Supplemental Information for "I Don't Know"

Matthew Backus* Andrew T. Little†

March 28, 2020

Contents

*Columbia University, NBER, and CEPR, matthew.backus@columbia.edu
†Corresponding author. UC Berkeley, andrew.little@berkeley.edu

1 Analysis of Other Cases

The main text(and associated proofs) contains a full analysis of the cases with no policy concerns and no validation/state validation, and the case with small policy concerns and difficulty validation. Here we first tie up the cases of difficulty and full validation with no policy concerns, and then the other validation cases with policy concerns.

1.1 Difficulty and Full Validation with No Policy Concerns

Difficulty validation A complete description of the MSE with difficulty validation proves challenging. However, with no policy concerns, we can show that there is never any information communicated about the state, though there can be information communicated about the difficulty of the problem:

Proposition S.1. *With no policy concerns and difficulty validation,*
i. in any MSE, $a^*(m) = p_1$ *for all on-path* m*, and*
ii. there is an MSE where the good uninformed types always admit uncertainty.

Proof. Given the payoff equivalence classes, the good and informed types must use the same mixed strategy. In any MSE, the posterior belief about the state upon observing an on-path message m can be written as a weighted average of the belief about the state conditional on being in each equivalence class, weighted by the probability of being in the

class:

$$\begin{aligned}Pr(\omega = 1|m) &= Pr(\omega = 1|m, \theta = g, s \in \{s_0, s_1\})Pr(\theta = g, s \in \{s_0, s_1\}|m) \\ &\quad + Pr(\omega = 1|m, \theta = g, s = s_\emptyset)Pr(\theta = g, s = s_\emptyset|m) \\ &\quad + Pr(\omega = 1|m, \theta = b)Pr(\theta = b|m) \\ &= p_1 Pr(\theta = g, s \in \{s_0, s_1\}|m) + p_1 Pr(\theta = g, s = s_\emptyset|m) + p_1 Pr(\theta = b|m) = p_1.\end{aligned}$$

For each equivalence class there is no information conveyed about the state, so these conditional probabilities are all p_1, and hence sum to this as well.

For part ii, we construct an equilibrium where the informed types always send m_e ("the problem is easy"), the good but uninformed types send m_h ("the problem is hard"), and the bad types mix over these two messages with probability $(\sigma_b(m_e), \sigma_b(m_h))$. Since m_h is never sent by the informed types, sending this message admits uncertainty.

There can be an equilibrium where both of these messages are sent by the bad types if and only if they give the same expected payoff. Writing the probability of sending m_e as $\sigma_b(m_e)$, this is possible if:

$$p_e \pi_g(m_e, e) + (1 - p_e)\pi_g(m_e, h) = p_e \pi_g(m_h, e)(1 - p_e)\pi_g(m_h, h),$$

– or, rearranged:

$$p_e \frac{p_g p_e}{p_g p_e + (1 - p_g)\sigma_b(m_e)} = (1 - p_e)\frac{p_g(1 - p_e)}{p_g(1 - p_e) + (1 - p_g)(1 - \sigma_b(m_e))}. \tag{1}$$

The left-hand side of this equation (i.e., the payoff to guessing the problem is easy) is decreasing in $\sigma_b(m_e)$, ranging from p_e to $p_e \frac{p_g p_e}{p_g p_e + (1 - p_g)}$. The right-hand side is increasing

in $\sigma_b(m_e)$, ranging from $(1-p_e)\frac{p_g(1-p_e)}{p_g(1-p_e)+(1-p_g)}$ to $1-p_e$. So, if

$$p_e \frac{p_g p_e}{p_g p_e + (1-p_g)} - (1-p_e) \geq 0, \tag{2}$$

then payoff to sending m_e is always higher. After multiplying through by $p_g p_e + (1-p_g)$, the left-hand side of (2) is quadratic in p_e (with a positive p_e term), and has a root at $\frac{2p_g - 1 + \sqrt{1+4p_g-4p_g^2}}{4p_g}$ which is always on $(1/2, 1)$, and a negative root.[1] So, when p_e is above this root, the payoff to sending m_e is always higher, and hence there is a MSE where the uninformed types always send this message.

On the other hand, if

$$(1-p_e)\frac{p_g(1-p_e)}{p_g(1-p_e)+(1-p_g)} - p_e \geq 0,$$

then the payoff for sending m_h is always higher, which by a similar argument holds if $p_e \leq \frac{2p_g+1-\sqrt{1+4p_g-4p_g^2}}{4p_g}$. However, if neither of these inequalities hold, then there is a $\sigma_b(m_e) \in (0,1)$ which solves (1), and hence there is an MSE where m_e is sent with this probability and m_h with complementary probability. Summarizing, there is an MSE where the bad type sends message m_e with probability:

$$\sigma_b^*(m_e) = \begin{cases} 0 & p_e \leq \frac{2p_g+1-\sqrt{1+4p_g-4p_g^2}}{4p_g} \\ \frac{p_e(p_e-p_g+2p_e p_g-2p_e^2 p_g)}{(1-p_g)(1-2p_e(1-p_e))} & p_e \in \left(\frac{2p_g+1-\sqrt{1+4p_g-4p_g^2}}{4p_g}, \frac{2p_g-1+\sqrt{1+4p_g-4p_g^2}}{4p_g}\right) \\ 1 & p_e \geq \frac{2p_g-1+\sqrt{1+4p_g-4p_g^2}}{4p_g} \end{cases}$$

and message m_h with probability $\sigma_b^*(m_h) = 1 - \sigma_b^*(m_h)$. □

This implies that, while we can learn something about the question difficulty with difficulty

[1] All of these observations follow from the fact that $1 + 4p_g - 4p_g^2 \in (1, (2p_g+1)^2)$.

validation alone, learning about the state (and attaining an honest equilibrium) require either nonzero policy concerns or state validation (or both).

Full Validation with No Policy Concerns With full validation, there are four possible validation results for each message. The expected payoff to sending message m given one's type and message is:

$$\sum_{\delta\in\{e,h\}}\sum_{\omega\in\{0,1\}} Pr(\delta|s,\theta)Pr(\omega|s,\theta)\pi_g(m,\omega,\delta).$$

No pair of types share the same $Pr(\omega|s,\theta)$ and $Pr(\delta|s,\theta)$, so none must be payoff equivalent. As a result, all types can use distinct strategies, and off-path beliefs are unrestricted.

In an honest equilibrium, upon observing $(m_0,0,e)$ or $(m_1,1,e)$, the DM knows that the expert is competent. Upon observing $(m_\emptyset,\omega,e)$ the DM knows that the expert is not competent, as a competent expert would have received and sent an informative message since the problem is easy. Upon observing $(m_\emptyset,\omega,h)$, the DM belief about the expert competence is the same as the prior, since if the problem is hard no one gets an informative message (and all send $m_\emptyset$).[2] So, the competence evaluations for the on-path messages are:

$$\pi_g(m_0,0,e)=1 \qquad \pi_g(m_1,1,e)=1$$
$$\pi_g(m_\emptyset,\omega,e)=0 \qquad \pi_g(m_\emptyset,\omega,h)=p_g$$

To make honesty as easy as possible to sustain, suppose that for any off-path message ("guessing wrong"), the competence evaluation is zero.

The informed types get a competence evaluation of 1 for sending their honest message, so face no incentive to deviate.

[2]Formally, applying Bayes' rule gives $Pr(\theta=g|m_\emptyset,\omega,h)=\frac{p_g(1-p_e)}{p_g(1-p_e)+(1-p_g)(1-p_e)}=p_g$.

A good but uninformed type knows the difficulty validation will reveal $\delta = h$, but does not know ω. Sending the honest message $m_\emptyset$ gives a competence payoff of p_g. However, sending either m_0 or m_1 will lead to an off-path message/validation combination, and hence a payoff of zero. So, these types face no incentive to deviate.

Finally, consider the bad uninformed types, who do not know what either the state or difficulty validation will reveal. If they send $m_\emptyset$, they will be caught as uninformed if the problem was in fact easy (probability p_e). However, if the problem is hard, the DM does not update about their competence for either state validation result. So, the expected payoff to sending $m_\emptyset$ is $(1 - p_e)p_g$.

If guessing m_1, the expert will be "caught" if either the problem is hard *or* the state is 0. However, if guessing correctly, the competence evaluation will be 1. So, the expected payoff to this deviation is $p_e p_1$. Similarly, the expected payoff to guessing m_0 is $p_e(1 - p_1) < p_e p_1$, so m_1 is the best deviation.

Honesty is possible if admitting uncertainty leads to a higher competence evaluation than guessing m_1, or:

$$(1 - p_e)p_g \geq p_e p_1 \implies p_e \leq \frac{p_g}{p_g + p_1}.$$

If this inequality does not hold, a fully honest MSE is not possible. However, there is always an MSE where the good but uninformed types always send $m_\emptyset$. In such an equilibrium, the bad types pick a mixed strategy over m_0, m_1, and $m_\emptyset$. Whenever the DM observes an "incorrect guess" she assigns a competence evaluation of zero. So, the informed types never deviate (as this ensures an incorrect guess), and good uninformed types have no reason to guess since they know the problem is hard. Returning to the derivation of the honest equilibrium, the off-path beliefs in this MSE are justified, in the sense that the good

types all have strict incentives to report their honest message, and the bad types are the only ones who potentially face an incentive to send m_0 or m_1 when the problem is hard or $m_\emptyset$ when the problem is easy.

