Too Good To Be True?
Retention Rules for Noisy Agents

Francisco Espinosa  Debraj Ray†

September 2020

Abstract. An agent who privately knows his type (good or bad) seeks to be retained by a principal. A principal seeks to retain good agents. Agents signal their type with some ambient noise, but can alter this noise, perhaps at some cost. Our main finding, that we examine in several extensions, is that in equilibrium, the principal treats extreme signals in either direction with suspicion, and retains the agent if and only if the signal falls in some intermediate bounded set. In short, she follows the maxim: “if it seems too good to be true, it probably is.” We consider extensions and applications, including non-normal signal structures, dynamics with term limits, risky portfolio management, and political risk-taking.

1. Introduction

Consider an agent who privately knows his type (good or bad) and seeks to be retained by a principal. The principal wishes to retain a good type, and to remove a bad type. The agent generates a noisy but informative signal of his type. He can choose to amplify or reduce the precision of this process, but there are two restrictions. First, the signal structure is constrained by the type; specifically, the mean of the signal is given by the type. Second, signal realizations cannot be tampered with ex post. That is, a specific realization cannot be augmented nor reduced: there is no “free disposal.” The principal observes the signal realization (but not the signal structure, or at least not fully), and makes a retention decision.

The equilibria of such a game — and some variants of it — form the subject matter of our paper. A central result that we examine from various angles is that in any equilibrium, the principal treats both kinds of excessive signals with suspicion, and retains the agent if and only if the signal falls in some intermediate bounded set. In short, she follows the maxim: “if it seems too good to be true, it probably is.”

In our baseline setting, the agent emits a normal signal centered around some mean, which is his type. This centering cannot be changed, but the variance can be altered at no cost, subject only to a minimum lower bound. The principal sees the outcome, and retains if and only if her posterior on a good type exceeds some threshold. Our initial discussion and Observation 1 together argue that there are three types of potential equilibria. The first is monotone retention, in which both types choose the same noise, and the principal retains if the signal is above some threshold. The second is bounded retention, in which

†Espinosa: Harris School of Public Policy, University of Chicago, fespinosa@uchicago.edu; Ray: New York University and University of Warwick, debraj.ray@nyu.edu. Ray acknowledges funding under National Science Foundation grant SES-1851758. We thank Dilip Abreu, Dhruva Bhaskar, Wioletta Dziuda, Gabriele Gratton, Emir Kamenica, Gaute Torvik and participants at the University of Chicago Harris Political Economy Lunches for useful comments.  indicates author names are in random order.
the bad agent chooses higher noise than the good agent, and the principal retains for intermediate signal realizations. The third is *bounded replacement*, in which all comparisons are reversed, and the principal replaces the agent for intermediate signal realizations.

Our baseline result is Proposition 1, which singles out *just* the bounded retention equilibrium when the ambient noise level is sufficiently low.

Our framework is stark and minimal, but easily extendable. In Section 6, we study some extensions that show the robustness of this observation and accommodate various ancillary features. These include costly noise, a dynamic setting with agent term limits, and non-normal signal structures. In the Supplementary Appendix, we study other extensions: costly mean-shifting noise, non-binary types, and *principals* who inject noise into their assessments. In a separate paper (Espinosa et al. 2020), we study an extension to commitment, in which the principal pre-announces mechanisms to assess agents.

We also consider two applications in Section 7. One is portfolio management, in which a money manager with career concerns might overload on risk. The principal — his client — might be able to verify the portfolio at any one point of time, so that there is no chance of ex-post “disposal” of financial returns, but may not be fully aware of the risk implications of any particular portfolio. The objective of the client is to make money, and so to find a durable relationship with a competent money manager. So the return today has intrinsic value, but also serves as a signal about the manager’s type. A second application is to a political leader, who seeks to be “retained” by the median voter (who plays here the role of the principal). A competent leader might play it safe by implementing solid but unambitious policies. In contrast, an incompetent leader can entertain a risky policy. These choices of risks may not be fully observable ex ante. Our model suggests that a striking success should be treated with a certain degree of reticence: it could be a sign of extreme competence, or the fortuitous outcome of a desperate move.

2. Related Literature

While our main results are new, we are far from the first to study models of deliberate risk or noise. The cheap talk literature beginning with Crawford and Sobel (1982) can be thought of as a leading example of noisy communication. In that literature, nothing binds the sender. In contrast, as explained above, our chosen communication structures have mean equal to the true state, the choice could be costly, and it is central that each individual chooses a *distribution* over signals, and cannot hide the outcome ex post.

The choice of an information structure is central to Kamenica and Gentzkow (2011). But no agent knows the true type ex-ante, and the chosen information *structure* is fully observed by the receiver. This last feature — an observed information structure — is shared by Degan and Li (2016), but the type of the agent is privately known, as in our model. In contrast, in our setting, the choice of information structure is not (fully) observed, only the signal, a feature that we share with DeMarzo et al. (2019). We return to the question of observability (and these references), first in Section 5.4, and then again in Section 7.

---

1 In this brief review we omit discussion of a related but distinct literature with *exogenous* noise, as in the limit pricing game studied by Matthews and Mirman (1983), the choice of mean return by managers of unknown quality who might seek to herd (Zwiebel, 1995), or inference settings when values have exogenous but unknown precision (Subramanyam, 1996).
Dewan and Myatt (2008) examine a model of leadership in which an individual’s clarity in communication is a virtue, but the leader also wishes to hold on to an audience for longer, to dissuade them from listening to others. Therefore extreme clarity is not chosen. Edmond (2013) also studies the obfuscation of states (say by a dictatorial regime), but restricts attention in his analysis (by assumption) to receiver-actions that are monotone in the signal realization. In contrast, in our setting, the non-monotonicity of receiver actions is a fundamental and robust outcome of the model.

Finally, Hvide (2002) studies tournaments with moral hazard where competitors can add noise to their actions; see also Palomino and Prat (2003) and Barron et al. (2017) who also study situations in which agents can inject noise into a moral hazard setting. Our adverse-selection setting is entirely distinct but shares the same feature of endogenous noise.\(^2\)

### 3. A Baseline Model

#### 3.1. Setting.

An agent works for a principal. The agent can be good (\(g\)) or bad (\(b\)). He knows his type. The principal doesn’t, but has a prior \(q \in (0, 1)\) that the agent is good. At the end of a single round of interaction, to be described below, the principal decides whether or not to retain the agent. Retention of an agent of type \(k = g, b\) yields an expected payoff of \(U_k\) to the principal, with \(U_g > U_b\). Non-retention yields the principal \(V \in (U_b, U_g)\). The type-\(k\) agent gets a payoff equal to 1 if he is retained and 0 otherwise. The agent therefore prefers to be retained regardless of type, while the principal prefers to retain the good agent.

The principal receives a signal from the agent, which is indicative of his type. Based on the realization of that signal, the principal decides whether or not to retain. The agent has some control over the distribution of this signal, but conditional on this, cannot alter the signal realization. Specifically, suppose that the signal is given by

\[
x = \theta_k + \sigma_k \epsilon,
\]

for \(k = g, b\), where \(\theta_k\) is a type-specific mean with \(\theta_g > \theta_b\), \(\epsilon \sim N(0, 1)\) is zero-mean normal noise, and \(\sigma_k\) is a term that scales the noise, that is chosen by the agent. That is, the agent cannot shift the mean of his signal,\(^3\) but he can modulate its precision. The principal does not observe \(\sigma_k\), but she observes the realization of the signal.

Our baseline setting assumes that the choice of noise is costless but bounded below: \(\sigma_k \geq \underline{\sigma}\) for some \(\underline{\sigma} > 0\). (We will add costly noise in Section 6.1.) Of course, a condition such as this is a minimal requirement for the problem to have any interest: otherwise, the high type can always reveal himself by choosing \(\sigma_g = 0\), and there is nothing to discuss. That said, we will think of \(\sigma\) as “small” (see below). Define \(p \in (0, 1)\) by

\[
pU_g + (1-p)U_b \equiv V;
\]

\(^2\)There is also a literature on policy uncertainty (see, for example, Shepsle, 1972; Campbell, 1983; Alesina and Cukierman, 1990; Glazer, 1990; Aragones and Neeman, 2000; Aragones and Postlewaite, 2002; Aragones, Palfrey and Postlewaite, 2007), where candidates offer deliberately ambiguous policy platforms, which generate uncertainty about the policies the candidate could implement were she to win the election.

\(^3\)For an extension to costly mean-shifting, see the Supplementary Appendix.
then $p$ is interpretable as an “outside option probability” that leaves the principal indifferent between retaining and replacing. A salient benchmark is $p = q$ (the balanced model). But if $V$ incorporates the option value of dealing with a new agent in a dynamic context, $p$ could exceed $q$ (see Section 6.2 on dynamics). This is a model with an optimistic future. Or if our current agent is an ongoing hire about whom some (positive) information has already been received, then $p$ could be smaller than $q$; call this a model with a pessimistic future.

3.2. Equilibrium. The principal observes a realized outcome $x$ from $N(\theta_k, \sigma^2_k)$, and uses Bayes’ Rule to retain the agent if (and modulo indifference, only if)

$$\Pr(k = g|x) = \frac{q \frac{1}{\sigma_g} \phi \left( \frac{x - \theta_g}{\sigma_g} \right)}{q \frac{1}{\sigma_g} \phi \left( \frac{x - \theta_g}{\sigma_g} \right) + (1 - q) \frac{1}{\sigma_b} \phi \left( \frac{x - \theta_b}{\sigma_b} \right)} \geq p,$$

where $\phi$ is the pdf of the standard normal. Rearranging, we have retention if and only if

$$\frac{\frac{1}{\sigma_b} \phi \left( \frac{x - \theta_b}{\sigma_b} \right)}{\frac{1}{\sigma_g} \phi \left( \frac{x - \theta_g}{\sigma_g} \right)} \leq \frac{1 - p}{p} \frac{q}{1 - q} =: \beta \in \mathbb{R}_+.$$

Simple algebra involving the normal density yields the equivalent expression

$$\left(\sigma^2_g - \sigma^2_b\right) x^2 + 2 \left(\sigma^2_g \theta_g - \sigma^2_b \theta_b\right) x + \left(\sigma^2_g \theta^2_g - \sigma^2_b \theta^2_b + 2A \sigma^2_g \sigma^2_b\right) \geq 0,$$

where $A := \ln \left(\frac{\beta \sigma_b}{\sigma_g}\right)$. The inequality (4) defines a retention regime, a zone $X$ of signals for which the principal will want to retain the agent. An equilibrium is a configuration $(\sigma_g, \sigma_b, X)$ such that given $(\sigma_g, \sigma_b)$, $X$ is the set of “retention signals” $x$ which solve (4), and given $X$, each type $k$ chooses $\sigma_k$ to maximize the probability of retention; that is,

$$\sigma_k \in \arg\max_{\sigma \geq 2} \int_X \frac{1}{\sigma} \phi \left( \frac{x - \theta_k}{\sigma} \right) dx.$$

3.3. A Remark on Interpretation. Principal-agent models are typically concerned with situations in which an agent takes an action that affects the payoff of the principal. That action could also influence the principal’s retention decision in a dynamic setting; see, for instance, Dutta et al. (1989) for a situation with agent moral hazard, and Banks and Sundaram (1998) for a situation with agent adverse selection. In the model we consider, the incentives to elicit current effort have been deliberately muted, so as to concentrate on the retention decision alone, and dynamic effects have also been suppressed through the device of an outside option. Our objective is to emphasize the choice of noise as a way of disguising one’s type, or revealing it to the extent possible.

That said, it may be useful to keep the following structure in mind. Agent-generated signals today are both signals and outcomes. The principal’s payoff depends on these outcomes or signals, and she is risk-neutral. So the chosen distribution of signals — or outcomes — is of no direct payoff-consequence to her; only the mean matters. Therefore her retention decision is entirely based on her update following the signal — she cares only about the mean outcome to be generated following retention. Our model fits this setting precisely.
This exact interpretation may be disturbed by a number of factors. First, the principal’s risk attitudes could matter. But in our analysis, the high type chooses a signal with lower variance, so our results extend to all risk-averse principals. Second, the agent could shift the mean outcome (relative to their “natural” type) by expending effort. This feature is easy to incorporate, as we show one in of the extensions in the Supplementary Appendix. Third, there could be dynamic considerations that overturn or unduly complicate the baseline reasoning. We discuss such an extension in Section 6.2. It fully supports our static arguments, and in a sense that we will make clear, even simplifies the statement of the results.

In summary, our model emphasizes retention, and the consequent incentives for agents to hide or reveal their types by injecting noise into their actions. Of course those actions could also have payoff consequences. But nothing we write is inconsistent with that fact.

