $\mu_{(\alpha=1)}(t_1 \mid R) = 0$, and $0.1\alpha + 0.9(1 - \alpha)$ is continuous with respect to $\alpha$, we can always find $\alpha \in [0, 1]$ such that $\mu_\alpha(t_1 \mid R) = \mu^*(t_1 \mid R)$ where $\mu^*_\alpha(t_1 \mid R) \in [0.475, 0.525]$. Hence, for any posterior belief of the pooling PBE, we can construct an HTE in which the alternative hypothesis is a mixed hypothesis $\tilde{\pi}$ justifying the posterior belief, thus leaving the PBE unrefined.

References