Summarizing:

Proposition S.2. *With no policy concerns and full validation, there is an MSE where the informed types send distinct messages and the good but uninformed types always admit uncertainty. If $p_e \leq \frac{p_g}{p_g+p_1}$, there is an honest MSE.*

Proof. The condition for the honest equilibrium is derived above. So what remains is to show there is always an MSE where the good but uninformed type always sends $m_\emptyset$.

In such an equilibrium, message/validation combinations $(m_0, 0, e)$, $(m_1, 1, e)$ and $(m_\emptyset, 0, h)$ and $(m_\emptyset, 1, h)$ are the only ones observed when the expert is competent. So, any other message/validation combination is either on-path and only sent by the bad types, in which case the competence assessment must be 0, or is off-path and can be set to 0.

The informed type observing s_0 knows the validation will be $(0, e)$, and $(m, 0, e)$ leads to competence assessment zero for $m \neq m_0$. So, this type has no incentive to deviate, nor does the s_1 type by an analogous argument. The good but uninformed type knows the validation will reveal h, and the DM observing (m_i, ω, h) for $i \in \{0, 1\}$ and $\omega \in \{0, 1\}$ will lead to a competence assessment of zero. So this type faces no incentive to deviate.

Now consider the bad type strategy. While explicitly deriving the equilibrium strategies here is tedious, a simple fixed point argument can be used to show existence. Write the whole strategy with $\sigma_b = (\sigma_b(m_0), \sigma_b(m_1), \sigma_b(m_\emptyset))$, and the bad type's expected competence assessment for sending each message when the DM expects strategy σ (averaging

over the validation result) as:

$$U_\theta(m_\emptyset, b, \sigma) \equiv p_e 0 + (1-p_e)\frac{p_g}{p_g + (1-p_g)\sigma_b(m_\emptyset)},$$
$$U_\theta(m_0, b, \sigma) \equiv p_e(1-p_1)\frac{p_g}{p_g + (1-p_g)\sigma_b(m_0)} + (1-p_e)0, \text{ and}$$
$$U_\theta(m_1, b, \sigma) \equiv p_e p_1 \frac{p_g}{p_g + (1-p_g)\sigma_b(m_1)} + (1-p_e)0.$$

Write the expected payoff to the bad expert choosing mixed strategy σ when the decision-maker expects mixed strategy $\hat{\sigma}_b$ as $U(\sigma, \hat{\sigma}) = \sum_{i \in \{0,1,\emptyset\}} \sigma_b(m_i) U_\theta(m_i; \hat{\sigma})$, which is continuous in all $\sigma_b(m_i)$, so optimizing this objective function over the (compact) unit simplex must have a solution. So, $BR(\hat{\sigma}_b) = \arg\max_\sigma U(\sigma; \hat{\sigma})$ is a continuous mapping from the unit simplex to itself, which by the Kakutani fixed point theorem must have a solution. So, the strategy (or strategies) given by such a fixed point are a best response for the bad type when the decision-maker forms correct beliefs given this strategy.

□

1.2 Nonzero Policy Concerns

Now we the case where the expert cares about the policy made, $\gamma > 0$. Not surprisingly, when policy concerns are "large", there is always an honest MSE since the expert primarily wants the DM to take the best possible action. Here we analyze how high policy concerns have to be in order to attain this honest equilibrium, and provide some results about what happens when policy concerns are not small but not large enough to induce honesty.

In Appendix 2, we show that with no validation and state validation, in any MSE which is not babbling, the types observing s_0 and s_1 cannot send any common messages. Combined with a relabeling argument, for all of the analysis (of these validation regimes) with policy

concerns we can again restrict attention to MSE where the informed types always send m_0 and m_1, respectively, and uninformed types send at most one other message $m_\emptyset$.

While we do not rely on this result for the difficulty and full validation cases (where we only focus on the existence of equilibria with certain properties), we analyze the analogous messaging strategies in these cases to facilitate comparison.

No Validation Informed types never face an incentive to deviate from the honest equilibrium: upon observing s_x for $x \in \{0, 1\}$, the DM chooses policy $a^*(s_x) = x$, and knows the expert is competent, giving the highest possible expert payoff.

Uninformed types, however, may wish to deviate. Upon observing $m_\emptyset$, the DM takes action $a = \pi_1 = p_1$, which gives expected policy value $1 - p_1(1 - p_1)$, and the belief about the competence is $\pi_g^\emptyset$. So, for the uninformed experts of either competence type, the payoff for reporting honestly and sending signal $m_\emptyset$ is:

$$\pi_g^\emptyset + \gamma(1 - p_1(1 - p_1)). \tag{3}$$

If the expert deviates to $m \in \{m_0, m_1\}$, his payoff changes in two ways: he looks competent with probability 1 (as only competent analysts send these messages in an honest equilibrium, and without validation this is always on path), and the policy payoff gets worse on average. So, the payoff to choosing m_1 is:

$$1 + \gamma p_1. \tag{4}$$

It i easy to check that the payoff to deviating to m_0 is weakly lower, and so m_1 is the

binding deviation to check. Preventing the uninformed type from guessing m_1 requires

$$\pi_g^{\emptyset} + \gamma(1 - p_1(1 - p_1)) \geq 1 + \gamma p_1.$$

Rearranging, define the threshold degree of policy concerns γ_{NV}^H required to sustain honesty by

$$\begin{aligned} \gamma &\geq \frac{1 - \pi_g^{\emptyset}}{(1 - p_1)^2} \\ &= \frac{(1 - p_g)}{(1 - p_g p_e)(1 - p_1)^2} \\ &\equiv \gamma_{NV}^H. \end{aligned} \tag{5}$$

If $\gamma < \gamma_{NV}^H$, the uninformed types strictly prefer sending m_1 to $m_{\emptyset}$ *if the DM expects honesty.* Given our concern with admission of uncertainty, it is possible that there is a mixed strategy equilibrium where the uninformed types sometimes send $m_{\emptyset}$ and sometimes send m_0 or m_1. However, as the following result shows, when policy concerns are too small to induce full honesty, the payoff for sending m_1 is always higher than the payoff for admitting uncertainty. Moreover, since γ_{NV}^H is strictly greater than zero, when policy concerns are sufficiently small some form of validation is required to elicit any admission of uncertainty.

Proposition S.3. *When $\gamma > 0$ and no validation:*
i. If $\gamma \geq \gamma_{NV}^H$, then there is an honest MSE,
ii. If $\gamma \in (0, \gamma_{NV}^H)$, then all non-babbling MSE are always guessing (i.e., $\sigma_{\emptyset}^*(m_{\emptyset}) = 0$)

Proof. Part i is shown above.

For part ii, it is sufficient to show that if $\gamma < \gamma_{NV}^H$, then in any proposed equilibrium where $\sigma_{\emptyset}(m_{\emptyset}) > 0$, the payoff for an expert to send m_1 is always strictly higher than the payoff to

sending $m_\emptyset$.

The competence evaluation upon observing m_1 as a function of the uninformed expert mixed strategy is:

$$\pi_g(m_1;\sigma_\emptyset(m_1)) = \frac{Pr(\theta = g, m_1)}{Pr(m_1)} = \frac{p_g p_1 p_e + p_g(1-p_e)\sigma_\emptyset(m_1)}{p_g p_1 p_e + (p_g(1-p_e) + (1-p_g))\sigma_\emptyset(m_1)}$$

– and the belief about the state is:

$$\pi_1(m_1;\sigma_\emptyset(m_1)) = \frac{Pr(\omega = 1, m_1)}{Pr(m_1)} = \frac{p_1(p_g p_e + (1-p_g p_e)\sigma_\emptyset(m_1))}{p_1 p_g p_e + (p_g(1-p_e) + (1-p_g))\sigma_\emptyset(m_1)}.$$

When observing $m_\emptyset$, the DM knows with certainty that the expert is uninformed, so $\pi_g(m_\emptyset) = \pi_g^\emptyset$ and $\pi_1(m_\emptyset) = p_1$.

Combining, the expected payoff for an uninformed type to send each message is:

$$\begin{aligned} U(m_1; s_\emptyset, \sigma_\emptyset(m_1)) = \pi_g(m_1;\sigma_\emptyset(m_1)) \\ + \gamma(1 - [p_1(1-\pi_1(m_1;\sigma_\emptyset(m_1)))^2 + (1-p_1)\pi_1(m_1;\sigma_\emptyset(m_1))^2]) \end{aligned}$$

and

$$U(m_\emptyset) = \pi_g^\emptyset + \gamma(1 - p_1(1-p_1)).$$

Conveniently, $U(m_\emptyset)$ is not a function of the mixed strategy.

If $\gamma = 0$, then $U(m_i;\sigma_i) > U(m_\emptyset)$ for both $i \in \{0,1\}$, because $\pi_g(m_i;\sigma_i) > \pi_g^\emptyset$. Further, by the continuity of the utility functions in γ and $\sigma_\emptyset(m_1)$, there exists a $\gamma^* > 0$ such that message m_1 will give a strictly higher payoff than $m_\emptyset$ for an open interval $(0, \gamma^*)$. The final

step of the proof is to show that this γ^* is exactly γ^H_{NV}.