4. Retention Regimes

4.1. Trivial Retention Regimes. Two examples of retention zones are (a) “always retain,” so that \( X = \mathbb{R} \), and (b) “always replace,” that is, \( X = \emptyset \). Both generate complete indifference across the two types as to the noise regime. With any cost function for noise that is minimized at some common value for both types, we then have \( \sigma_g = \sigma_b \), but then (4) must alter sign over \( x \), a contradiction (Section 6.1 shows this explicitly). Even without any cost of noise, these equilibria are eliminated in a dynamic setting (Section 6.2). So in the benchmark model, we ignore such trivial and delicately supported regimes.

4.2. Monotone Retention Regimes. An equilibrium regime is monotone if there is a finite threshold \( x^* \) such that the principal replaces the agent for signals on one side of \( x^* \), and retains him for signals to the other side of \( x^* \). See Figure 1. A monotone retention regime arises (and can only arise) when both types transmit with the same noise \( \sigma_g = \sigma_b = \sigma \). Then (4) reduces to the condition

\[
 x \geq x^*(\sigma) := \frac{\theta_g + \theta_b}{2} - \frac{\sigma^2}{\theta_g - \theta_b} \ln(\beta) .
\]

Loosely, \( x^*(\sigma) \) is the threshold above which the principal deduces that a signal from two possible noisy sources of equal variance is more likely to be coming from the higher-mean source. This is the exact

\footnote{If \( \sigma_g \neq \sigma_b \), then by condition (4), the resulting retention regime is either trivial or non-monotone.}
interpretation of \( x^*(\sigma) \) in the balanced model, for then \( \beta = 1 \) and

\[
x^*(\sigma) = \frac{\theta_g + \theta_b}{2},
\]

which is the mid-point between the two means. When \( p \neq q \), retention is not simply dependent on relative likelihoods, but also on how pessimistic or optimistic the principal feels about future agents, which is measured by the ratio of \( q \) to \( p \), as proxied by \( \beta \) in (5).

That said, consider any monotone equilibrium with threshold between \( \theta_b \) and \( \theta_g \). Then the good type will want to minimize the noise of his signal, while the bad type will want to maximize it. But this immediately destroys the putative equilibrium: when the bad type chooses higher noise than the good type, there cannot be a single threshold for retention. Good news — but only moderately good news — offer the best likelihood ratios in favor of the good type, and will generate retention. But a high “good signal” will be regarded as too good to be true: for those signals, the higher chosen variance of the bad type will dominate the lower mean, leading to a higher likelihood that the signal was emitted by the bad type. The formal analysis in the rest of this section, and in Section 6.1, extends these arguments to all single-threshold equilibria, arguing that if the minimal noise level \( \sigma \) is small enough or if there is a cost of noise, no monotone equilibrium can exist, whether the retention threshold lies between the means of the two types, or to one side of these.

4.3. Non-Monotone Retention Regimes. When different types transmit at different noises, the corresponding best response for the principal is never monotone. Figure 2 illustrates this (for the balanced case). In Panel A, \( \sigma_b > \sigma_g \), and in Panel B, \( \sigma_b < \sigma_g \). In each case, the signal densities cross precisely twice. In Panel A, the principal retains the agent for all signals in between the two intersections, and in Panel B, she does so for all signals not in between those intersections. We make these observations more formal in Observation 1 below. The general point is that one of the two zones must be defined by a bounded zone of signals. It is convenient to use the notation \([x_-, x_+]\) to denote the relevant interval when bounded retention occurs, and by \([x_+, x_-]\) to denote the interval when bounded replacement occurs. Obviously,
\(x_+\) and \(x_-\) are the two roots of (4), which means that

\[
(6) \quad \beta \frac{1}{\sigma_g} \phi \left( \frac{x - \theta_g}{\sigma_g} \right) = \frac{1}{\sigma_b} \phi \left( \frac{x - \theta_b}{\sigma_b} \right)
\]

represents the equalization of weighted likelihoods for both types at \(x = x_-, x_+\). Furthermore, the weighted likelihood for the good type must have a higher slope in \(x\) relative to that for the bad type, evaluated at \(x_-\), so that retention occurs to the right of \(x_-\). That means

\[
\beta \frac{1}{\sigma_g} \phi' \left( \frac{x_+ - \theta_g}{\sigma_g} \right) > \frac{1}{\sigma_b} \phi' \left( \frac{x_- - \theta_b}{\sigma_b} \right).
\]

Because \(\phi(z) = (1/ \sqrt{2\pi}) \exp[-z^2/2]\) satisfies \(\phi'(z) = -z\phi(z)\), this is equivalent to:

\[
(7) \quad \beta \phi \left( \frac{x_- - \theta_g}{\sigma_g} \right) \frac{x_- - \theta_g}{\sigma^3_g} - \phi \left( \frac{x_- - \theta_b}{\sigma_b} \right) \frac{x_- - \theta_b}{\sigma^3_b} < 0.
\]

Exactly the opposite slope condition holds at \(x_+\), so that

\[
(8) \quad \beta \phi \left( \frac{x_+ - \theta_g}{\sigma_g} \right) \frac{x_+ - \theta_g}{\sigma^3_g} - \phi \left( \frac{x_+ - \theta_b}{\sigma_b} \right) \frac{x_+ - \theta_b}{\sigma^3_b} > 0.
\]

Use (6) for \(x = x_-\) in equation (7) to obtain

\[
\left( \sigma^2_b - \sigma^2_g \right) x_- < \sigma^2_g \theta_g - \sigma^2_g \theta_b.
\]

In the same way, use (6) for \(x = x_+\) in equation (8) to see that

\[
\left( \sigma^2_b - \sigma^2_g \right) x_+ > \sigma^2_g \theta_g - \sigma^2_g \theta_b.
\]

Combining these two inequalities, we must conclude that

\[
(9) \quad \left( \sigma^2_b - \sigma^2_g \right) (x_+ - x_-) > 0
\]

in any non-monotonic equilibrium. We summarize the above discussion as:

**Observation 1.** Bounded retention with \(x_+ > x_-\) is associated with \(\sigma_b > \sigma_g\), while bounded replacement with \(x_- > x_+\) is associated with \(\sigma_b < \sigma_g\).

### 5. Bounded Retention Equilibrium

Our main result, that we extend in several directions, is that there is a unique nontrivial equilibrium if the ambient noise \(\sigma\) is small, and in it the principal uses a bounded retention zone. She is suspicious of both bad signals and excessively good signals, and follows the maxim: “If it seems too good to be true, it probably is.” In what follows we emphasize both our results and make explicit some qualifications.

#### 5.1. No Bounded Replacement

First, we eliminate bounded replacement equilibria. In such an equilibrium the principal replaces the agent when the signal falls inside \([x_-, x_+\], with \(x_+ < x_-\). Observation 1 tells us that this regime is associated with \(\sigma_b < \sigma_g\). It is easy to see that, in this case, the retention probability of any type converges to 1 as \(\sigma_k \to \infty\), and therefore in equilibrium, it can never be that \(\sigma_b < \sigma_g\). We remark that this argument is straightforward only in our baseline model where any level
of noise above some minimum can be costlessly chosen. We will need to revisit it when the choice of noise is costly.

5.2. Small Ambient Noise. Recall that \( \beta = \frac{1-p}{p} \times \frac{q}{1-q} \). For \( \beta \in (0,1) \) (that is, for \( p > q \) or an optimistic future), define \( \alpha(\beta) \) by the unique solution to

\[
\beta \equiv \frac{1}{\alpha(\beta) + \sqrt{1 + \alpha(\beta)^2}} \exp \left[ -\frac{\alpha(\beta)}{\alpha(\beta) + \sqrt{1 + \alpha(\beta)^2}} \right].
\]

Notice that \( \alpha(\beta) \) is well-defined, that \( \alpha(\beta) > 0 \) for all \( \beta \in (0,1) \) and \( \alpha(\beta) \to 0 \) as \( \beta \to 1 \). We will assume that \( \sigma \) is small enough so that:

\[
\frac{\sigma}{\theta_g - \theta_b} < \frac{1}{2} \alpha(\beta)^{-1}.
\]

On the other hand, with a pessimistic future or \( \beta > 1 \), we will use this bound:

\[
\frac{\sigma}{\theta_g - \theta_b} < \left[ \sqrt{2 \ln(\beta)} \right]^{-1}.
\]

These bounds progressively weaken as we move to the balanced case. At or near the balanced case, no restrictions are imposed at all; both right-hand side terms in (11) and (12) diverge to infinity. Moreover, in Section 6.2, which studies a dynamic extension with fixed term limits for agents, we show that these restrictions on \( \sigma \) will be automatically satisfied.

5.3. The Salience of Bounded Retention. Our baseline result can now be stated:

**Proposition 1.** (i) With an optimistic future (\( \beta < 1 \)), a nontrivial equilibrium exists if and only if (11) holds. When it exists, it is unique, and has bounded retention.

(ii) With a pessimistic future (\( \beta \geq 1 \)), a unique nontrivial equilibrium exists. This equilibrium has bounded retention if (12) holds. Otherwise it has monotonic retention.

(iii) In a nontrivial equilibrium with bounded retention, the good type chooses \( \sigma_g = \sigma \), the bad type chooses higher but finite noise \( \sigma_b > \sigma_g \), and the principal employs a strategy of the form: retain if and only if the signal \( x \) lies in some bounded interval \( [x_-, x_+] \).

The proposition distinguishes between two cases. In the first case, the future is optimistic. Equivalently, there is a low prior that our current agent is good, possibly due to unsatisfactory past performance (not modeled here). So a bad type is quite desperate to imitate his good counterpart, and takes on more noise than does the good type, who is equally desperate to reveal himself by reducing noise. So \( \sigma_b > \sigma_g \), which serves to uniquely precipitate the bounded retention equilibrium, whenever an equilibrium exists.

In the second case, the future is pessimistic. That is, our current agent is doing well relative to the market, so the bad type can afford to take less risk. If both types minimize risk that resulting equilibrium will involve monotone retention; this is part (ii) when (12) fails. However, the incentive for the bad type to take on more noise than the good type is never entirely absent, and it will invariably appear provided the minimal feasible noise is low enough. This is at the heart of the argument: even in (ii), if the minimal
feasible noise falls below the bound described by (12), then the bad type will never want to follow the
good type all the way to minimum noise, and no monotone equilibrium can exist.

How permissive is the bound in (12)? One way to view it is to study the perfectly balanced case in which
there is neither pessimism nor optimism. Then the right-hand side of (12) is infinitely large, and we can
unequivocally assert that the unique nontrivial equilibrium involves bounded retention. Now move
away from the balanced case by placing greater faith in the current agent, so that (12) begins to bite.
Suppose that an agent is equally likely to be good or bad, but our agent is 75% likely to be good. Then
(12) implies that the unique equilibrium involves bounded retention as long as the standard deviation
of the signal can be brought below approximately 2/3 the difference between the two means.

Two extensions continue to underscore the salience of bounded retention. In Section 6.1, we replace
the costless choice of noise by a cost function. That effectively compactifies the space of noises, but the
cost function is smooth and not “L-shaped” as in this, our baseline model. Proposition 2 proves that
a monotone equilibrium generically cannot exist. In fact, this is a more uncompromising prediction
than the one made by our baseline model, at least as far as the impossibility of monotone equilibrium
is concerned.

Second, in Section 6.2, we describe a dynamic model in which \( \beta \) drifts over time in line with Bayes’
Rule. Specifically, we consider an infinite-horizon setting in which each agent faces a two-term limit.
It turns out that conditions (11) and (12) automatically hold in that exercise, and there is no monotone
equilibrium. However, a fuller accounting of the dynamic case is beyond the scope of the present
exercise, and is not pursued in this paper.

In summary, Proposition 1 and our subsequent discussion argue that when the ambient level of noise is
positive but small, we are left with our case of central interest: an equilibrium in which the types choose
different noise levels, the bad noise higher than the good. The principal does not use a “one-sided”
retention strategy. She looks for good signals to retain the agent, but distrusts signals that are extremely
positive, because she suspects that bad types are injecting noise into the system, and the good types are
not. That suspicion will justifiably yield a bounded retention zone, because far enough out, the higher
variance of the bad-type signal will dominate the lower mean in determining relative likelihoods.\(^5\)

In this setting, a basic single-crossing property is missing. Yet the model itself is tractable. Specifically,
low types choose a larger variance in a bid to convince the sender that she is of higher mean. But that
also enhances the relative likelihood for the low type under extreme signal realizations. The argument
is delicate — and therefore complex — because the sender understands the previous sentence, and so
dislikes such realizations. Nonetheless, the low type continues to choose higher noise in equilibrium.