To show this, let $\sigma^{\text{cand}}(\gamma)$ be the candidate value of $\sigma_\emptyset(m_1)$ that solves $U(m_1; s_\emptyset, \sigma_\emptyset(m_1)) = U(m_\emptyset)$. Rearranging, and simplifying this equality gives:

$$\sigma^{\text{cand}}(\gamma) = -\frac{p_1 p_g p_e}{1 - p_g p_e} + \gamma \frac{p_1 p_g p_e (1 - p_1)^2}{1 - p_g}$$

which is linear in γ. When $\gamma = 0$, $\sigma^{\text{cand}}(\gamma)$ is negative, which re-demonstrates that with no policy concerns the payoff to sending m_1 is always higher than $m_\emptyset$. More generally, whenever $\sigma^{\text{cand}}(\gamma) < 0$, the payoff to sending m_1 is always higher than $m_\emptyset$ so there can be no admission of uncertainty. Rearranging this inequality gives:

$$\begin{aligned} &-\frac{p_1 p_g p_e}{1 - p_g p_e} + \gamma \frac{p_1 p_g p_e (1 - p_1)^2}{1 - p_g} < 0 \\ \Leftrightarrow\ &\gamma < \frac{1 - p_g}{(1 - p_g p_e)(1 - p_1)^2} = \gamma^H_{NV}, \end{aligned}$$

completing part ii.

Now that we have demonstrated any MSE is always guessing, we can prove proposition S.3. As $\gamma \to 0$, the condition for an equilibrium where the uninformed types send both m_0 and m_1 is that the competence assessments are the same. Writing these out gives:

$$\begin{aligned} \pi_g(m_0; \sigma_\emptyset) &= \pi_g(m_1; \sigma_\emptyset) \\ \frac{p_g(1 - p_1)p_e + p_g(1 - p_e)\sigma_\emptyset(m_0)}{p_g(1 - p_1)p_e + (p_g(1 - p_e) + (1 - p_g))\sigma_\emptyset(m_0)} &= \frac{p_g p_1 p_e + p_g(1 - p_e)\sigma_\emptyset(m_1)}{p_g p_1 p_e + (p_g(1 - p_e) + (1 - p_g))\sigma_\emptyset(m_1)} \end{aligned}$$

which, combined with the fact that $\sigma_\emptyset(m_1) = 1 - \sigma_\emptyset(m_0)$ (by part ii) is true if and only if $\sigma_\emptyset(m_0) = 1 - p_1$ and $\sigma_\emptyset(m_1) = p_1$. There is no equilibrium where $\sigma_\emptyset(m_0) = 0$; if so, $\pi_g(m_0; \sigma_\emptyset) = 1 > \pi_g(m_1; \sigma_\emptyset)$. Similarly, there is no equilibrium where $\sigma_\emptyset(m_1) = 0$. $\square$

An intuition for this result is as follows. As uninformed types send m_1 more often, this has two affects on the appeal of sending this message. First, there is a complementarity where sending m_1 more often makes the policy response to this message less extreme, which the uninformed types like. Second, there is a substitution effect where it makes those sending m_1 look less competent. While these effects go in the opposite direction, the substitution effect that makes sending m_1 less appealing when other uninformed types do so is only strong when policy concerns are weak, which is precisely when sending m_1 is generally preferable to $m_\emptyset$ regardless of the uninformed type strategy.

State Validation. Suppose there is an honest equilibrium with state validation.

As in the case with no policy concerns, upon observing message $(m_0, 0)$ or $(m_1, 1)$ the DM knows the expert is competent and takes an action equal to the message, and upon $(m_\emptyset, 0)$ or $(m_\emptyset, 1)$ takes action p_1 and knows the expert is uninformed, giving competence evaluation $\pi_g^\emptyset$. So, the payoff for an uninformed type to send the equilibrium message is:

$$\pi_g^\emptyset + \gamma(1 - p_1(1 - p_1)). \tag{6}$$

By an identical argument to that made with no policy concerns, upon observing an off-path message, the payoff equivalence of the good and bad uninformed types implies the belief about competence in an MSE must be greater than or equal to $\pi_g^\emptyset$. So, the payoff to deviating to m_1 must be at least

$$p_1 + (1 - p_1)\pi_g^\emptyset + \gamma p_1$$

–and the corresponding policy concerns threshold to prevent this deviation is:

$$\pi_g^{\emptyset} + \gamma(1 - p_1(1 - p_1)) \geq p_1 + (1 - p_1)\pi_g^{\emptyset} + \gamma p_1$$

– which reduces to

$$\begin{aligned} \gamma &\geq p_1 \gamma_{NV}^H \\ &\equiv \gamma_{SV}^H \end{aligned} \tag{7}$$

Adding state validation weakens the condition required for an honest equilibrium, particularly when p_1 is close to $1/2$. However, this threshold is always strictly positive, so for small policy concerns there can be no honesty even with state validation.

As shown in the proof of the following, if this condition is not met, then as with the no validation case there can be no admission of uncertainty. Further, since adding policy concerns does not change the classes of payoff equivalence, the case as $\gamma \to 0$ is the same as $\gamma = 0$.

Proposition S.4. *With policy concerns and state validation:*
i. If $\gamma \geq \gamma_{SV}^H = p_1 \gamma_{NV}^H$, then there is an honest MSE,
ii. If $\gamma \in (0, \gamma_{SV}^H)$, then all non-babbling MSE are always guessing (i.e., $\sigma_{\emptyset}^*(m_{\emptyset}) = 0$).

Proof. Part i is demonstrated above

For part ii, our strategy mirrors the proof with no validation – that is, by way of contradiction, if the constraint for honesty is not met, then the payoff to sending m_1 is always strictly higher than $m_{\emptyset}$. As above, in any MSE where $\sigma_{\emptyset}(m_1) > 0$, the payoff for sending $m_{\emptyset}$ is

$\pi_g^{\emptyset} + \gamma(1 - p_1(1 - p_1))$. The payoff to sending m_1 is:

$$p_1\pi_g(m_1, 1) + (1 - p_1)\pi_g^{\emptyset} + \gamma(1 - p_1(1 - \pi_1(m_1, \sigma_{\emptyset}(m_1)))^2 + (1 - p_1)\pi_1(m_1, \sigma_{\emptyset}(m_1))^2).$$

Next, the posterior beliefs of the decision-maker are the same as in the no validation case except:

$$\pi_g(m_1, 1) = \frac{Pr(\theta = g, m_1, \omega = 1)}{Pr(m_1, \omega = 1)} = \frac{p_1 p_g p_e + p_1 p_g (1 - p_e)\sigma_{\emptyset}(m_1)}{p_1 p_g p_e + p_1(1 - p_g p_e)\sigma_{\emptyset}(m_1)} = \frac{p_g p_e + p_g(1 - p_e)\sigma_{\emptyset}(m_1)}{p_g p_e + (1 - p_g p_e)\sigma_{\emptyset}(m_1)}.$$

The difference between the payoffs for sending m_1 and $m_{\emptyset}$ can be written:

$$p_e p_g p_1 \frac{z(\sigma_{\emptyset}(m_1); \gamma)}{(1 - p_e p_g)(p_e p_g(1 - \sigma_{\emptyset}(m_1)) - \sigma_{\emptyset}(m_1))(p_e p_g(p_1 - \sigma_{\emptyset}(m_1)) + \sigma_{\emptyset}(m_1))^2}$$

– where

$$\begin{aligned} z(\sigma_{\emptyset}(m_1); \gamma) = \gamma p_e p_g(-1 + p_e p_g)(-1 + p_1)^2 p_1(p_e p_g(-1 + \sigma_{\emptyset}(m_1)) - \sigma_{\emptyset}(m_1)) \\ + (-1 + p_g)(p_e p_g(p_1 - \sigma_{\emptyset}(m_1)) + \sigma_{\emptyset}(m_1))^2). \end{aligned}$$

So any equilibrium where both m_1 and $m_{\emptyset}$ are sent is characterized by $z(\sigma_{\emptyset}(m_1); \gamma) = 0$. It is then sufficient to show that for $\gamma < \gamma_{SV}^H$, there is no $\sigma_{\emptyset}(m_1) \in [0, 1]$ such that $z(\sigma_{\emptyset}(m_1); \gamma) = 0$.

Formally, it is easy to check that z is strictly decreasing in γ and that $z(0, \gamma_{SV}^H) = 0$. So, $z(0, \gamma) > 0$ for $\gamma < \gamma_{SV}^H$. To show z is strictly positive for $\sigma_{\emptyset}(m_1) > 0$, first observe that:

$$\left.\frac{\partial z}{\partial \sigma_{\emptyset}(m_1)}\right|_{\gamma = \gamma_{SV}^H} = (1 - p_g)(1 - p_e p_g)(p_e p_g(2 - p_1)p_1 + (2 - 2p_e p_g)\sigma_{\emptyset}(m_1)) > 0$$

– and

$$\frac{\partial^2 z}{\partial \sigma_\emptyset(m_1)\partial\gamma} = -p_e p_g (1 - p_e p_g)^2 (1 - p_1)^2 p_1 < 0.$$

Combined, these inequalities imply $\frac{\partial z}{\partial \sigma_\emptyset(m_1)} > 0$ when $\gamma < \gamma^H_{SV}$. So, $z(\sigma_\emptyset(m_1), \gamma) > 0$ for any $\sigma_\emptyset(m_1)$ when $\gamma < \gamma^H_{SV}$, completing part ii. □

Difficulty Validation. As shown in the main text, the condition for an honest equilibrium with difficulty validation and policy concerns is.

$$\begin{aligned}(1 - p_e)p_g + \gamma(1 - p_1(1 - p_1)) &\geq p_e + \gamma p_1 \\ \gamma &\geq \frac{p_e(1 + p_g) - p_g}{(1 - p_1)^2} \equiv \gamma^H_{DV}.\end{aligned}$$

As discussed in the main text, γ^H_{DV} can be negative, meaning that there is an honest equilibrium even with no policy concerns.

Proposition S.5. *With policy concerns and difficulty validation:*
i. If $\gamma \geq \gamma^H_{DV}$, then there is an honest MSE.
ii. If $\gamma \leq \gamma^H_{DV}$, then there is an MSE where the uninformed good types admit uncertainty, and if $p_e \geq \frac{p_1}{2-p_1}$ there is an MSE where all of the good types send their honest message.

Proof. Part i is shown above. For part ii, first note the equilibrium constructed in proposition S.1 also holds with policy concerns: the policy choice upon observing both equilibrium messages is p_1, so each type's relative payoff in this equilibrium is unaffected by the value of γ. Since the good uninformed types always admit uncertainty in this equilibrium, this demonstrates the first claim.

Now suppose the good types all send their honest message. By the same fixed point argu-

ment as proposition S.2, the bad types must have at least one mixed strategy which is a best response given the good types strategy and DM strategy. What remains is to show the good types have no incentive to deviate from the honest message.

The message/validation combinations (m_0, e), (m_1, e), and $(m_\emptyset, h)$ are on-path and yield competence evaluations which are all strictly greater than zero.

Message/validation combinations (m_0, h), (m_1, h), and $(m_\emptyset, e)$ are never reached with a good type. So, if the bad types send those respective messages, they are on-path and the competence assessment must be zero. If these information sets are off-path the competence assessment can be set to zero.

Since only uninformed types send $m_\emptyset$, the policy choice upon observing $m_\emptyset$ must be $a^*(m_\emptyset) = p_1$. The m_0 message is sent by the informed type who knows $\omega = 0$, and potentially also by uninformed bad types, so $a^*(m_0) \in [0, p_1)$. Similarly, $a^*(m_1) \in (p_1, 1]$. So $a^*(m_0) < a^*(m_\emptyset) < a^*(m_1)$.

The good and uninformed type has no incentive to deviate from sending message $m_\emptyset$ because for $m \in \{m_0, m_1\}$, $\pi_g(m_\emptyset, h) > \pi_g(m, h)$ and $v(a^*(m_\emptyset), p_1) > v(a^*(m), p_1)$.

The s_0 type has no incentive to deviate to $m_\emptyset$ since $\pi_g(m_0, e) > \pi_g(m_\emptyset, e) = 0$ and $v(a^*(m_0), 0) > v(a^*(m_\emptyset), 0)$. Similarly, the s_1 type has no incentive to deviate to $m_\emptyset$.

So, the final deviations to check are for the informed types switching to the message associated with the other state; i.e., the s_0 types sending m_1 and the s_1 types sending m_0. Preventing a deviation to m_1 requires:

$$\begin{aligned} \pi_g(m_0, e) + \gamma v(a^*(m_0), 0) &\geq \pi_g(m_1, e) + \gamma v(a^*(m_1), 0) \\ \Delta_\pi + \gamma \Delta_v(0) &\leq 0, \end{aligned} \tag{8}$$

where $\Delta_\pi \equiv \pi_g(m_1, e) - \pi_g(m_0, e)$ is the difference in competence assessments from sending m_1 versus m_0 (when the problem is easy), and $\Delta_v(p) \equiv v(a^*(m_1), p) - v(a^*(m_0), p)$ is the difference in the expected quality of the policy when sending m_1 vs m_0 for an expert who believes $\omega = 1$ with probability p. This simplifies to:

$$\Delta_v(p) = (a^*(m_1) - a^*(m_0))(2p - a^*(m_1) - a^*(m_0)).$$

Since $a^*(m_1) > a^*(m_0)$, $\Delta_v(p)$ is strictly increasing in p, and $\Delta_v(0) < 0 < \Delta_v(1)$.

The analogous incentive compatibility constraint for the s_1 types is:

$$\Delta_\pi + \gamma \Delta_v(1) \geq 0 \tag{9}$$

If the bad types never send m_0 or m_1, then $\Delta_\pi = 0$, and (8)-(9) both hold. So, while not explicitly shown in the main text, in the honest equilibrium such a deviation is never profitable.

Now consider an equilibrium where the bad types send both m_0 and m_1, in which case they must be indifferent between both messages:

$$\begin{aligned} p_e \pi_g(m_0, e) + \gamma v(a^*(m_0), p_1) &= p_e \pi_g(m_1, e) + \gamma v(a^*(m_1), p_1) \\ p_e \Delta_\pi + \gamma \Delta_v(p_1) &= 0 \end{aligned} \tag{10}$$

Substituting this constraint into (8) and (9) and simplifying gives:

$$p_e \Delta_v(0) - \Delta_v(p_1) \leq 0 \tag{11}$$

$$p_e \Delta_v(1) - \Delta_v(p_1) \geq 0. \tag{12}$$

If $\Delta_v(p_1) = 0$ the constraints are both met. If $\Delta_v(p_1) < 0$ then the second constraint is

always met, and the first constraint can be written:

$$p_e \geq \frac{\Delta_v(p_1)}{\Delta_v(0)} = \frac{a^*(m_0) + a^*(m_1) - 2p_1}{a^*(m_0) + a^*(m_1)} \equiv \check{p}_\delta \tag{13}$$

This constraint is hardest to meet when $\check{p}_\delta$ is large, which is true when $a^*(m_0) + a^*(m_1)$ is high. The highest value this sum can take on is $p_1 + 1$, so $\check{p}_\delta \leq \frac{1-p_1}{1+p_1}$.

If $\Delta_v(p_1) > 0$, then the first constraint is always met, and the second constraint becomes:

$$p_e \geq \frac{\Delta_v(p_1)}{\Delta_v(1)} = \frac{2p_1 - (a^*(m_0) + a^*(m_1))}{2 - (a^*(m_0) + a^*(m_1))} \equiv \hat{p}_\delta \tag{14}$$

This is hardest to meet when $a^*(m_0) + a^*(m_1)$ is small, and the smallest value it can take on is p_1. Plugging this in, $\hat{p}_\delta \geq \frac{p_1}{2-p_1} \geq \check{p}_\delta$.

For $p_1 \geq 1/2$, $\hat{p}_\delta \geq \check{p}_\delta$. Without placing any further restrictions on the value of $a^*(m_0) + a^*(m_1)$, this constraint ranges from $\hat{p}_\delta \in (1/3, 1)$. Still, if p_e is sufficiently high, the informed types never have an incentive to deviate when the bad types send both m_0 and m_1.

If the bad types only send m_1 but not m_0, then the s_0 types get the highest possible payoff, so the relevant deviation to check is the s_1 types switching to m_0. The bad types sending weakly preferring m_1 implies $p_e\Delta_\pi + \gamma\Delta_v(p) \geq 0$, and substituting into equation 12 gives the same $p_e \geq \hat{p}_\delta$. Similarly, if the bad types only send m_0 but not m_1, then the relevant constraint is the s_0 types sending m_1, for which $p_e \geq \check{p}_\delta$ is sufficient.

Summarizing, a sufficient condition for the existence of a MSE where the good types report honestly (for any value of γ) is $p_e \leq p_g/(1 + p_g)$ (in which case $\gamma \leq \gamma^H_{DV}$), or $p_e \geq \frac{p_1}{2-p_1}$. This completes part ii.

Now to prove proposition 4 we first characterize the optimal strategy for the bad types as $\gamma \to 0$, assuming the good types send their honest message. If sending $m_\emptyset$, the expert will reveal his type if $\delta = e$, but appear partially competent if $\delta = h$, giving expected payoff

$$(1 - p_e)\frac{p_g}{p_g + (1 - p_g)\sigma_b(m_\emptyset)}.$$

When sending m_0, the expert will reveal his type if $\delta = h$ (as only bad types guess when the problem is hard), but look partially competent if $\delta = e$:

$$p_e\frac{p_g(1 - p_1)}{p_g(1 - p_1) + (1 - p_g)\sigma_b(m_0)}.$$

and when sending m_1 the expect payoff is:

$$p_e\frac{p_g p_1}{p_g p_1 + (1 - p_g)\sigma_b(m_1)}.$$

setting these three equal subject to $\sigma_b(m_0) + \sigma_b(m_1) + \sigma_b(m_\emptyset) = 1$ gives:

$$\begin{aligned}
\sigma_b(m_\emptyset) &= \frac{1 - p_e(1 + p_g)}{1 - p_g}; \\
\sigma_b(m_0) &= \frac{(1 - p_1)(p_e - p_g(1 - p_e))}{1 - p_g} \\
\sigma_b(m_1) &= \frac{p_1(p_e - p_g(1 - p_e))}{1 - p_g}.
\end{aligned}$$

These are all interior if and only if:

$$0 < \frac{1 - p_e(1 + p_g)}{1 - p_g} < 1 \Longrightarrow \frac{p_g}{1 + p_g} < p_e < \frac{1}{1 + p_g}.$$

If $p_e \leq \frac{p_g}{1+p_g}$, then there can be no equilibrium where the bad expert uses a fully mixed strategy because he would always prefer to send $m_\emptyset$; and recall this is exactly the condition for an honest equilibrium with no validation. If $p_e \geq \frac{1}{1+p_g}$, then the bad type always

guesses. Setting the payoff for a bad type sending m_0 and m_1 equal along with $\sigma_b(m_0) + \sigma_b(m_1) = 1$ gives the strategies in the statement of the proposition.