5.4. A Remark on the Unobservability of Noise. Our result depends on the assumption that the choice of
agent noise is not observable. Consider the opposite presumption that the noise is observable. Then
every type must choose the same noise; i.e., separation is impossible via the choice itself — the bad type
will deviate by mimicking this choice. So, with observable noise, agent types must choose the same

\(^5\)Of course, the principal also distrusts signals that are bad: after all, lower mean and higher variance are particularly
synergistic in producing lower signals.
noise, and then the principal will use a monotone retention rule to retain or replace the agent, retaining if the signal realization is good enough (see Degan and Li, 2016\textsuperscript{6}). If risk choices are costly, as they will be in Section 6.1, the same argument applies as long as the cost function for noise is the same for both types — again, there must be pooling in observed components.\textsuperscript{7}

It is easy to augment this variation with unobserved noise, resulting in exactly our model.\textsuperscript{8} Observed noise must be chosen identically by all types. The exact value is immaterial. (If noise is costly, it would be set to the minimum-cost level in any equilibrium refined by intuitive off-path restrictions on beliefs.) The remainder of the analysis then proceeds with no change. To summarize: (a) complete lack of observability is not needed for our results; (b) there will be pooling on the observable components if the choice of risk is costless or uncorrelated with agent type; and (c) our results then apply to the unobserved components of risk. The singular nature of Proposition 1 is rooted in the presumption that there is some unobserved component of the signal structure, not that the entire structure is unobservable. It is important to appreciate this for the applications in Section 7.

6. Extensions

Section 6.1 introduces costly noise. Section 6.2 analyzes a dynamic version with agent term limits. Section 6.3 drops the normality assumption. Section 6.4 briefly describes other extensions, covered in depth in the Supplementary Appendix.

6.1. Costly Noise. Suppose there is a cost to modulating precision $\sigma$. Specifically, consider a strictly convex cost function $c(\sigma)$, with a minimum at some $\sigma_\ast$, with $c(\sigma_\ast) = 0$, and $c(0) = c(\infty) = \infty$. While we use the same notation, $\sigma$ is no longer the minimum possible variance but just some ambient noise that reflects the usual frequency of communication glitches, errors of perception, and so on. Deviations from this ambient noise are costly in either direction. That is, it is costly both to fully reveal one’s type, or to fully hide it. An equilibrium is a configuration $(\sigma_g, \sigma_b, X)$ such that given $(\sigma_g, \sigma_b), x \in X$ solves (4), and given $X$, each type $k$ chooses $\sigma_k$ to maximize the probability of retention, net of cost:

$$\sigma_k \in \arg \max_{\sigma \geq \sigma_\ast} \left[ \int_X \frac{1}{\sigma} \phi \left( \frac{x - \theta_k}{\sigma} \right) dx - c(\sigma) \right].$$

This version of the model presents some new features. First, trivial equilibria never exist. If the retention regime is trivial, both types must choose the lowest cost signal, which in turn makes the signal

\textsuperscript{6}For a related exercise, see Titman and Trueman (1986), in which observed auditor quality is used to signal firm valuation during an initial public offering. (Higher-quality auditors provide more precise information, by assumption.) An entrepreneur with more favorable private information about the value of his firm will choose a higher-quality auditor than will an entrepreneur with less favorable private information.

\textsuperscript{7}If the cost function for risk choices is systematically connected with agent type, then there may be separation achieved via costly signaling using observable components. In this case the fact that the action set is a choice of risk is of no separate importance. It is just one of many abstract ways to achieve separation.

\textsuperscript{8}In an otherwise different setting, DeMarzo et al. (2019) also work with the choice of an unobserved information structure — correctly guessed at in equilibrium. A sender gathers information about the quality of an object by selecting a test, which might return a null result. He can choose to disclose (verifiable) information, but also to suppress test results. The receiver understands this, so there is no value in observably choosing a test. But there is some value to choosing an unobserved test — the authors describe this optimal choice.
informative, a contradiction. Second, as we shall see, monotone regimes are generically impossible in equilibrium. But third, the model does opens up the possibility of the existence of bounded replacement equilibria, which were easily ruled out in the benchmark model. Now we expand on some of these points, beginning with the agent’s best response mapping before moving to a fuller description of equilibrium.

The Agent’s Best Response. We already know that nontrivial equilibria are either monotone or have interval cutoffs. Type \( k \) chooses \( \sigma_k \) to maximize \( \Phi \left( \frac{x_+ - \theta_k}{\sigma_k} \right) - \Phi \left( \frac{x_- - \theta_k}{\sigma_k} \right) - c(\sigma_k) \). In a monotone regime, \( x_+ = \pm \infty \) and \( x_- = x^* \). First-order conditions are

\[
(13) \quad \phi \left( \frac{x_+ - \theta_k}{\sigma_k^2} \right) \left( \frac{x_- - \theta_k}{\sigma_k^2} \right) - \phi \left( \frac{x_+ - \theta_k}{\sigma_k^2} \right) \left( \frac{x_- - \theta_k}{\sigma_k^2} \right) = c' \left( \sigma_k \right)
\]

for each type \( k = g, b \). Optimally chosen noise now moves in a subtle and quite complicated way with the location of a player’s type. Figure 3, Panel A, illustrates this for a monotone retention threshold. When a player’s type is far from the retention threshold, it takes large noise to generate (with any significant probability) a signal within the retention zone. That’s costly, so noise converges to the zero-cost choice \( \sigma \) as the type moves far from the retention zone. Moving closer to the zone, chosen noise increases, but must decline again: after all, when the type is on the edge of the zone, noise makes no difference to the chances of retention, so chosen noise is back to \( \sigma \) again. As the type moves into the retention zone, noise can only throw him out it, so chosen noise now falls below \( \sigma \). But the downward movement does not continue forever. Deep in the retention zone, the type is confident of remaining there, and so noise goes up again, converging to \( \sigma \), this time from below.

With bounded retention zones, the choice function exhibits even more non-monotonicities. Panel B of Figure 3 shows that there will generally be five turning points. There is one each for either side of the retention zone, for the same reason as in the earlier discussion. There are three more within the retention zone: noise initially falls as an agent with type close to the edge avoids escape from the zone; then rises in the middle of the zone as the risk of escape falls, then falls again as the risk goes up, and finally rises as we approach the edge. (The noise choice at the edges is below \( \sigma \), because the retention zone is bounded.)

\[\text{Figure 3. How Choice of Noise Varies With Agent Type}\]
This behavior is consistent with empirical findings on risk-taking. Genakos and Pagliero (2012) find that risk-taking in weightlifting contests exhibits an inverted-U relationship between risk and rank, with the peak reached around rank 6. Figueiredo et al. (2015) observe that risk-taking by portfolio managers is non-monotonic: managers significantly below a compensation threshold reduce risk-taking relative to those who are relatively close. These findings are consistent with our predictions when agents are to the left of the retention threshold (Panel A in Figure 3). With costly noise, both monotone and bounded retention can generate this type of behavior. That said, monotone retention cannot be an equilibrium outcome of this model, except in degenerate cases.

No Monotone Retention.

Proposition 2. Generically, a monotone equilibrium can not exist in the costly noise model. Specifically, there is at most one value of \( \sigma \) that both types must choose in any monotone equilibrium, but this value is determined independently of the noise cost function.

For some intuition, consider any single retention threshold \( x^* \) as in Figure 1, produced by some common value \( \sigma_g = \sigma_b = \sigma \). The first-order condition (13) becomes

\[
\phi \left( \frac{x^* - \theta_k}{\sigma_k} \right) \left( \frac{x^* - \theta_k}{\sigma_k^2} - c' (\sigma_k) \right) = 0,
\]

where, in equilibrium, \( x^* \) is given by (5). Setting \( \sigma_g = \sigma_b = \sigma \), we can see that the two first-order conditions cannot hold simultaneously when \( x^* \in (\theta_b, \theta_g) \). But it’s possible that both types lie on the same side of the threshold. Defining \( \Delta := \theta_g - \theta_b \), we can rewrite the first-order condition for good and bad types as

\[
\phi \left( \frac{\sigma}{\Delta} \ln (\beta) + \frac{\Delta}{2\sigma} \right) \left( \frac{\sigma}{\Delta} \ln (\beta) + \frac{\Delta}{2\sigma} \right) = \phi \left( \frac{\sigma}{\Delta} \ln (\beta) - \frac{\Delta}{2\sigma} \right) \left( \frac{\sigma}{\Delta} \ln (\beta) - \frac{\Delta}{2\sigma} \right) = -\sigma c'(\sigma).
\]

Equation (15) tells us to study \( \phi(z)z \); Figure 4 does so. Denote \( \frac{\sigma}{\Delta} \ln (\beta) - \frac{\Delta}{2\sigma} \) by \( z_1 \) and \( \frac{\sigma}{\Delta} \ln (\beta) + \frac{\Delta}{2\sigma} \) by \( z_2 \). Given the shape of \( \phi(z)z \), Figure 4 indicates how \( z_1 \) and \( z_2 \) must be located: they must both have the same sign and the same “height.” With an optimistic future (\( \ln \beta < 0 \)), both \( z_1 \) and \( z_2 \) are negative; see Panel A. With a pessimistic future, both \( z_1 \) and \( z_2 \) are positive as in Panel B. In each case, there is only
one value of $\sigma$ that can solve this requirement; i.e., just one value that fits the first equality in (15). It is entirely independent of the cost function for noise, and so the second equality generically can not hold.

This contrasts with the case of costless noise, where corner responses are possible at lower bounds of noise (or upper bounds, if those exist), thereby permitting monotone thresholds.

*Bounded Retention and Replacement Equilibria.* We are left with equilibria in which the principal employs bounded intervals for retention or replacement. To analyze these, we sidestep a technical complication that now arises. The noise distribution generates nonconvexities in the agent’s optimization problem, which raises the possibility that an agent’s choice could be multi-valued. These can be possibly be handled using mixtures, but for monotone or bounded retention regimes, such multi-valuedness is more a technical nuisance than a feature of any economic import, and we rule it out by assumption:

[U] For every monotone or bounded retention zone and for each agent type, the optimal choice of noise is unique.\(^{10}\)

It is possible to deduce [U] by placing alternative primitive restrictions on the parameters of the model. One is that the curvature of the cost function is large enough. The Supplementary Appendix shows that a sufficient condition for [U] is

\[
 c''(\sigma) > \frac{K}{\sigma^2} \text{ for all } \sigma \in [\sigma_*, \sigma^*],
\]

where $K \approx 0.6626$, and $\sigma_*$ and $\sigma^*$ are two distinct lower and upper bounds on noise that straddle $\varrho$, such that $c(\sigma_*) = c(\sigma^*) = 1$.

Next, recall $\sigma_*$ and $\sigma^*$ from (16). These are lower and upper bounds on noise that straddle $\varrho$, with $c(\sigma_*) = c(\sigma^*) = 1$. No agent would ever transmit noise outside $[\sigma_*, \sigma^*]$. Suppose both types transmit common noise equal to $\sigma^*$; then the principal responds by choosing a single threshold $x^*(\sigma^*)$ for retention, as in equation (5). We impose

[T] The threshold $x^*(\sigma^*)$ lies in $[\theta_b, \theta_g]$.

Condition T is automatically satisfied in the balanced case: there, $x^*(\sigma^*) = (\theta_g + \theta_b)/2$. Indeed, [T] can be viewed as a restriction on the extent to which $\beta$ can depart from 1 on either side of the balanced case. Subtract the formula for $x^*(\sigma^*)$ — see (5) — from $\theta_b$ and then $\theta_g$ to obtain an equivalent form of [T]:

\[
-\frac{(\theta_g - \theta_b)^2}{2\sigma^2} \leq \ln(\beta) \leq \frac{(\theta_g - \theta_b)^2}{2\sigma^2}.
\]

**Proposition 3.** Under Conditions U and T, there is an equilibrium with bounded retention.

The proof has economic intuition, so we outline it. The first box in Figure 5 shows the domain of a suitable fixed-point mapping, with agent noise lying between $\sigma_*$ and $\sigma^*$. The mapping is derived as follows: for each $(\sigma_g, \sigma_b)$, find the principal’s retention decision, shown in the middle graph ($x_-$ and $x_+$), and then record the best response to that decision, shown by the continuation mapping into the last box, a replica of the one we started from.

\(^{10}\)With bounded replacement, multiple responses are more compelling. An agent located in one of the two retention zones, but close to the replacement zone, could be indifferent between small and large noise.
The problem is that this fixed point mapping is not well-behaved. For any $\sigma_b < \sigma_g$, the principal best-responds with bounded replacement, and the “subsequent” response that completes the mapping is generally discontinuous in $(\sigma_g, \sigma_b)$. This problem is endemic. However, to probe the existence of a bounded retention equilibrium, we can start from a smaller domain: the shaded triangle over which $\sigma_b \geq \sigma_g$. On this subdomain, the principal chooses bounded retention (or a monotone threshold), and the subsequent best response by the agents is unique (by Condition U) and continuous. But it may be that the mapping slips out of the smaller domain (see lower pair of arrows in Figure 5).

Our assumed condition (17) guarantees that this cannot happen. To see why, study the upper pair of arrows in Figure 5. The first arrow maps a point on the principal diagonal (where $\sigma_b = \sigma_g$) to a monotone retention regime; that is, $(x_-, x_+)$ is of the form $(x', \infty)$. By (17), $x'$ must lie between $\theta_b$ and $\theta_g$. In response, the good type will want to reduce noise as much as possible, while the bad type will want to increase it. Therefore $\sigma_g < \sigma$, while the opposite is true of the bad type. But that implies $\sigma_b > \sigma_g$, which takes us back into the starting subdomain from its boundary. A fixed point theorem due to Halpern and Bergman (1968) then completes the argument, establishing the existence of a bounded retention equilibrium when $\beta$ does not take on “extreme” values.

In summary, when the future is neither too optimistic nor too pessimistic — and certainly when it is balanced — a bounded retention equilibrium must exist. Indeed, it could be the only equilibrium. After all, it can be seen that moderate degrees of optimism or pessimism about the future are not only conducive to the existence of a bounded retention equilibrium, they push against the existence of a bounded replacement equilibrium. For instance, assume a sizable difference between the two types; specifically, that

\begin{equation}
\theta_g - \theta_b \geq \sigma^*,
\end{equation}

where recall that $\sigma^*$ is defined by the larger of the two solutions to $c(\sigma) = 1$.

**Proposition 4.** Assume (17) and (18). Then only bounded retention equilibria exist.
The argument emphasizes the location of types relative to replacement and retention zones. When the conditions for Proposition 4 fail, Figure 6 shows how bounded replacement might arise. The density for the bad type is the thicker line in both panels. The figure shows that $\beta$ must be so large or so small (that is, the future is either so optimistic or so pessimistic) that the intersections of the two weighted densities are either on one side of both the mean types, or straddle them both. These are the only two possible kinds of bounded replacement equilibria. For completeness, the Appendix provides examples in each case.\footnote{In the first kind of bounded replacement equilibrium, both types are in the retention zone as in Panel A of Figure 6, with $x_+ < x_\ast < \theta_b < \theta_g$. Because they want to remain there, both want noise lower than the ambient level. But the bad type is closer to the edge, so he will make a bigger effort than the good type to stay safe, and $\sigma_b < \sigma_g$. To justify this configuration as an equilibrium, the future must be super-pessimistic: $q \gg p$. In the second case, shown in Panel B of Figure 6, both $\theta_b$ and $\theta_g$ lie in the replacement zone, with $x_+ < \theta_b < \theta_g < x_\ast$, and both exert costly effort to escape it. The good type is embedded closer to the edge of the zone and has a high marginal benefit of noise, while the bad type is embedded deep in the zone and has only a low marginal benefit. The good type therefore exerts greater noise. The principal reacts by choosing a bounded replacement zone. To implement this equilibrium, the future must be super-optimistic: $p \gg q$.}

6.2. Dynamics With Term Limits. We solve for the “outside option probability” $p$ in a dynamic setting. We assume that the agent has a two-period “term limit.” We consider stationary equilibrium, in which every new agent of a given type takes the same action for the same value of $p$. Given $\sigma_k$ for each type, and a realization $x$, the update on $q$ is

$$q(x) := \frac{q\pi_g(x)}{\pi(x)},$$

where for each $k$, the density of signal $x$ is given by $\pi_k(x) = (1/\sigma_k)\phi([x-\theta_k]/\sigma_k)$, and where $\pi(x) = q\pi_g(x) + (1-q)\pi_b(x)$ is the overall density of signal $x$.

Let $M(q') := q'U_g + (1-q')U_b$ be the expected payoff to the principal in any period when her prior is given by $q'$. (This equals $q$ for a fresh draw from the pool.) At the end of term 1, a signal $x$ is generated, and $q$ is updated to $q(x)$. If $V$ denotes the lifetime payoff to the principal starting from a fresh agent, the retention zone $X$ is the set of all $x$ for which $(1-\delta)M(q(x)) + \delta V \geq V$. Let $\Pi_k := \int_x \pi_k(x)dx$ be the type-dependent probability of retention, and $\Pi := q\Pi_g + (1-q)\Pi_b$ the overall probability of retention.
Then
\[
V = (1 - \delta)M(q) + \delta \int_x [(1 - \delta)M(q(x)) + \delta V] \pi(x)dx + \delta \int_x V \pi(x)dx
\]
\[
= (1 - \delta) \left[ q(1 + \delta \Pi_g)U_g + (1 - q)(1 + \delta \Pi_b)U_b \right] + \delta [1 - (1 - \delta)\Pi] V.
\]

Transposing terms, we see that \(V\) is a convex combination of baseline utilities \(U_g\) and \(U_b\); i.e., \(V = pU_g + (1 - p)U_b\), where
\[
p = \frac{q \left( 1 + \delta \Pi_g \right)}{1 + \delta \left[ q \Pi_g + (1 - q) \Pi_b \right]}.
\]

We can rewrite this expression to obtain a “general equilibrium formula” for the ratio \(\beta\):

\[
(20) \beta = \frac{q}{1 - q} \frac{1 - p}{p} = \frac{1 + \delta \Pi_b}{1 + \delta \Pi_g}.
\]

Now observe that in any equilibrium, \(\Pi_g \geq \Pi_b\), because the principal will choose a retention zone that retains the high type at least as often than the low type. Indeed, \(\beta\) cannot even equal 1 in any equilibrium.\(^{12}\) Additionally, (20) shows how to solve the two-term dynamic extension of our model. For some value of \(\beta\), solve the equilibrium in the baseline model. That equilibrium generates retention probabilities \(\Pi_g\) and \(\Pi_b\). The circle is then closed by the additional condition that \((\beta, \Pi_g, \Pi_b)\) must solve (20). Formally:

**Proposition 5.** When agents can be hired for up to two terms, and the principal can replace agents with a new draw from a stationary pool, there is a unique equilibrium with all the properties of the bounded retention equilibrium identified in Proposition 1. This equilibrium endogenously has an optimistic future, and (11) and (12) do not need to be assumed.

Under a two-term constraint, Proposition 5 eliminates all monotone and trivial equilibria. That said, we note that a full dynamic extension of our model is beyond the scope of this paper, and in a more general version, could display regions on the equilibrium path in which monotone retention is used, along with bounded retention elsewhere.

6.3. Beyond Normal Signals. Suppose that for each type \(k\), the signal \(x\) is given by \(x = \theta_k + \sigma \varepsilon\), where \(\sigma\) is a parameter (“noise”) to be chosen by the agent, subject to \(\sigma \geq \sigma > 0\), and \(\varepsilon\) is distributed according to some differentiable density function \(f\) with support on all of \(\mathbb{R}\). The resulting density for \(x\) is given by:

\[
\tilde{f}(x|k, \sigma) = \frac{1}{\sigma} f \left( \frac{x - \theta_k}{\sigma} \right).
\]

The familiar monotone likelihood ratio property (MLRP) guarantees that when two types transmit with the same noise, higher signals are increasingly likely to be associated with the higher type; that is \(f(z - a)/f(z)\) is increasing in \(z\) whenever \(a > 0\). We assume a stronger version of this, which is automatically satisfied in the normal case, and guarantees a single, finite threshold for retention when

\(^{12}\)Suppose \(\beta = 1\). Then \(p = q\), and we know that in the static model only bounded retention equilibria are possible. But in that situation the principal can strictly discriminate in favor of the good type, since there will always exist two distinct real roots to (4). But now \(\Pi_g > \Pi_b\), which contradicts our starting point that \(\beta = 1\).
both types use the same noise, no matter how optimistic or pessimistic the principal’s prior is regarding agent types:

**Strong MRLP.** \( f(z-a)/f(z) \) is increasing in \( z \) whenever \( a > 0 \), with

\[
(21) \quad \lim_{z \to \infty} \frac{f(z-a)}{f(z)} = \infty \quad \text{and} \quad \lim_{z \to -\infty} \frac{f(z-a)}{f(z)} = 0.
\]

By MLRP, \( f \) is single-peaked; it will be expositionally convenient to place this peak at 0, so \( f'(z) < 0 \) for all \( z > 0 \) and \( f'(z) > 0 \) for all \( z < 0 \). Define \( \sigma(\beta) \) by

\[
(22) \quad \beta f \left( -\frac{\theta_g - \theta_b}{\sigma(\beta)} \right) \equiv f(0),
\]

for all \( \beta > 1 \), and set \( \sigma(\beta) = \infty \) otherwise. This function is well-defined and unique because we place the peak of \( f \) at zero and because \( f(z) \to 0 \) as \( |z| \to \infty \). The Supplementary Appendix establishes the following two-part proposition:

**Proposition 6.** Assume strong MLRP on signal densities. Then:

(i) A bounded replacement equilibrium cannot exist.

(ii) A monotone retention equilibrium exists if and only if \( \sigma \geq \sigma(\beta) \). In particular, monotone retention equilibria fail to exist when \( \sigma \) is low, and never exist when \( \beta \leq 1 \).

Strong MLRP delivers the observation that “spreads dominate means,” which ensures that likelihood ratios for extreme signals move in favor of the type using the higher spread. The boundedness of either retention or replacement zones is an easy consequence. This has two implications. First, bounded replacement equilibria do not exist (part (i) of the Proposition). For if such an equilibrium were to exist, then by “spreads dominate means,” it must be that \( \sigma_b < \sigma_g \). But, then, by deviating to some \( \sigma \neq \sigma_b \), the bad type can assure retention with probability approaching 1 as \( \sigma \to \infty \). This is a profitable deviation.

Second, “spreads dominate means” implies that a monotone equilibrium can only exist if both types choose the same level of noise. Just as in our benchmark model, that can only happen if both types choose \( \sigma \) and the putative retention threshold lies below \( \theta_b \). However, for small \( \sigma \), that cannot happen — the relative likelihood of the bad type at \( x = \theta_b \) is just too high. This rules out monotone equilibria when the lower bound on noise is small; specifically, when the condition identified in Proposition 6 holds.

The question that remains is whether a bounded retention equilibrium exists, and whether (not as central but still of interest) the retention regime has an interval structure. The following proposition provides sufficient conditions.

**Proposition 7.** Consider any signal density satisfying strong MLRP, and with its single peak at 0. Assume that either \( \beta \geq 1 \), or that \( \frac{\partial \ln(f(x))}{\partial x} \) is convex for all \( x > 0 \). Then:

(i) There exists \( \hat{\sigma} > 0 \) such that a nontrivial equilibrium exists if and only if \( \sigma \in (0, \hat{\sigma}) \). When it exists, the equilibrium is unique.
(ii) There exists $\tilde{\sigma} > 0$ such that if $\sigma \in (0, \min\{\tilde{\sigma}, \hat{\sigma}\})$, the nontrivial equilibrium involves bounded retention. In it, the good type chooses $\sigma_g = \sigma$, the bad type chooses higher but finite noise $\sigma_b > \sigma_g$, and the principal employs a strategy of the form: retain if and only if the signal $x$ lies in some bounded interval $[x_-, x_+]$.

(iii) In the balanced case or with an optimistic future ($\beta \leq 1$), the condition $\sigma \in (0, \tilde{\sigma})$ automatically holds, and the nontrivial equilibrium must involve bounded retention.

6.4. Other Extensions. Our baseline model is both rich and tractable, and therefore easy to extend in several other ways. We consider some extensions in the Supplementary Appendix, but mention them here.

Costly shifting of the mean signal. Suppose that each type can exert unobservable effort to shift the mean value of his signal, at a cost. The cost function, viewed as a translation of that mean from true type, is assumed to be the same for both types, and it is nondecreasing and strictly convex in effort. So the good type has an advantage on account of his “initial location.” The environment is otherwise exactly the same as in the baseline setting. The main result is that in any equilibrium we have $\theta_g > \theta_b$, and therefore the choice of noise and principal retention decisions are the same as in the benchmark model.

Noise created by principals. Suppose that agent effort is separately valuable to the principal. Then she may have an interest in choosing ambient noise level $\sigma$. The rest of the game is the same, with the agent expending effort to choose a signal, and the principal retaining or replacing after observing the signal realization. At $\sigma = 0$, when the two agent types are sufficiently separated (in terms of the mean signal values when they both exert no effort), there can be only separating equilibria in which both types exert zero effort. The principal will therefore want to inject noise into the environment. Noise serves here as a commitment device: by making it impossible to perfectly identify the agent type ex post, the principal gives both types a chance to be retained, thus incentivizing them to exert effort.