The final step is to ensure the informed types do not send the message associated with the other state. Recall the IC constraints depend on $a^*(m_0) + a^*(m_1)$, which we can now restrict to a narrower range given the bad type strategy:

$$\begin{aligned} a^*(m_0) + a^*(m_1) &= \frac{(1-p_g)p_1(1-p_1)(1-\sigma_b(m_\emptyset))}{p_e p_g(1-p_1+(1-p_g)(1-p_1)(1-\sigma_b(m_\emptyset)))} \\ &\quad + \frac{p_e p_g p_1 + (1-p_g)p_1 p_1(1-\sigma_b)(m_\emptyset)}{p_e p_g p_1 + (1-p_g)p_1(1-\sigma_b)(m_\emptyset)} \\ &= \frac{p_e p_g + (1-\sigma_b(m_\emptyset))(1-p_g)2p_1}{p_e p_g + (1-\sigma_b(m_\emptyset))(1-p_g)}. \end{aligned}$$

This can be interpreted as weighted average of 1 (with weight $p_e p_g$) and $2p_1 > 1$ (with weight $(1-\sigma_b(m_\emptyset)(1-p_g))$, and so must lie on $[1, 2p_1]$. So, (14) is always the binding constraint, and is hardest to satisfy when $a^*(m_0)+a^*(m_1) \to 1$, in which case the constraint becomes $\hat{p}_\delta = 2p_1 - 1$. So, $p_e \geq 2p_1 - 1$ is a sufficient condition for the informed types to never deviate. For any $p_e > 0$, this holds for p_1 sufficiently close to $1/2$, which completes the proof of proposition 4.

□

Here is an example where the constraint on the informed types is violated. Suppose p_1 is close to 1, and the bad types usually send m_1, and rarely m_0. Then the tradeoff they face is that sending m_1 leads to a better policy, but a lower competence payoff when the problem is easy (when the problem is hard, the competence payoff for either guess is zero). Now consider the good expert who observes signal s_1. Compared to the bad expert, this type has a marginally stronger incentive to send m_1 (since p_1 is close to 1). However, this type *knows* that he will face a reputational loss for sending m_1 rather than m_0, while the

bad type only experiences this loss with probability p_e. So, the bad type being indifferent means the type who knows the state is 1 has a strict incentive to deviate to m_0. In general, this deviation is tough to prevent when p_e is low and p_1 is close to 1, hence the condition in the proposition.

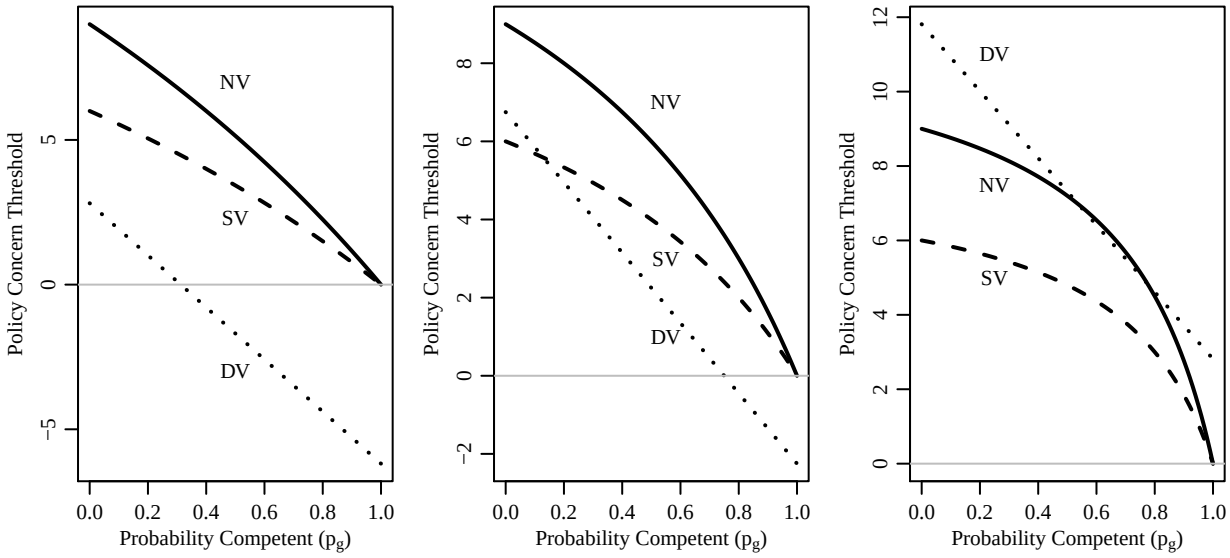
Comparative Statics: Difficulty Validation Can be the Wrong Kind of Transparency. As long as policy concerns are strictly positive but small, difficulty validation is more effective at eliciting honesty than state validation.

For larger policy concerns the comparison becomes less straightforward. Figure S.1 shows the policy concern threshold for honesty under no validation (solid line), state validation (dashed line), and difficulty validation (dotted line) as a function of the prior on the expert competence, when the problem is usually hard ($p_e = 0.25$, left panel), equally likely to be easy or hard ($p_e = 0.5$, middle panel) and usually easy ($p_e = 0.75$, right panel). In all panels $p_1 = 0.67$; changing this parameter does not affect the conclusions that follow.[3] For intuition, difficulty validation makes it hard to compensate bad experts for saying "I don't know," as there are fewer good experts who don't know. For very easy problems difficulty validation can be *worse* than no validation. This mirrors the result in Prat (2005), where transparency can eliminate incentives for bad types to pool with good types by exerting more effort.

This figure illustrates several results. First, in all cases, the policy concern threshold required is decreasing in p_g, which means it is easier to sustain honesty when the prior is that the expert is competent. This is because when most experts are competent in general, most uninformed experts are competent as well, and so there is less of a penalty for admitting uncertainty. Second, the threshold with state validation is always lower than the threshold

[3] In general, honesty is easier to sustain under all validation regimes when p_1 is lower, with state validation being particularly sensitive to this change.

Figure S.1: Comparative Statics of Honesty Threshold



Notes: Comparison of threshold in policy concerns for full honesty under different validation regimes as a function of p_g. The panels vary in the likelihood the problem is solvable, which is 0.25 in the left panel, 0.5 in the middle panel, and 0.75 in the right panel.

with no validation, though these are always strictly positive as long as $p_g < 1$. Further, for most of the parameter space these thresholds are above two, indicating the expert must care twice as much about policy than about perceptions of his competence to elicit honesty. On the other hand, in the right and middle panels there are regions where the threshold with difficulty validation is below zero, indicating no policy concerns are necessary to induce admission of uncertainty (in fact, the expert could want the decision-maker to make a *bad* decision and still admit uncertainty).

Finally, consider how the relationship between the thresholds changes as the problem becomes easier. When problems are likely to be hard (left panel), difficulty validation is the best for eliciting honesty at all values of p_g. In the middle panel, difficulty validation is always better than no validation, but state validation is best for low values of p_g. When the problem is very likely to be easy, difficulty validation is always worse than state validation and is even worse than even no validation other than for a narrow range of p_g.

However, even in this case difficulty validation still can elicit honesty from good but uninformed experts when policy concerns are not high enough, while there is no admission of uncertainty at all when policy concerns are not high enough with no validation and state validation.

2 Relabeling

Some are our formal results only rely on the existence of equilibria with certain properties. For these results the fact that we often restrict attention to the $(m_0, m_1, m_\emptyset)$ message set poses no issues: it is sufficient to show that there is an equilibrium of this form with the claimed properties. However, propositions 2, 3, 10ii-iii, and 11, make claims that all (non-

babbling) MSE have certain properties.[4] The proofs show that all equilibrium where the s_0 and s_1 types send distinct and unique messages (labelled m_0 and m_1) and there is at most one other message (labelled $m_\emptyset$) have these properties. Here we show this is WLOG in the sense that with no validation or state validation, any non-babbling equilibrium can be relabeled to an equilibrium of this form.

Consider a general messaging strategy where $M \subseteq \mathcal{M}$ is the set of messages sent with positive probability. Write the probability that the informed types observing s_0 and s_1 and $\sigma_0(m)$ and $\sigma_1(m)$. When the good and bad uninformed types are not necessarily payoff equivalent we write their strategies $\sigma_{\theta,\emptyset}(m)$. When these types are payoff equivalent and hence play the same strategy, we drop the θ: $\sigma_\emptyset(m)$. Similarly, let M_0 and M_1 be the set of messages sent by the respective informed types with strictly positive probability, and $M_{g,\emptyset}$, $M_{b,\emptyset}$, and $M_\emptyset$ the respective sets for the uninformed types, divided when appropriate.

As is standard in cheap talk games, there is always a babbling equilibrium:

Proposition S.6. *There is a class of babbling equilibria where* $\sigma_0(m) = \sigma_1(m) = \sigma_{g,\emptyset}(m) = \sigma_{b,\emptyset}(m)$ *for all* $m \in M$.

Proof. If all play the same mixed strategy, then $\pi_g(m, \mathcal{I}_{DM2}) = p_g$ and $a^*(m, \mathcal{I}_{DM}) = p_1$ for any $m \in M$ and $\mathcal{I}_{DM}$. Setting the beliefs for any off-path message to be the same as the on-path messages, all types are indifferent between any $m \in \mathcal{M}$. □

The next result states that for all cases with either state validation or policy concerns, in any non-babbling equilibrium the informed types send no common message (note this result does *not* hold with difficulty validation; in fact, the proof of proposition S.1 contains a counterexample):

[4] Proposition 8 also makes a claim about all equilibria, but this is already proven in Appendix B of the published version.

Proposition S.7. *With either no validation or state validation (and any level of policy concerns), any MSE where $M_0 \cap M_1 \neq \emptyset$ is babbling, i.e., $\sigma_0(m) = \sigma_1(m) = \sigma_{g,\emptyset}(m) = \sigma_{b,\emptyset}(m)$ for all $m \in M$.*

Proof. We first prove the result with state validation, and then briefly highlight the aspects of the argument that differ with no validation.