More than one agent. This accommodates several new scenarios, such as electoral competition. The principal knows that one, and only one of two candidates is good. The other is bad. The agents know their types, and therefore the types of their opponents. Each agent chooses noise as before, and in so doing, sends the principal a signal. The principal decides which agent to retain; her outside option in our baseline model is now replaced by the value of the discarded agent. The retention regimes must be redefined, because the principal now observes two signals. In this setting, a monotone regime is a retention rule in which the principal retains the agent with the higher signal value. In a bounded retention (replacement) regime, the principal keeps the agent whose signal value falls closer to (further away from) an endogenous threshold. If an equilibrium exists, it must have bounded retention.

Several agent types. The distribution of agent types is given by some density $q(\theta)$ on $\mathbb{R}$. The principal obtains a payoff of $u(\theta)$ if she retains an agent of type $\theta$, where $u$ is nondecreasing, bounded, and continuous. The principal’s (exogenous) outside option is $V$, which falls somewhere between the inferior and superior limits of $u(\theta)$. Then, using the costly noise model, we show that monotone equilibria cannot generically exist. This result throws some light on Edmond (2013)’s incisive analysis of information manipulation by dictatorial regimes of unknown strength. When considering the choice
of signal precision, Edmond presumes that monotonic retention regimes are employed by each citizen. Our result suggests that this assumption may not be without loss of generality.

Commitment. The case of principal commitment to retention mechanisms is the subject of a separate paper; see Espinosa et al. (2020). We consider a more general mechanism design problem within a setting studied by Ray et al. (2020). A version of our model, but with principal commitment, emerges as a special case of that framework. The main result is that the principal gains nothing from commitment relative to a model with no commitment, such as the one studied in this paper.

7. Applications

Our theory separates three features: the choice of risk, the signal realization, and subsequent inference and decision by the principal. A central implication is that a signal may be “good” — even in the sense of generating high payoffs for the principal today — while it simultaneously serves as a cautionary indicator for excessive risk-taking by the agent. That may sound contradictory, but as long as we properly separate the current payoff-relevance of a signal realization from its role qua signal, there is no inconsistency here.

The potential relevance of our model should be assessed by the following considerations: (a) whether the choice of action by the agent corresponds, at least in part, to obscure or clarify his ability, (b) whether it is reasonable to suppose that the resulting choice of noise cannot be observed ex ante by the principal, at least in part, and (c) whether the outcome, apart from being intrinsically good or bad, serves ex post as an indicator for the extent of risk-taking, thereby leading to some form of inference about the agent’s competence. It is important to appreciate the emphasized phrase in part (b). It is only necessary that there be some significant unobserved component to the choice of risk, not that every aspect of that choice be unobserved. Moreover, as argued in the Supplementary Appendix, it is irrelevant for our purposes whether the agent can affect the mean of the signal. The relevant criterion that determines the applicability of our model is whether the agent affect risk in a way that’s at least partially unobservable. We briefly describe two potential applications.

7.1. Risk-Taking in Delegated Portfolio Management. A risk-neutral investor is looking for a good money manager who will help her invest her money. This is easier said than done. To be sure, there are persistent differences in managerial skill across funds (Chevalier and Ellison, 1999; Berk and van Binsbergen, 2015), but assessing them ex-ante is no trivial task. In large part this is because noise or “luck” appears to dominate skill, at least in the short term (Kritzman, 1987; Fama and French, 2010), and because differences in managerial skills arise from differences in the acquisition and use of specialized knowledge (Coval and Moskowitz, 2001; Kacperczyk et al., 2005; Cohen et al., 2008; Shumway et al., 2011), and this information is hard to access either ex-ante or ex-post.

Our investor’s filtering problem is exacerbated by the fact that underperforming funds are known to inject risk in their midyear portfolios in the hope of catching up with the winners (Brown et al., 1996; Chevalier and Ellison, 1997; Koski and Pontiff, 1999) — a strategy colorfully referred to as “gambling for resurrection.” Presumably, a good current outcome will make our investor happy. But remember
that her main goal is to find a durable relationship with a competent money manager that will deliver higher expected returns, not just now but in the future, and in this sense the (possibly welcome) return today is also a signal about the manager’s type. This specific concern should be separated from one of designing a performance contract for a manager, though it is reasonable to conjecture that parallel considerations would emerge there as well.\textsuperscript{13} For our investor, the compensation scheme is fixed, and we emphasize the reputational dimension of the interaction: in particular, she can terminate or prolong the contractual relationship.

As far as the observability of risk-taking is concerned, we argue that investors do not have easy access to good measures of just how much risk is being taken on, even if they can see the choice of portfolio. After all, if they could fully assess such attributes in real time, they would presumably not need a money manager to begin with. As Palomino and Prat (2003) observe, “most smaller investors do not have the time or the knowledge to perform the monitoring and do not observe the distribution of the portfolio the agent chooses but only the realized return on the portfolio.” This is also true of specialized actors; witness the subprime crisis of 2008.\textsuperscript{14} It is generally hard for investors to infer the level of risk in a given portfolio both ex-ante and ex-post (for other references, see Kritzman, 1987; Sirri and Tufano, 1998; Fama and French, 2010).

This non-observability of risk (or more generally, the distribution function of the signal) is in line with analyses of delegated investment or active asset management (Penno, 1996; Palomino and Prat, 2003; Makarov and Plantin, 2015), delegated project management (DeMarzo et al., 2014; Barron et al., 2017), and optimal security design (Hébert, 2018). It certainly does not mean that all aspects of risk-taking are unobservable. Surely, a financially literate investor should be able to infer something from, say, the extent of diversification in the portfolio or the degree of its departure from well-known index compositions. But as we have argued earlier, this makes absolutely no difference at all to the applicability of our framework, as long as some significant component of risk-taking is unobserved ex ante.\textsuperscript{15}

Our bounded retention equilibrium implies a novel prediction for the literature on mutual funds and investors’ behavior: the probability of assets flowing out of a mutual fund as a function of performance (excess returns) should eventually increase. To our knowledge, this question has not been systematically studied, especially for younger money managers for whom reputation-building is presumably a serious concern. What we do have is evidence of a positive flow-performance correlation at the aggregate level.

\textsuperscript{13}See Barron et al. (2017) for the analysis of optimal incentive schemes in a principal-agent model in which the agent can engage in risk-taking.

\textsuperscript{14}The U.S. Financial Crisis Inquiry Commission determined that no one involved understood the risks they were taking: “The captains of finance and the public stewards of our financial system ignored warnings and failed to question, understand and manage evolving risks within a system essential to the well-being of the American public” (Commission, 2010, emphasis added).

\textsuperscript{15}Dasgupta and Prat (2006) study a setting in which a fund manager faces career concerns. He trades — or passively holds — an asset for a principal who decides whether to retain the manager or not. The good managers know the precise value of the asset, to be later revealed to all. The bad manager is uninformed. In this setting, the manager’s actions are fully observed and understood by the principal. As we’ve already noted in Section 5.4, if actions are observed and can be fully interpreted — e.g., if buying or selling is known to be generically optimal — then a separating equilibrium cannot exist in which the bad type is passive. So that type randomizes between selling and buying, or “churns.” Yet, this kind of excessive trading apart, a good outcome is always a reason for retention in the Dasgupta-Prat analysis.
(Ippolito, 1992; Chevalier and Ellison, 1997; Sirri and Tufano, 1998),\textsuperscript{16} which appears to go in favor of “monotone retention regimes.” Our theory does not rule out such regimes, especially if in the mutual fund industry, the natural level of risk is large relative to the differences in expected returns that bad and good mutual funds can deliver. There is, in fact, some evidence that such differences are “small” (see Fama and French, 2010).

That said, in the absence of more detailed study, we are agnostic about this evidence, which is very much at the aggregative level. First, the relevant retention/replacement decision could be made at the level of the fund. That is, investors pick funds, and funds pick managers. Then the retention-replacement problem is one that faces the fund and not the investor. Hvide (2002) provides anecdotal evidence in favor of this: the CEO of Skandia Fund Management (SFM) confessed to the author that SFM “first selects an initial pool of fund managers and then gradually terminates the relationship with the managers whose return are too high or too low as compared with an index return.”

Second, Chevalier and Ellison (1997) study the response of asset flows to mutual fund performance and the consequent risk-taking behavior of these funds. This relationship is highly nonlinear, and especially so for young funds, which presumably have stronger reputational concerns than their older counterparts. In this paper, the authors conclude: “The one clear regularity in the data that is somewhat puzzling in contrast with our earlier results is that higher excess returns are clearly correlated with larger risk increases.” This is certainly consistent with the predictions of the model for a bounded retention equilibrium, but not so for a monotone threshold equilibrium, where risk choices must be identical for all firm types.

These preliminary remarks do not substitute for a careful study of managerial risk-taking and the proper design of performance contracts (or renewal decisions) aimed at curtailing excessive risk-taking.\textsuperscript{17} Ideally, such a model would allow managers to choose both the mean and the variance of the portfolio. The latter is costless and fits our baseline scenario — imagine loading on pure risk by the use of options, for instance. The former would require costly effort that would vary with manager type, an extension that we briefly explore in Section ?? of the Supplementary Appendix. The proposed analysis would presumably have an empirical component as well, that follows up on Chevalier and Ellison (1997) and related literature. Anecdotal evidence on financial ventures or Ponzi schemes that promise (and initially deliver) high rates of return suggests that careful individuals often stay away from such ventures — to be sure, others don’t.

\textbf{7.2. Risky Politics.} The unobservability of risk is a salient feature of situations when the observer either does not fully know or cannot judge the full set of consequences (and associated likelihoods) of an observed action. This is true, for instance, of risky political actions. An observer might be able to “compute” the risk that political of different competencies are likely to be taking, just as an agent

\textsuperscript{16}Given that persistence in performance is rather weak (Gruber, 1996; Zheng, 1999; Bollen and Busse, 2001), except for the worst performing funds (Hendricks et al., 1993; Carhart, 1997; Berk and van Binsbergen, 2015), it is unclear that such behavior is rational, though see Berk and Green (2004).

\textsuperscript{17}The use of risky gambles by managers with career concerns is studied in a dynamic setting by Makarov and Plantin (2015).
computes equilibrium play from her beliefs about opponent strategies, but that is different from actually observing that risk ex-ante.

There are two main reasons that sustain this line of argument: either voters can not fully comprehend the implications of a given policy (much as our investor in the previous example), or they are severely underinformed about policy. The first argument is part of the seminal work of Arnold (1990), who analyzes congressional action. For instance, most citizens prefer less inflation to more, but at the same time support price controls to fight high inflation, a position that stems from simplistic or even erroneous views of the underlying mechanics (or causal relationships) of the problem. The second reason is based on empirical evidence (see, for example, Delli Carpini and Keeter, 1996; Somin, 2013; Baum and Kernell, 1999; Prior, 2007), typically collected through surveys, that shows public unawareness of policy, even around local issues could affect their everyday life. Perhaps the acquisition of information is costly, and the benefits are perceived to be distant or indirect. As Schumpeter (1942) notes (p. 261): “[national and international affairs] seem so far off; they are not at all like a business proposition; dangers may not materialize at all and if they should they may not prove so very serious; one feels oneself to be moving in a fictitious world.” In addition, the sheer knowledge of a policy does not imply knowledge of all the potential implications of that policy. In effect, and in the language of this paper, a voter may not fully observe the risk of a policy.

Now think of a political leader, the assessment of whose competence is currently important, and who seeks to be “retained” by the median voter (who plays here the role of the principal). If that leader is competent, he can attempt to play it safe by implementing reliable but unspectacular policies, and so the better will be the fix that the public obtains about his true type after a policy outcome is realized — though convergence to that understanding may be far from total. In contrast, the incompetent leader can entertain an alternative policy which he knows to be riskier than the unambitious policy of the competent leader. For instance — and only speaking hypothetically — he might attempt to conduct a denuclearization summit with the authoritarian leader of a rogue state. When observing this policy choice, the median voter is not aware of all the risks entailed, but she can evaluate the policy ex-post in terms of its success (or lack thereof).