Recall that for this case the good and bad uninformed types are payoff equivalent, so we write their common message set and strategy $M_\emptyset$ and $\sigma_\emptyset(m)$ The proof proceeds in three steps.

Step 1: If $M_0 \cap M_1 \neq \emptyset$, then $M_0 = M_1$. Let $m_c \in M_0 \cap M_1$ be a message sent by both informed types. Suppose there is another message sent only by the s_0 types: $m_0 \in M_0 \backslash M_1$. For the s_0 type to be indifferent between m_0 and m_c:

$$\pi_g(m_c, 0) + \gamma v(a^*(m_c), 0) = \pi_g(m_0, 0) + \gamma v(a^*(m_0), 0).$$

For this equation to hold, it must be the case that the uninformed types send m_0 with positive probability: if not, then $\pi_g(m_c, 0) \leq \pi_g(m_0, 0) = 1$, but $v(a^*(m_c), 0) < 1 = v(a^*(m_0), 0)$, contradicting the indifference condition.

For the uninformed types to send m_0, it must also be the case that his expected payoff for sending this message, which can be written

$$p_1(\pi_g(m_0, 1) + \gamma v(a^*(m_0), 1)) + (1 - p_1)(\pi_g(m_0, 0) + v(a^*(m_0), 0))$$

– is at least his payoff for sending m_c:

$$p_1(\pi_g(m_c, 1) + \gamma v(a^*(m_c), 1)) + (1 - p_1)(\pi_g(m_c, 0) + v(a^*(m_c), 0)).$$

The second terms, which both start with $(1 - p_1)$, are equal by the indifference condition for s_0 types, so this requires:

$$\pi_g(m_0, 1) + \gamma v(a^*(m_0), 1) \geq \pi_g(m_c, 1) + \gamma v(a^*(m_c), 1).$$

Since m_0 is never sent by the s_1 types, $\pi_g(m_0, 1) = \pi_g^{\emptyset}$, while $\pi_g(m_c, 1) > \pi_g^{\emptyset}$. So, this inequality requires $v(a^*(m_0), 1) > v(a^*(m_c), 1)$, which implies $a^*(m_0) > a^*(m_c)$. A necessary condition for this inequality is $\frac{\sigma_\emptyset(m_0)}{\sigma_0(m_0)} > \frac{\sigma_\emptyset(m_c)}{\sigma_0(m_c)}$, which also implies $\pi_g(m_c, 0) > \pi_g(m_0, 0)$. But if $a^*(m_0) > a^*(m_c)$ and $\pi_g(m_c, 0) > \pi_g(m_0, 0)$, the s_0 types strictly prefer to send m_c rather than m_0, a contradiction. By an identical argument, there can be no message in $M_1 \setminus M_0$, completing step 1.

Step 2: If $M_0 = M_1$, then $\sigma_0(m) = \sigma_1(m)$ for all m. If $M_0 = M_1$ is a singleton, the result is immediate. If there are multiple common messages and the informed types do not use the same mixed strategy, there must be a message m^0 such that $\sigma_0(m^0) > \sigma_1(m^0) > 0$ and another message m^1 such that $\sigma_1(m^1) > \sigma_0(m^1) > 0$. (We write the message "generally sent by type observing s_x" with a superscript to differentiate between the subscript notation referring to messages always sent by type s_x.) The action taken by the DM upon observing m^0 must be strictly less than p_1 and upon observing m^1 must be strictly greater than p_1,[5] so $a^*(m^0) < a^*(m^1)$.

[5]The action taken upon observing m can be written $\mathbb{P}(s_1|m) + p_1\mathbb{P}(s_\emptyset|m)$. Rearranging, this is greater than p_1 if and only if $\frac{\mathbb{P}(s_1,m)}{\mathbb{P}(s_1,m)+\mathbb{P}(s_0,m)} > p_1$ which holds if and only if $\sigma_1(m) > \sigma_0(m)$.

Both the s_1 and s_0 types must be indifferent between both messages, so:

$$\pi_g(m^0, 0) + \gamma v(a^*(m^0), 0) = \pi_g(m^1, 0) + \gamma v(a^*(m^1), 0)$$
$$\pi_g(m^0, 1) + \gamma v(a^*(m^0), 1) = \pi_g(m^1, 1) + \gamma v(a^*(m^1), 1)$$

Since $v(a^*(m^0), 0) > v(a^*(m^1), 0)$, for the s_0 to be indifferent it must be the case that $\pi_g(m^0, 0) < \pi_g(m^1, 0)$. Writing out this posterior belief:

$$\mathbb{P}(\theta = g|m, 0) = \frac{(1-p_1)(p_g(p_e\sigma_0(m) + (1-p_e)\sigma_\emptyset(m))}{(1-p_1)(p_g p_e\sigma_0(m) + (1-p_g p_e)\sigma_\emptyset(m)}.$$

Rearranging, $\pi_g(m^0, 0) < \pi_g(m^1, 0)$ if and only if $\frac{\sigma_0(m^0)}{\sigma_0(m^1)} < \frac{\sigma_\emptyset(m^0)}{\sigma_\emptyset(m^1)}$. Similarly, it must be the case that $\pi_g(m^1, 1) < \pi_g(m^0, 1)$, which implies $\frac{\sigma_1(m^0)}{\sigma_1(m^1)} > \frac{\sigma_\emptyset(m^0)}{\sigma_\emptyset(m^1)}$. Combining, $\frac{\sigma_0(m^0)}{\sigma_0(m^1)} < \frac{\sigma_1(m^0)}{\sigma_1(m^1)}$, which contradicts the definition of these messages. So, $\sigma_0(m) = \sigma_1(m)$ for all m.

Step 3: If $M_0 = M_1$ and $\sigma_0(m) = \sigma_1(m)$, then $M_\emptyset = M_0 = M_1$ and $\sigma_\emptyset(m) = \sigma_0(m) = \sigma_1(m)$. By step 2, it must be the case that $a^*(m) = p_1$ for all messages sent by the informed types. So, the uninformed types can't send a message not sent by the informed types: if so, the payoff would be at most $\pi_g^\emptyset + \gamma v(p_1, p_1)$, which is strictly less than the payoff for sending a message sent by the informed types. If there is only one message in M then the proof is done. If there are multiple types, all must be indifferent between each message, and by step 2 they lead to the same policy choice. So, they must also lead to the same competence assessment for each revelation of ω, which is true if and only if $\sigma_\emptyset(m) = \sigma_0(m) = \sigma_1(m)$. □

Next, consider the no validation case. For step 1, define m_0 and m_1 analogously. The uninformed types must send m_0 by the same logic, and these types at least weakly prefer

sending this to m_c (while the s_0 types are indifferent) requires:

$$\pi_g(m_0) + \gamma v(a^*(m_0), 1) \geq \pi_g(m_c) + \gamma v(a^*(m_c), 1).$$

This can hold only weakly to prevent the s_1 types from sending m_0 (as required by the definition). Combined with the s_0 indifference condition:

$$\pi_g(m_0) - \pi_g(m_c) = \gamma v(a^*(m_c), 1) - \gamma v(a^*(m_0), 1) = \gamma v(a^*(m_c), 0) - \gamma v(a^*(m_0), 0),$$

which requires $a^*(m_0) = a^*(m_c)$. Since the s_1 types send m_c but not m_0 this requires $\frac{\sigma_\emptyset(m_0)}{\sigma_0(m_0)} > \frac{\sigma_\emptyset(m_c)}{\sigma_0(m_c)}$, which implies $\pi_g(m_0) < \pi_g(m_c)$, contradicting the s_0 types being indifferent between both messages.

Steps 2 and 3 follow the same logic.

□

Finally, we prove that any MSE where the messages sent by the s_0 and s_1 types do not overlap is equivalent to an MSE where there is only one message sent by each of these types and only one "other" message. This provides a formal statement of our claims about equilibria which are "equivalent subject to relabeling":

Proposition S.8. *Let $M_U = M_\emptyset \backslash (M_0 \cup M_1)$ (i.e., the messages only sent by the uninformed types). With no validation or state validation:*
i. In any MSE where $M_0 \cap M_1 = \emptyset$, for $j \in \{0, 1, U\}$, and any $m', m'' \in M_j$, $a^*(m') = a^*(m'')$ and $\pi_g(m', \mathcal{I}_{DM2}) = \pi_g(m'', \mathcal{I}_{DM2})$
ii. Take an MSE where $|M_j| > 1$ for any $j \in \{0, 1, U\}$, and the equilibrium actions and posterior competence assessments for the messages in this set are $a^*(m_i)$ and $\pi_g(m_i, \mathcal{I}_{DM2})$ (which by part i are the same for all $m_i \in M_j$). Then there is another MSE where $M_j =$

$\{m\}$*, and equilibrium strategy and beliefs* a^*_{new} *and* $\pi_{g,new}$ *such that* $a^*(m_i) = a^*_{new}(m)$*, and* $\pi_g(m_i, \mathcal{I}_{DM2}) = \pi_{g,new}(m, \mathcal{I}_{DM2})$

Proof. For part i, first consider the message in M_U. By construction the action taken upon observing any message in this set is p_1. And since the good and bad uninformed types are payoff equivalent and use the same strategy, the competence assessment upon observing any message in this set must be $\pi_g^{\emptyset}$.