We reiterate: to the extent that the implications of the policies can be observed ex-ante, both types of leader must pool on those observable risks — with binary types, separation cannot occur before the realization of the policy. We would therefore have monotone retention. However, when observability is imperfect, and especially when the voter feels optimistic about future political candidates, and the difference in competence of the two leaders is large enough, the incompetent leader will choose the policy that he knows to be riskier — in the language of our example, he will pursue the denuclearization summit. Then, a striking success from such a policy — if, continuing the hypothetical streak, one were to occur — should be treated with a certain degree of reticence by the median voter. It could be a sign of extreme competence. It could also be sign of a desperate move by a largely incompetent individual, which happened to pay off. That outcome, if it occurs, may be good for society. But it may not be a good signal on which to base re-election.
8. Summary

We’ve studied a model in which an agent who seeks to be retained by a principal might deliberately inject noise into a process that signals his type. Possible equilibrium regimes include monotone retention, in which a principal retains if an agent’s signal is high enough, and various non-monotone regimes. Of these, we argue that bounded retention is the salient equilibrium regime. In it, different types of agents choose different degrees of noise, with worse agents behaving more noisily. The resulting equilibrium has a “double-threshold” property: the principal retains the agent if the signal is good, but neither too good nor too bad. We discuss extensions to a variant with costly noise, to a dynamic version with agent term limits, and to non-normal signal structures.

At the heart of our argument is a fundamental failure of “single-crossing.” In our setting, we know that with any reasonable assumptions on the signal distribution, higher means are stochastically associated with better signals, in the sense that the likelihood ratio of the high mean (relative to the low mean) rises with the emitted signal. But once the choice of noise enters the picture, single-crossing is irretrievably damaged. Types with lower means are more likely to choose higher noise, and the likelihood ratio behave in more complex, non-monotone ways as a function of the signal realization. Such a failure is a feature that generally renders a full analysis intractably hard. In our setting, it leads to a simple yet rich model in which equilibria can be described — and have interesting properties.

We believe that the deliberate injection of ambiguity or noise is a central feature of many principal-agent interactions. Throughout, we make the central assumption that the extent of noise cannot be fully observed by the principal, and must be inferred, at least to some degree. We believe this assumption holds in many settings, in which the receiver does not fully understand, ex ante, the full range of possible options available to the agent. In this paper, we have discussed two such applications — risky portfolio management, and the choice of risky political strategy. But there is a plethora of other situations that our analysis could fit: a non-governmental organization of unknown competence seeking funding from donors, risky versus safe strategies in the deliberate generation of leaks, a government under pressure which might inject noise into official statistics, an individual taking risky steps to bolster a cv for an upcoming promotion or interview, a less-than-competent lawyer calling a high-risk witness (who could destroy the case or win it), an athlete who might engage in doping, a news media outlet using sensationalist headlines to get readership, and so on. In all these situations, full observability of strategic risk would restore single-crossing, and generate standard results. However, when there are constraints on the observability of risk, our framework makes a new contribution towards the understanding of such environments.

Appendix: Main Proofs

Proof of Proposition 1. The proof of this proposition is long and contains several steps, with many technical details relegated to the Supplementary Appendix. Recall that the discussion in Section 5.1 eliminates all bounded replacement equilibria. With that out of the way, we focus on monotone and bounded retention regimes, and agent responses to them.
Lemma 1. With bounded retention, $\sigma_b > \sigma_g$, and $X = [x_-, x_+]$, where $\theta_g < \frac{x_+ + x_-}{2} < x_+$.

Proof. When $\sigma_b \neq \sigma_g$, and $x_-$ and $x_+$ are both finite and given by (4), one can check that

$$\frac{x_+ + x_-}{2} = \frac{\sigma^2_b / \theta_g - \sigma^2_g / \theta_b}{\sigma^2_b - \sigma^2_g / \theta^2_b}.$$ 

So if $\sigma_b > \sigma_g$ then $x_+ > \frac{x_+ + x_-}{2} > \theta_g$.

Lemma 2. In a bounded retention equilibrium with thresholds $x_-$ and $x_+$, and for each $k$,

$$(23) \quad \phi\left(\frac{x_- - \theta_k}{\sigma_k}\right) > \phi\left(\frac{x_+ - \theta_k}{\sigma_k}\right).$$

Proof. With bounded retention, $\sigma_b > \sigma_g$, and $(x_+ + x_-)/2 > \theta_k$ by Lemma 1, and so

$$\frac{x_+ - \theta_k}{\sigma_k} > \frac{\theta_k - x_-}{\sigma_k},$$

which implies, using single-peakedness and symmetry of $\phi$ around 0, along with $x_+ > x_-$, that

$$\phi\left(\frac{x_+ - \theta_k}{\sigma_k}\right) < \phi\left(\frac{\theta_k - x_-}{\sigma_k}\right) = \phi\left(\frac{x_- - \theta_k}{\sigma_k}\right),$$

which establishes (23).

Lemma 3. (i) If $X = [x^*, \infty)$ and $\theta_k > x^*$, the agent chooses $\sigma_k = \sigma$; if $\theta_k < x^*$, the problem has no solution, in particular, the agent always wants to inject additional noise; if $\theta_k = x^*$, the agent is indifferent across all choices of $\sigma$.

(ii) Assume a retention zone of the form $[x_-, x_+]$ with $x_- < x_+$. If $x_- \leq \theta_k$, then $\sigma_k = \sigma$.

(iii) Assume a retention zone of the form $[x_-, x_+]$ with $x_- < x_+$. If $x_+ > \theta_k$, then for each $k$ define

$$(24) \quad d_k(\sigma_k) := \phi\left(\frac{x_- - \theta_k}{\sigma_k}\right)(x_+ - \theta_k) - \phi\left(\frac{x_+ - \theta_k}{\sigma_k}\right)(x_- - \theta_k) \quad \text{for all } \sigma_k > 0.$$ 

Then $d_k$ is continuous, initially positive then negative, with a unique root to $d_k(\sigma_k) = 0$, given by

$$\sigma^*_k = \sqrt{\frac{(x_+ - x_-)(x_+ + x_- - \theta_k)}{\ln(x_+ - \theta_k) - \ln(x_- - \theta_k)}} \in (x_- - \theta_k, x_+ - \theta_k),$$

and agent $k$ sets $\sigma_k = \max(\sigma, \sigma^*_k)$.

Proof. (i) In the case of monotone retention, the first-order derivative with respect to $\sigma_k$ is

$$\phi\left(\frac{x^* - \theta_k}{\sigma_k}\right) \left(\frac{x^* - \theta_k}{\sigma_k^2}\right).$$

It is always negative if $x^* < \theta_k$, so $\sigma_k = \sigma$; always positive if $x^* > \theta_k$, so the agent always wants to increase the noise and the problem has no solution; and always equal to 0 if $x^* = \theta_k$, so the agent is indifferent across all choices of $\sigma$. 


(ii) A type-$k$ agent wishes to maximize the probability of being in the retention zone $[x_-, x_+]$, so he chooses $\sigma_k \geq \sigma$, to maximize

\[
\Phi \left( \frac{x_+ - \theta_k}{\sigma_k} \right) - \Phi \left( \frac{x_- - \theta_k}{\sigma_k} \right),
\]

where $\Phi$ is the cdf of the standard normal. The first-order derivative of the objective function with respect to $\sigma_k$ is

\[
d_k(\sigma_k) = \frac{1}{\sigma_k^2} \left[ \phi \left( \frac{x_- - \theta_k}{\sigma_k} \right) (x_- - \theta_k) - \phi \left( \frac{x_+ - \theta_k}{\sigma_k} \right) (x_+ - \theta_k) \right],
\]

where $d_k$ is defined in (24). By Lemma 1, $x_+ > \theta \geq \theta_k$ for any $k$. If in addition, $x_- \leq \theta_k$, then the sign of the derivative is always negative, so $\sigma_k = \sigma$.

(iii) When $\theta_k < x_- < x_+$, the sign of the derivative depends on the value of $\sigma_k$. After some elementary manipulation, we see that

\[
d_k(\sigma_k) = \phi \left( \frac{x_+ - \theta_k}{\sigma_k} \right) (x_+ - \theta_k) \left\{ \exp \left[ \frac{x_+ - x_-}{\sigma_k^2} \left( \frac{x_- + x_+}{2} - \theta_k \right) \right] \left( \frac{x_- - \theta_k}{x_+ - \theta_k} \right) - 1 \right\}.
\]

The term inside the curly brackets is the only one that can change sign. Moreover, this term is continuous and strictly decreasing in $\sigma_k$, with limit $\frac{x_- - \theta_k}{x_+ - \theta_k} - 1 < 0$ when $\sigma_k \to \infty$, and with limit $\infty$ as $\sigma_k \to 0$. So $d_k$ has all the claimed properties, and there exists a unique $\sigma_k^*$ that solves (26), given by setting the term within curly brackets equal to zero, which yields:

\[
\sigma_k^* = \sqrt{\frac{(x_+ - x_-)(\frac{x_- + x_+}{2} - \theta_k)}{\ln (x_+ - \theta_k) - \ln (x_- - \theta_k)}}.
\]

Therefore, the agent will optimally choose $\sigma_k = \max \{ \sigma, \sigma_k^* \}$.

To show that $\sigma_k^* \in (x_- - \theta_k, x_+ - \theta_k)$, first define $\hat{x}_k := [(x_+ - \theta_k)/(x_- - \theta_k)]^2 \in (1, \infty)$. Provided $x_- > \theta_k$, we will have $\theta_k + \sigma_k^* > x_-$ if and only if $\hat{x}_k - 1 > \ln (\hat{x}_k)$, which is always true because equality holds at $\hat{x}_k = 1$ and then the left-hand side increases at a rate of 1, whereas the right-hand side increases at a rate of $1/\hat{x}_k < 1$. Similarly, $\theta_k + \sigma_k^* < x_+$ iff $1 - (1/\hat{x}_k) < \ln (\hat{x}_k)$. The condition holds with equality for $\hat{x}_k = 1$, and the derivatives of the left and right-hand sides are $1/\hat{x}_k^2$ and $1/\hat{x}_k$, respectively, making the condition valid for any $\hat{x}_k \in (1, \infty)$.

We will use part (iii) of Lemma 3 to construct our fixed point map. But first we note:

**Lemma 4.** In any non-trivial equilibrium, $\sigma_g = \sigma$.

**Proof.** From (4) it is clear that the principal employs a monotone retention regime if and only if both agent types choose the same level of noise, $\sigma_g = \sigma_b = \sigma$. In fact, by Lemma 3(i), $\sigma_g = \sigma_b = \sigma$. Otherwise, a non-trivial equilibrium must have bounded retention, in which case $\sigma_g < \sigma_b$ by Observation 1. Suppose,
on the contrary, that $\sigma < \sigma_g$. Then both choices of noise are interior, and so agent optimality requires
\[
\phi \left( \frac{x_+ - \theta_b}{\sigma_b} \right) (x_+ - \theta_b) = \phi \left( \frac{x_- - \theta_b}{\sigma_b} \right) (x_- - \theta_b),
\]
\[
\phi \left( \frac{x_- - \theta_g}{\sigma_g} \right) (x_- - \theta_g) = \phi \left( \frac{x_+ - \theta_g}{\sigma_g} \right) (x_+ - \theta_g).
\]
Combining these equations with the principal’s indifference condition (6), we obtain
\[
\phi \left( \frac{x_- - \theta_g}{\sigma_g} \right) = \phi \left( \frac{x_+ - \theta_g}{\sigma_g} \right),
\]
which contradicts Lemma 2.

Lemmas 3 and 4 help us introduce a mapping, the fixed point(s) of which will be interpreted as equilibrium; conditions (11) and (12) will enter the discussion here. Consider a self-map $\Psi$ on $(\sigma, \infty)$, with domain to be interpreted as the principal’s conjecture about the noise used by the low type, and range as the subsequent optimal choice of noise by the bad type, in response to the retention decision. (Throughout, informed by Lemma 4, $\sigma_g = \sigma$.) Guided by part (iii) of Lemma 3, our self-map is:

\[
(27) \quad \Psi(\sigma) \equiv \max \left\{ \sqrt{\frac{[x_+(\sigma) - x_-(\sigma)](x_+(\sigma) + x_-(\sigma) - \theta_b)}{[\ln(x_+(\sigma) - \theta_b) - \ln(x_-(\sigma) - \theta_b)]}} \right\},
\]
where for any $\sigma > \sigma_g$,

\[
(28) \quad x_-(\sigma) := \frac{\sigma^2 \theta_g - \sigma^2 \theta_b - \sigma \sigma R(\sigma)}{\sigma^2 - \sigma^2} \quad \text{and} \quad x_+(\sigma) := \frac{\sigma^2 \theta_g - \sigma^2 \theta_b + \sigma \sigma R(\sigma)}{\sigma^2 - \sigma^2},
\]
with

\[
(29) \quad R(\sigma) := + \sqrt{\left( \frac{\theta_g - \theta_b}{\sigma} \right)^2 + \left( \frac{\sigma^2 - \sigma^2}{\sigma^2 - \sigma^2} \right)^2 \ln \left( \frac{\beta \sigma}{\sigma_g} \right)}.
\]
To interpret these objects, notice that $x_-(\sigma)$ and $x_+(\sigma)$ are the roots to

\[
(30) \quad \beta \frac{1}{\sigma} \phi \left( \frac{x - \theta_g}{\sigma} \right) = \frac{1}{\sigma} \phi \left( \frac{x - \theta_b}{\sigma} \right),
\]
so these bound the retention regime $X$ when the principal expects $(\sigma_b, \sigma_g) = (\sigma, \sigma)$. (We will verify that these bounds are well-defined.) Given these thresholds, type $b$ reacts as in Lemma 3(iii). So $\Psi(\sigma)$ can be interpreted as $b$’s reaction to a chain that starts with a conjecture about $b$’s action ($\sigma$), travels via the principal’s thresholds, and culminates in that type’s optimal reaction to those thresholds. Hence a fixed point of $\Psi$ must correspond to an equilibrium with bounded retention, and all such equilibria can be described in this way.