For M_0, first note that for any $m', m'' \in M_0$, it can't be the case that the uninformed types only send one message but not the other with positive probability; if so, the message not sent by the uniformed types would give a strictly higher payoff for the s_0 types, and hence they can't send both messages. So, either the uninformed types send neither m' nor m'', in which case the result is immediate, or they send both, in which case they must be indifferent between both. As shown in the proof of proposition S.7, this requires that the action and competence assessment are the same for both m' and m''. An identical argument holds for M_1, completing part i.

For part ii and $M_{\emptyset}$, the result immediately follows from the same logic as part i.

For M_0, if the uninformed types do not send any messages in M_0, then the on-path response to any $m_0^j \in M_0$ are $a^*(m_0^j) = 0$ and $\pi_g(m_0^j, 0) = 1$. Keeping the rest of the equilibrium fixed, the responses in a proposed MSE where the s_0 types always send m_0 are also $a^*_{\text{new}}(m_0) = 0$ and $\pi_{g,\text{new}}(m_0^j, 0) = 1$. So there is an MSE where the s_0 types all send m_0 which is equivalent to the MSE where the s_0 types send multiple messages.

If the uninformed types do send the messages in M_0, then part i implies all messages must lead to the same competence evaluation, which implies for any $m_0', m_0'' \in M_0$, $\frac{\sigma_{\emptyset}(m_0')}{\sigma_0(m_0')} = \frac{\sigma_{\emptyset}(m_0'')}{\sigma_0(m_0'')} \equiv r_0$. In the new proposed equilibrium where $M_0 = \{m_0\}$, set $\sigma_{0,\text{new}}(m_0) = 1$ and

$\sigma_{\emptyset,\text{new}}(m_0) = r_0$. Since $\frac{\sigma_{\emptyset,\text{new}}(m_0)}{\sigma_{0,\text{new}}(m_0)} = \frac{\sigma_\emptyset(m_0')}{\sigma_0(m_0')}$, $a^*_{\text{new}}(m_0) = a^*(m_0')$ and $\pi_{g,\text{new}}(m_0', 0) = 1$, and all other aspects of the MSE are unchanged. □

3 Alternative Signal Structure

In this section, we consider two alternative signal specifications.

3.1 More general binary signal structure

Here is a more general formulation of the signal structure. We again assume a binary incumbent type $\theta \in \{g, b\}$ and problem difficulty $\delta \in \{e, h\}$. Now assume that the signal is given by:

$$s = \begin{cases} s_\omega & \text{with probability } P(\theta, \delta) \\ s_\emptyset & \text{o.w} \end{cases} \tag{15}$$

where $P(g, \delta) \geq P(b, \delta)$ (with the inequality strict for at least one δ) and $P(\theta, e) \geq P(\theta, h)$ (with the inequality strict for at least one θ). That is, more competent experts are (weakly) more likely to get an informative signal for either problem difficulty, and easy problems are (weakly) more likely to result in an informative signal. We assume that at least one of the inequalities is strict so that both variable "matter".

All other aspects of the model are the same as in the main text.

The analysis in the main text is a special case of these assumptions where $P(\theta, \delta)$ is equal to 1 if $\theta = g$ and $\delta = e$ and zero otherwise. With the more general signal structure, there

are now up to 6 potential types, as a function of the expert signal and competence. A type of competence θ who observes $s_x, x \in \{0,1\}$ (if such a signal is possible for type θ; recall this is not possible in the main formulation for $\theta = b$ types) knows that $\omega = x$, and his posterior belief about the problem difficulty is:

$$Pr(\delta = e|s_x, \theta) = \frac{p_e p_x p_\theta P(\theta, e)}{p_e p_x p_\theta P(\theta, e) + p_h p_x p_\theta P(\theta, h)} = \frac{p_e p_\theta P(\theta, e)}{p_e p_\theta P(\theta, e) + p_h p_\theta P(\theta, h)}. \quad (16)$$

Since $P(\theta, e) \geq P(\theta, h)$, $Pr(\delta = e|s_x, \theta) \geq p_e$, and if $P(\theta, e) > P(\theta, h)$ the inequality is strict. That is, since each type is (weakly) more likely to observe an informative signal when the problem is easy, they are (weakly) more likely to believe the problem is easy given an informative signal.

Also important for what comes, both types have an equal belief about the problem difficulty if and only if $\frac{P(g,e)}{P(g,h)} = \frac{P(b,e)}{P(b,h)}$. This condition might hold. For example, suppose $P(b,h) = 1/4$, $P(b,e) = 1/2$, $P(g,h) = 1/2$, and $P(g,e) = 1$. Then both types are twice as likely to receive an informative signal when the problem is easy, and hence learn the same about the problem difficulty from getting an informative signal. However, this is a knife-edged condition, and if $\frac{P(g,e)}{P(g,h)} > \frac{P(b,e)}{P(b,h)}$ the competent type will update about the easiness of the problem more sharply and if $\frac{P(g,e)}{P(g,h)} < \frac{P(b,e)}{P(b,h)}$ the bad type will update more in the positive direction.

A type of competence θ who observes $s = s_\emptyset$ maintains his prior belief about the state ($Pr(\omega = 1|s_\emptyset) = p_1$) and his belief about the problem difficulty becomes:

$$Pr(\delta = e|s_\emptyset, \theta) = \frac{p_e p_\theta (1 - P(\theta, e))}{p_e p_\theta (1 - P(\theta, e)) + p_h p_\theta (1 - P(\theta, h))}. \quad (17)$$

Following a similar logic as the above, both sides will (weakly) come to believe the problem is less likely to be easy when getting an uninformative signal. These updates are equal if

and only if $\frac{1-P(g,e)}{1-P(g,h)} = \frac{1-P(b,e)}{1-P(b,h)}$

As with the example in the main text, MSE helps quickly pin down which types can send different messages in equilibrium.

No policy concerns, no validation The analysis with no policy concerns or validation is identical: all types are payoff equivalent, and any equilibrium is babbling.

No policy concerns, state validation With state validation, types who observe different signals have different beliefs about ω, but types observing the same signal are still payoff equivalent regardless of their competence.[6] However, this will never induce honesty for a similar reason as the main model. In an honest equilibrium, the posterior belief about the expert competence when observing $(m_1, \omega = 1)$ is:

$$Pr(\theta = g|m_1, \omega = 1) = \frac{p_1 p_g (p_e P(g,e) + p_h P(g,h))}{p_1 p_g (p_e P(g,e) + p_h P(g,h)) + p_1 p_b (p_e P(b,e) + p_h P(b,h))} > p_g, \tag{18}$$

where the inequality follows from the assumption that our assumption that $P(g,\delta) \geq P(b,\delta)$ for both δ and one of the inequalities is strict, and hence $(p_e P(g,e) + p_h P(g,h)) > (p_e P(b,e) + p_h P(b,h))$.

Similarly, $Pr(\theta = g|m_1, \omega = 1) > p_g$ and $Pr(\theta = g|m_\emptyset) < p_g$. Given the Markov strategies requirement, Markov consistency implies that the worst inference the DM can make about the expert competence upon observing something off path is $Pr(\theta = g|s =$

[6]Depending on the P function they might have different views of the problem difficulty for some signals, but since there is no difficulty validation this does not affect their expected payoff for any possible DM strategy.

$s_\emptyset) = Pr(\theta = g|m_\emptyset)$. So, the expected utility of sending m_1 is:

$$p_1 Pr(\theta = g|m_1, \omega = 1) + (1 - p_1)Pr(\theta = g|s_\emptyset) > Pr(\theta = g|m_\emptyset) \tag{19}$$

and hence this is a profitable deviation. So, there is no honest MSE with state validation alone.

Difficulty validation and small policy concerns With difficulty validation and small policy concerns, no two types with different beliefs about the difficulty of the problem are always payoff equivalent. So, as long as $\frac{P(g,e)}{P(g,h)} \neq \frac{P(b,e)}{P(b,h)}$ and $\frac{1-P(g,e)}{1-P(g,h)} \neq \frac{1-P(b,e)}{1-P(b,h)}$, the Markov strategies and Markov consistency requirements have no bite, making it possible to punish those who guess incorrectly with a belief that they are competent with probability zero.

However, an important aspect for this to make honesty possible is for sending an informative message when the problem is hard is actually off-path. In a proposed honest equilibrium with just difficulty validation, if $P(\theta, h) > 0$ for some θ, then not only are message/validation combinations (m_1, h) and (m_1, h) on path, but tend to be reached when the expert is competent.

So, an important assumption to make difficulty (and small policy) concerns effective at inducing honesty is that $P(\theta, h) = 0$ (as was true in the main model). In words, this implies that there are not only relatively hard or easy questions that the expert might be asked, but there are *impossible* questions. We think in most domains this is reasonable, particularly if we interpret informative messages as stating that the state is zero or one with certainty.

If so, the key constraint for sustaining an honest equilibrium is that good and bad un-

informed experts face no incentive to guess. The payoff to admitting uncertainty (i.e., sending $m_\emptyset$) for type θ is:

$$Pr(\delta = e|s_\emptyset, \theta)Pr(\theta = g|s_\emptyset, \delta = e) + Pr(\delta = h|s_\emptyset, \theta)Pr(\theta = g|s_\emptyset, \delta = h). \tag{20}$$

The $Pr(\delta|s_\emptyset, \theta)$ are derived above, and the second halves are:

$$Pr(\theta = g|s_\emptyset, \delta) = \frac{p_g(1 - P(g, \delta))}{p_g(1 - P(g, \delta) + (1 - p_g)(1 - P(b, \delta))}. \tag{21}$$

The payoff to guessing m_1 (which is a better deviation than m_0 as long as $p_1 \geq 1/2$.) is the probability that his guess is correct and the problem is easy (since an informative signal with a hard problem is off path), times the competence evaluation in this circumstance:

$$p_1 Pr(\delta = e|s_\emptyset, \theta)Pr(\theta = g|s_1, m_1, \delta = e), \tag{22}$$

where

$$Pr(\theta = g|s_1, m_1, \delta = e) = \frac{p_g P(g, e)}{p_g P(g, e) + (1 - p_g)P(b, e)}. \tag{23}$$

So, as $\gamma \to 0$, the condition for an honest equilibrium is that:

$$\begin{aligned} &Pr(\delta = h|s_\emptyset, \theta)Pr(\theta = g|s_\emptyset, \delta = h) \geq \\ &Pr(\delta = e|s_\emptyset, \theta)(p_1 Pr(\theta = g|s_1, m_1, \delta = e) - Pr(\theta = g|s_\emptyset, \delta = e)) \end{aligned} \tag{24}$$

for $\theta \in \{g, b\}$. While the algebra is messier, the core idea is just like in the main case. The trade-off here is that if the problem turns out to be easy it can be more profitable to guess, while when the problem turns out to be hard it is better to admit uncertainty.

Full Validation, no policy concerns Finally, consider the full validation case. Because of state validation it is possible for experts with different views about the state to send different messages, and as long as $\frac{P(g,e)}{P(g,h)} \neq \frac{P(b,e)}{P(b,h)}$ and $\frac{1-P(g,e)}{1-P(g,h)} \neq \frac{1-P(b,e)}{1-P(b,h)}$ it is possible to set any off path beliefs to zero.

This allows for the possibility of an honest equilibrium even without the assumption that some problems are impossible, since in an honest equilibrium incorrect guesses are never on path, and unlike the case with just state validation can be punished with an off-path belief that the expert must be the bad type, regardless of what the difficulty validation says.

The utility for admitting uncertainty in such an honest equilibrium is again given by equation (20). The expected utility for guessing 1 (assuming that sending an informative signal when validation reveals that $\delta = h$ is on-path) is:

$$p_1(p_e Pr(\theta = g|s_1, \delta = e) + p_h Pr(\theta = g|s_1, \delta = h)) \tag{25}$$

so an honest equilibrium is possible if:

$$\begin{aligned} &p_e(Pr(\theta = g|s_\emptyset, \delta = e) - p_1 Pr(\theta = g|s_1, \delta = e)) \\ &+ p_h(Pr(\theta = g|s_\emptyset, \delta = h) - p_1 Pr(\theta = g|s_1, \delta = h)) \geq 0. \end{aligned} \tag{26}$$

As $p_1 \to 1$, this inequality never holds, but for smaller p_1 it is possible.

3.2 Continuous expert competence and question difficulty

Now let the competence of the expert θ and the difficulty of the problem δ both be uniform on $[0, 1]$. Let the private signal to the expert be:

$$s = \begin{cases} s_0 & \omega = 0, \theta > \beta\delta \\ s_1 & \omega = 1, \theta > \beta\delta \\ s_\emptyset & \text{o/w} \end{cases}$$

where $\beta > 0$.

If $\beta < 1$, then even the hardest problems ($\delta = 1$) are solvable by a strictly positive proportion of experts. If $\beta > 1$, then there are some problems which are so difficult that no expert can solve them. For reasons which will become apparent, we focus on the case where $\beta > 1$, and so $\overline{\delta} = 1/\beta < 1$ is the "least difficult unsolvable problem".

The expert learns s and θ, which is always partially informative about δ. In particular, an expert who gets an informative signal knows that $\delta \in [0, \theta/\beta]$, and an expert who does not get an informative signal knows that $\delta \in [\theta/\beta, 1]$. An interesting contrast with the binary model is that better experts don't always know more about the problem difficulty: when they learn the state, the range of possible values of δ is increasing in θ. However, when the expert learns the state (particularly with state validation) knowing the difficulty is not particularly relevant. On the other hand, when the expert is uninformed those who are more competent can restrict the difficulty of the problem to a smaller interval.

As with the binary model, we search for honest equilibria in the sense that the expert fully separates with respect to their signal (if not with respect to their competence). Here we only consider the case with no policy concerns.

No validation, State validation, Difficulty validation Even though the state space is much larger in this version of the model, with no validation (and no policy concerns) all types are still payoff equivalent. So any MSE must be babbling.

Similarly, with state validation there are now multiple types (differentiated by θ) who learn the state and multiple types who do not learn the state. However, since the knowledge of the state is the only payoff-relevant component of the type space, all types observing a particular s must play the same strategy in an MSE. So, by the same logic as the binary model, there is no honest MSE.

Difficulty validation alone (and again with no policy concerns; with small policy concerns honesty might possible for some parameters) also hits the same problem as in the binary model. Among the informed types with competence θ, those observing s_0 and those observing s_1 are payoff equivalent, and so no information can be communicated about the state. It is possible that information about the difficulty of the problem can be conveyed.

Full Validation Now consider the full validation case. No pairs of types are payoff equivalent, since even those observing the same signal have different beliefs about what the difficulty validation will reveal for each value of θ. So, it is possible to use punitive off-path beliefs where those who guess incorrectly or when no expert could solve the problem, which is possible when $\beta > 1$.

We now show an honest equilibrium can be possible in this case. First, consider the on-path inferences by the DM. When seeing a correct message and difficulty δ, the DM knows the expert competence must be on $[\beta\delta, 1]$, and so the average competence assessment is:

$$\pi_g(m_1; \omega = 1, \delta) = \pi_g(m_0, \omega = 0, \delta) = \frac{\beta\delta + 1}{2}$$

which is at least $1/2$, and increasing in δ.

Upon observing $m_\emptyset$ and δ, there are two possible cases. If $\delta > 1/\beta$, then no expert could have solved the problem, and so there is no information conveyed about the expert competence. If $\delta < 1/\beta$, then the DM learns that the expert competence is uniform on $[0, \beta\delta]$. Combining:

$$\pi_g(m_\emptyset, \omega, \delta) = \begin{cases} 1/2 & \delta > 1/\beta \\ \frac{\beta\delta}{2} & \delta \leq 1/\beta \end{cases}.$$

All other message and validation combinations are off-path, and can be set to zero.

Now consider the expert payoffs.

An informed expert (of any competence) gets a payoff of $\frac{\beta\delta+1}{2} > 1/2$ for sending the equilibrium message, 0 for sending the other informed message (i.e., m_1 rather than m_0 when $s = s_0$), and $\pi_g(m_\emptyset, \omega, \delta) \leq 1/2$ for sending $m_\emptyset$. So these types never deviate.

Uninformed experts know the difficulty – which again will be revealed to the DM – is uniform on $[\theta/\beta, 1]$. Note that for all but the (measure zero) $\theta = 1$ types, $1/\beta$ lies on this interval. So, all but the most competent experts don't know for sure if the problem is solvable by some experts, though very competent experts can become nearly certain the problem is unsolvable.

We can write the expected competence for admitting uncertainty to be the probability that $\delta \geq 1/\beta$ times $1/2$, plus the probability that $\delta < 1/\beta$ times the average competence assigned on this interval. Since the competence assessment is linear in δ on this interval, ranging from $\frac{\beta(\theta/\beta)}{2} = \frac{\theta}{2}$ to $1/2$, this average is $\frac{\theta+1}{4}$. Combining, the expected competence

for sending $m_\emptyset$ is:

$$\begin{aligned}\mathbb{E}_\delta[\pi_g(m_\emptyset, \omega, \delta)] &= Pr(\delta \leq 1/\beta)\frac{\theta+1}{4} + Pr(\delta > 1/\beta)\frac{1}{2} \\ &= \frac{1/\beta - \theta\beta}{1-\theta/\beta}\frac{\theta+1}{4} + \frac{1-1/\beta}{1-\theta/\beta}\frac{1}{2} \\ &= \frac{1-\theta}{\beta-\theta}\frac{\theta+1}{4} + \frac{\beta-1}{\beta-\theta}\frac{1}{2}\end{aligned}$$

which is increasing in θ.

Next consider the payoff for sending m_1; as before, this is the "better guess" since it is more likely to be matched by the state validation. This will lead to a competence evaluation of $\frac{\beta\delta+1}{2}$ if $\omega = 1$ (probability p_1) and if $\delta < 1/\beta$ (probability $\frac{1-\theta}{\beta-\theta}$), and 0 otherwise. Since the guessing correct payoff is linear in δ and the belief about δ conditional on a solvable problem is uniform on $[\theta/\beta, 1/\beta]$, the average competence assessment when getting away with a guess is:

$$\frac{\frac{\theta+1}{2}+1}{2} = \frac{3+\theta}{4}.$$

So the payoff to this deviation is:

$$p_1\frac{1-\theta}{\beta-\theta}\frac{3+\theta}{4}$$

which is decreasing in θ.

So, the binding constraint is that the $\theta = 0$ prefers sending $m_\emptyset$, which again reinforces the

assumption made about off-path beliefs. Honesty is possible when:

$$\frac{1}{\beta}\frac{1}{4} + \frac{\beta - 1}{\beta}\frac{1}{2} \geq p_1 \frac{1}{\beta}(3/4)$$

$$\beta \geq (3/2)p_1 + 1/2.$$

Since $p_1 \in [1/2, 1]$, this threshold ranges from $5/4$ to 2. That the threshold in β is strictly greater than 1 means that there must be some possibility of getting caught answering an unanswerable question. The threshold is lower when p_1 is lower since this makes guessing less attractive as one is more likely to be caught guessing wrong.

References

Prat, A. (2005). The wrong kind of transparency. *American Economic Review*, 95(3):862–877.