Our first task is to make sure that $x_-(\sigma)$ and $x_+(\sigma)$ are well-defined and distinct for $\sigma > \sigma_g$. The following lemma relates this to condition (11).

**Lemma 5.** If $\beta \geq 1$, $x_-(\sigma)$ and $x_+(\sigma)$ are well-defined and distinct for $\sigma > \sigma_g$. If $\beta < 1$, $x_-(\sigma)$ and $x_+(\sigma)$ are well-defined and distinct for $\sigma > \sigma_g$ if and only if (11) holds.
Proof. When $\beta \geq 1$, it is clear that the term within the square root in (29) is strictly positive for all $\sigma > \sigma$. In the Supplementary Appendix we show that, when $\beta < 1$, this term is strictly positive for all $\sigma > \sigma$ if and only if (11) holds.

As already mentioned, we follow the lead of Lemma 4 in holding $\sigma_g$ at $\sigma$ throughout. Nevertheless, when all is said and done, we must make sure that the good type willingly chooses this value when confronted with the principal’s retention strategy. We get this out of the way before proceeding any further.

**Lemma 6.** If $\sigma_b = \sigma$ satisfies $d_b(\sigma) = 0$ and $\{x_-(\sigma), x_+(\sigma)\}$ are the roots to (30), then the good type optimally chooses $\sigma_g = \sigma$.

**Proof.** By Lemma 1, $x_+ - \theta_b(\sigma) > \theta_g$. If, in addition, $x_- - \theta_b(\sigma) \leq \theta_g$, then by Lemma 3 (ii), type $g$ chooses $\sigma_g = \sigma$, and we are done.

Otherwise, $x_- - \theta_g > \theta_g$. Then by Lemma 3 (iii), there is a unique $\sigma_g$ (not worrying about the lower bound $\sigma$) maximizing $g$’s probability of retention. This solves $d_g(\sigma_g) = 0$, where $d_g$ is defined in (24). We claim that this value is smaller than $\sigma$. By Lemma 3 (iii), it will suffice to show that $d_g(\sigma) < 0$.

Because $d_b(\sigma) = 0$, we see from (24) that

$$
\phi\left(\frac{x_- - \theta_b}{\sigma}\right)(x_- - \theta_b) = \phi\left(\frac{x_+ - \theta_b}{\sigma}\right)(x_+ - \theta_b).
$$

(31)

It follows that

$$
d_g(\sigma) = \phi\left(\frac{x_- - \theta_g}{\sigma}\right)(x_- - \theta_g) - \phi\left(\frac{x_+ - \theta_g}{\sigma}\right)(x_+ - \theta_g)
$$

$$
= \frac{\sigma}{\beta} \left[ \phi\left(\frac{x_- - \theta_b}{\sigma}\right) \frac{x_- - \theta_g}{\sigma} - \phi\left(\frac{x_+ - \theta_b}{\sigma}\right) \frac{x_+ - \theta_g}{\sigma} \right]
$$

$$
= \frac{\sigma}{\beta} \phi\left(\frac{x_- - \theta_b}{\sigma}\right) \left[ \frac{x_- - \theta_g}{x_- - \theta_b} - \frac{x_+ - \theta_g}{x_+ - \theta_b} \right] < 0,
$$

where the second equality follows from (30), the third equality from (31), and the very last inequality from $\theta_g > \theta_b$ and $x_+(\sigma) > x_-(\sigma)$.

With the good type dealt with, we return to the fixed point problem for the bad type. In preparation for the steps ahead, the two retention thresholds $x_-(\sigma)$ and $x_+(\sigma)$ are shown as the lowest and highest curves in Figure 7. These mark the principal’s best-response thresholds for every $\sigma_b = \sigma > \sigma$ (remember that type-$g$ is kept fixed at $\sigma_g = \sigma$ in line with Lemma 4). Now consider type b’s best response to these thresholds. Lemma 3 (iii) tells us that this best response plus $\theta_b$ must lie strictly between the $x_-(\sigma)$ and $x_+(\sigma)$ loci. This is shown as the thick intermediate curve. Our fixed point(s) will be determined by the intersection(s) between this curve and the $\theta_b + \sigma$ line (depicted as a shifted diagonal line). The analysis below will tell us the conditions under which these intersections will or will not be possible, and will also establish uniqueness (conditional on existence). These observations together constitute the foundations of the statement: “A nontrivial equilibrium exists if and only if (11) is satisfied, and it
is then unique.” We begin with a lemma that serves as formal description of the shapes of $x_-(\sigma)$ and $x_+ (\sigma)$ in the figure.

**Lemma 7.** Assume that either $\beta \geq 1$ or $\beta < 1$ and (11) holds. Then:

(i) $\lim_{\sigma \to \sigma} x_-(\sigma) = x^*(\sigma)$ and $\lim_{\sigma \to \sigma} x_+(\sigma) = \infty$, where $x^*(\sigma)$ is defined in (5).

(ii) $\lim_{\sigma \to \infty} x_-(\sigma) = -\infty$ and $\lim_{\sigma \to \infty} x_+(\sigma) = \infty$.

(iii) If $\beta \geq 1$ and (12) fails, then $x_-(\sigma) < \theta_b$ for all $\sigma > \sigma$.

**Proof.** See Supplementary Appendix.

With Lemmas 5 and 7 in hand, we can state:

**Lemma 8.** If $\beta \leq 1$ and (11) holds, or $\beta \geq 1$ and (12) holds, there is a unique non-trivial equilibrium. It has bounded retention.

**Proof.** By Lemma 7 (i), $\lim_{\sigma \to \sigma} x_-(\sigma) = x^*(\sigma)$. Inspect the definition of $x^*(\sigma)$ in (5) and note that if $\beta \leq 1$ or if $\beta > 1$ and (12) holds, then $x^*(\sigma) > \theta_b$. Also by Lemma 7 (i), $\lim_{\sigma \to \sigma} x_+(\sigma) = \infty$. Using this information in (27), we see that $\lim_{\sigma \to \sigma} \Psi(\sigma) = \infty$.

Next, by Lemma 7 (ii), the interval $(x_-(\sigma), x_+(\sigma))$ must contain $\theta_b$ for all $\sigma$ large, so that by Lemma 3 (ii), $\Psi(\sigma) = \sigma$ for all such $\sigma$.

Moreover, by Lemma 5, $x_-(\sigma)$ and $x_+(\sigma)$ are well-defined and distinct for every $\sigma > \sigma$, and these values move continuously with $\sigma$. Consequently, so does $\Psi(\sigma)$. The above end-point verifications and continuity guarantee that $\Psi$ has at least one fixed point.

At any such fixed point $\sigma$, we have $\sigma < \sigma = \Psi(\sigma)$. Consequently, the first term on the right hand side of (27) must bind. It follows that $\Psi(\sigma)$ solves $d_b(\Psi(\sigma)) = 0$, where $d_b$ is defined in (24), so that

$$(32) \quad \phi \left( \frac{x_+(\sigma) - \theta_b}{\Psi(\sigma)} \right) (x_+(\sigma) - \theta_b) = \phi \left( \frac{x_-(\sigma) - \theta_b}{\Psi(\sigma)} \right) (x_-(\sigma) - \theta_b).$$
Figure 8. $x_-(\sigma)$ and $x_+(\sigma)$ are not always well-defined if (11) fails.

Equation (32) can be used to compute $\Psi'(\sigma)$. The Supplementary Appendix indicates the steps and shows that this derivative is strictly negative at any fixed point. So $\Psi(\sigma)$ is strictly decreasing at any fixed point, and therefore can have just one fixed point $\sigma^+ > \bar{\sigma}$, as asserted. At this fixed point, both the principal and the bad type are playing best responses. That the good type is also playing a best response is guaranteed by Lemma 6. Therefore $\sigma^+ > \bar{\sigma}$ is the only equilibrium with bounded retention.

It remains to eliminate the monotone equilibrium, which must involve $\sigma_b = \sigma_g$ and therefore (by Lemma 4) a common value of $\sigma$. Because both types must play a best response, it follows from Lemma 3(i) that

$$x^*(\sigma) = \frac{\theta_g + \theta_b}{2} - \frac{\sigma^2}{\theta_g - \theta_b} \ln(\beta) \leq \theta_b$$

or

$$\ln(\beta) \geq \frac{\Delta^2}{2\sigma^2},$$

which would contradict (12) when $\beta \geq 1$, or is impossible under $\beta \leq 1$. So only bounded retention equilibria can exist.

Lemma 9. If $\beta \geq 1$ and (12) fails, there is a unique non-trivial equilibrium. It has monotone retention.

Proof. If $\beta \geq 1$ and (12) fails, then $x_-(\sigma) < \theta_b$ for all $\sigma > \bar{\sigma}$ by Lemma 7 (iii). At the same time, by Lemma 1, $\theta_b < x_+(\sigma)$, so $\theta_b \in (x_-(\sigma), x_+(\sigma))$ for all $\sigma \geq \bar{\sigma}$. So by Lemma 3(ii), $\Psi(\sigma) = \bar{\sigma}$ for all $\sigma > \bar{\sigma}$ and has no fixed point with $\sigma > \bar{\sigma}$.

We separately verify that there is an equilibrium with monotone retention and both types choosing $\sigma$. If $(\sigma_b, \sigma_g) = (\bar{\sigma}, \bar{\sigma})$, then the planner uses the monotone retention strategy with threshold $x^*(\sigma)$. Because $\beta \geq 1$ and (12) fails, $x_-(\sigma) \leq \theta_b < \theta_g$. By Lemma 3(i), it is a best response for both types to choose $\sigma$.

To complete our characterization of equilibrium using Conditions (11) and (12), we note:

Lemma 10. If $\beta < 1$ and (11) fails, a non-trivial equilibrium does not exist.
Proof. The details of this argument center around establishing the validity of Figure 8. When \( \beta < 1 \) and (11) fails, the roots to the principal’s indifference condition, \( x_{-}(\sigma) \) and \( x_{+}(\sigma) \), are not always well-defined or distinct. But matters are more subtle than that: for the values of \( \sigma \) at which they are well defined and distinct, a fixed point is impossible. This happens because at any such value of \( \sigma \) we either have that \( \theta_b + \sigma < x_{-}(\sigma) < x_{+}(\sigma) \), or \( x_{-}(\sigma) < x_{+}(\sigma) < \theta_b + \sigma \) (look at Figure 8 again). But type-\( b' \)'s best response in a bounded retention equilibrium regime must satisfy \( x_{-}(\sigma) < \theta_b + \sigma_b < x_{+}(\sigma) \), as asserted by Lemma 3 (iii). Therefore, no bounded retention equilibrium is possible. See the Supplementary Appendix for the formal details.

We can now complete the proof of Proposition 1. Part (i) is a direct consequence of Lemmas 8 and 10. Part (ii) is proved immediately by Lemmas 8 and 9. The description of bounded retention equilibria in Part (iii) is proved by combining Lemmas 1, 4 and 3 (iii).

Proof of Proposition 2. Recall (15); this is the equation that \( \sigma \) must satisfy if it commonly chosen by both types:

\[
\phi(z_1)z_1 = \phi(z_2)z_2 = -\sigma c'(\sigma),
\]

where \( z_1 = (\sigma/\Delta) \ln(\beta) - (\Delta/2\sigma) \) and \( z_2 = (\sigma/\Delta) \ln(\beta) + (\Delta/2\sigma) \). The function \( \phi(z)z \) has the shape shown in Figure 4, reaching maxima and minima at \( z = 1 \) and \( z = -1 \) respectively, and exhibiting “negative symmetry” around 0. Using (15), this tells us that there are two exclusive possibilities: (i) either \( \beta > 1 \) and \( \sigma < g \), or (ii) either \( \beta < 1 \) and \( \sigma > g \). We study (i); Case (ii) is dealt with in the same way.

In Case (i), elementary computation shows that \( z_2 \), viewed as a function of \( \sigma \) (holding all other terms constant) starts from infinity as \( \sigma = 0 \), declines to a minimum of \( \sqrt{2 \ln(\beta)} \), and then climbs monotonically again to \( \infty \) as \( \sigma \to \infty \). Meanwhile, \( z_1 \) is always increasing in \( \sigma \), and is exactly zero when \( z_2 \) reaches its minimum. From this point on, \( \phi(z_1)z_1 \) climbs from 0 to its maximum value of \( \phi(1) \) and then falls, while \( \phi(z_2)z_2 \) falls monotonically from a positive value to zero. Finally, we note that in the phase where \( \phi(z_1)z_1 \) falls, we have \( \phi(z_1)z_1 > \phi(z_2)z_2 \) throughout. Putting these observations together, we must conclude that there is a unique value of \( \sigma \) such that the first equality in (33) holds, and it is independent of the cost function \( c \).

Proof of Proposition 3. Recall that \( \sigma_s < g \) and \( \sigma^* > g \) are the two solutions to \( c(\sigma) = 1 \). Let \( \Sigma := [\sigma_s, \sigma^*]^2 \), and define

\[
\Sigma^+ := \{(\sigma_b, \sigma_b) \in \Sigma | \sigma_b \geq \sigma_g\}.
\]

For each \( \sigma \in \Sigma^+ \), define \( x_{-} \) and \( x_{+} \) by the distinct lower and upper roots to (4) if \( \sigma_b > \sigma_g \); otherwise, if \( \sigma_b = \sigma_g = \sigma \), set \( x_{-} = x'(\sigma) \) as defined in (5) and \( x_{+} = \infty \). Interpret \( [x_{-}, x_{+}] \) as the retention zone. Call this map \( \Psi_1 \). As discussed in the main text, this map is well-defined when \( \sigma_b = \sigma_g \). To check that \( \Psi_1 \) is also well-defined when \( \sigma_b > \sigma_g \), we must show that there are two distinct real roots to the quadratic in (4), or equivalently, using the elementary formula for quadratic roots, that the expression

\[
\Delta^2 + (\sigma_b^2 - \sigma_g^2) 2 \ln\left(\frac{\sigma_b}{\sigma_g}\right)
\]
is strictly positive. But (17) tells us that \( \ln(\beta) \geq -[\Delta^2]/2\sigma^2 \), and so
\[
\Delta^2 + \left( \sigma_b^2 - \sigma_g^2 \right) 2 \ln \left( \frac{\sigma_b}{\sigma_g} \right) = \Delta^2 + \left( \sigma_b^2 - \sigma_g^2 \right) 2 \ln \left( \frac{\sigma_b}{\sigma_g} \right) \\
\geq \Delta^2 + \left( \sigma_b^2 - \sigma_g^2 \right) 2 \ln (\beta) \\
\geq \Delta^2 \left[ 1 - \frac{\sigma_b^2 - \sigma_g^2}{\sigma^2} \right] > 0,
\]
where the very last inequality uses \( \sigma' \geq \sigma_b > \sigma_g \). So there are distinct roots \( x_- < x_+ \), and by exactly the same logic as for Observation 1, the zone \([x_-, x_+] \) must involve retention.

Next, for each pair \((x_-, x_+) \) with \( x_+ > x_- \) and with \( x_+ \) possibly infinite, define \((\sigma'_g, \sigma'_b) \) to be the best-response choices of noise by the bad and good types who face the retention zone \([x_-, x_+] \). By condition [U], these choices are well-defined and unique. Call this map \( \Psi_2 \).

Finally, define a map \( \Psi \) with domain \( \Sigma^+ \) and range \( \Sigma \) by \( \Psi := \Psi_2 \circ \Psi_1 \). We claim that \( \Psi \) is continuous. We first argue that \( \Psi_1 \) is continuous in the extended reals. That is:

(i) if \((\sigma'^n_g, \sigma'^n_b) \to (\sigma_g, \sigma_b) \) with \( \sigma_b > \sigma_g \), then \( \Psi_1(\sigma_g, \sigma_b) = (x_-, x_+) \) with \( x_- < x_+ < \infty \), and it is obvious that \( \Psi_1(\sigma'^n_g, \sigma'^n_b) \to \Psi_1(\sigma_g, \sigma_b) \).

(ii) if \((\sigma'^n_g, \sigma'^n_b) \to (\sigma_g, \sigma_b) \) with \( \sigma_b = \sigma_g \), then \( \Psi_1(\sigma_g, \sigma_b) = (x_-, \infty) \). In this case, an inspection of the quadratic condition (4) (the roots of which yield \( x_- \) and \( x_+ \)) reveals that \( \Psi_1(\sigma'^n_g, \sigma'^n_b) = (x'^n_-, x'^n_+) \) must satisfy \( x'^n_+ \to \infty \).

Now we turn to the map \( \Psi_2 \). As already mentioned, condition [U] guarantees that best-response noise choices are unique, as long as \( x_+ > x_- \). They are fully characterized by the first-order condition (13), which we reproduce here for convenience:

\[
\phi \left( \frac{x_- - \theta_k}{\sigma_k} \right) \left( \frac{x_- - \theta_k}{\sigma_k} \right) - \phi \left( \frac{x_+ - \theta_k}{\sigma_k} \right) \left( \frac{x_+ - \theta_k}{\sigma_k} \right) = \sigma_k c' \left( \sigma_k \right)
\]
where we include the possibility that \( x_+ = \infty \) by setting \( \phi(z)z = 0 \) when \( z = \infty \).

Pick any sequence \((x'^n_-, x'^n_+) \) that converges in the extended reals. That is, either the sequence converges to \((x_-, x_+) \) with \( x_+ < \infty \), or it converges to a limit of the form \((x_-, \infty) \). Let \( \sigma'^n_k \) be the best responses for an agent of type \( k \), and let \( \sigma_k \) be the best response at the limit value \((x_-, x_+) \). When \( x_+ < \infty \), it is obvious from (34) that \( \sigma'^n_k \to \sigma_k \). In the latter case, the fact that \( \sigma'^n_k \to \sigma_k \) follows from the additional observation that \( \phi(z)z^n \to 0 \) for any sequence \( z^n \to \infty \).

We claim that \( \Psi \) is inward pointing; that is, for every \((\sigma_g, \sigma_b) \in \Sigma^+ \), there exists \( a > 0 \) such that
\[
(\sigma_g, \sigma_b) + a[\Psi(\sigma_g, \sigma_b) - (\sigma_g, \sigma_b)] \in \Sigma^+.
\]

First observe that for every \((\sigma_g, \sigma_b) \in \Sigma^+ \), we have \((\sigma', \sigma) \leq \Psi(\sigma_g, \sigma_b) \leq (\sigma^*, \sigma^*) \). Therefore, if \((\sigma_g, \sigma_b) \in \Sigma^+ \) with \( \sigma_b > \sigma_g \), (35) is easily seen to hold: for \( a > 0 \) and small, it must be that both components of the vector
\[
(\sigma_g, \sigma_b) + a[\Psi(\sigma_g, \sigma_b) - (\sigma_g, \sigma_b)]
\]
lie in \([\sigma_*, \sigma'^*] \), and the second component is larger than the first. The remaining case is one in which \((\sigma_g, \sigma_b) \in \Sigma^+ \) with \( \sigma_b = \sigma_g \). In this case, we know from condition (17) that \( \Psi_1(\sigma_g, \sigma_b) \) is of the form
\((x_-, x_+) = (x^*, \infty)\), where \(x^* \in [\theta_b, \theta_g]\). From the first-order conditions that describe each type — see (14) — it is easy to see that \(\sigma_k \geq \sigma\) when \(x^* \geq \theta_k\). Therefore \(\Psi_2(x^*, \infty) = (\sigma'_g, \sigma'_b)\) must have the property that \(\sigma'_b > \sigma'_g\) (and of course each component lies between \(\sigma\) and \(\sigma^\star\)). It follows that for every \(a \in (0, 1)\), (35) holds, and the claim is proved.

To summarize, we have: \(\Sigma^+\) is a nonempty, compact, convex subset of Euclidean space, and \(\Psi\) is continuous on \(\Sigma^+\). In general, however, \(\Psi\) will fail to map from from \(\Sigma^+\) to \(\Sigma^+\). However, the map is inward pointing in the sense of Halpern (1968) and Halpern and Bergman (1968); for an exposition, see Aliprantis and Border (2006, Definition 17.53). By the Halpern-Bergman fixed point theorem (see Aliprantis and Border, 2006, Theorem 17.54), there exists \((\sigma_g, \sigma_b)\) \(\in \Sigma^+\) such that \(\Psi(\sigma_g, \sigma_b) = (\sigma_g, \sigma_b)\). It is easy to see that \((\sigma_g, \sigma_b)\), along with the associated bounded retention zone \(\Psi_1(\sigma_g, \sigma_b)\), forms an equilibrium.

For proving Proposition 4, we first state the following two results.

**Lemma 11.** In a bounded replacement equilibrium with thresholds \(x_-\) and \(x_+\), for each \(k\),

\[
\phi\left(\frac{x_- - \theta_k}{\sigma_k}\right) > \phi\left(\frac{x_+ - \theta_k}{\sigma_k}\right).
\]

**Proof.** When \(\sigma_b \neq \sigma_g\), and \(x_-\) and \(x_+\) are both finite and given by (4), we have that

\[
\frac{x_+ + x_-}{2} = \frac{\sigma^2_g \theta_b - \sigma^2_b \theta_g}{\sigma^2_b - \sigma^2_g}.
\]

So if \(\sigma_b < \sigma_g\) then \(x_+ < \frac{x_+ + x_-}{2} < \theta_b\). Then,

\[
\frac{x_+ - \theta_k}{\sigma_k} < \frac{\theta_k - x_-}{\sigma_k},
\]

which implies, by single-peakedness and symmetry of \(\phi\) around 0, and \(x_+ < x_-\), that

\[
\phi\left(\frac{x_+ - \theta_k}{\sigma_k}\right) < \phi\left(\frac{\theta_k - x_-}{\sigma_k}\right) = \phi\left(\frac{x_- - \theta_k}{\sigma_k}\right),
\]

which establishes (36).

**Lemma 12.** Under (17) and (18), \(x_+ < \theta_b < x_- < \theta_g\) in bounded replacement equilibrium.

**Proof.** Consider a bounded replacement equilibrium. Then \(\sigma_g > \sigma_b\). Recall (4), which states that retention is strictly optimal if

\[
(\sigma_g^2 - \sigma_b^2) x^2 + 2(\sigma_b^2 \theta_g - \sigma_g^2 \theta_b) x + (\sigma_g^2 \theta_b^2 - \sigma_b^2 \theta_g^2 + 2A \sigma_g^2 \sigma_b^2) > 0,
\]

where \(A = \ln(\beta_{\sigma_b/\sigma_g})\), and replacement is strictly optimal if the opposite strict inequality holds. Putting \(x = \theta_b\) in (37) and simplifying, we see that replacement is strictly optimal at \(\theta_b\) if

\[
\beta < \frac{\sigma_g}{\sigma_b} \exp\frac{\Lambda^2}{2\sigma_g^2},
\]

but this is guaranteed by the right hand inequality of (17), because \(\sigma^* \geq \sigma_g > \sigma_b\). Therefore \(\theta_b\) lies in the interior of the replacement zone, or put another way, \(x_+ < \theta_b < x_-\).
Now putting \( x = \theta_g \) in (37) and simplifying, we see that retention is strictly optimal at \( \theta_g \) if
\[
\frac{\Lambda^2}{2\sigma_b^2} + \ln(\sigma_b) - \ln(\sigma_g) > -\ln(\beta).
\]
(38)

The derivative of the left hand side of (38) with respect to \( \sigma_b \) is given by
\[
\frac{1}{\sigma_b} \left( \frac{1 - \frac{\Lambda^2}{\sigma_b^2}}{1} \right)
\]
which is strictly negative given (18) and \( \sigma_b \leq \sigma^* \), so it follows that the left hand side of (38) is minimized by setting \( \sigma_b = \sigma_g = \sigma^* \). To establish (38), then, it is sufficient to have
\[
\frac{\Lambda^2}{2\sigma^2} \geq -\ln(\beta),
\]
but this is guaranteed by the left hand inequality of (17). Consequently, the principal strictly prefers to retain the agent if she observes \( x = \theta_g \). Given \( x_+ < \theta_b < x_- \), this can only mean that \( x_- < \theta_g \), and the proof is complete.

**Proof of Proposition 4.** In a bounded replacement equilibrium, \( \sigma_g > \sigma_b \) and \( x_- > x_+ \). By Lemma 12, \( \theta_g \geq x_- \geq \theta_b > x_+ \). Define \( B_k(\sigma) \) to be type-\( k \)'s marginal benefit of noise:
(39)
\[
B_k(\sigma) := \phi\left( \frac{x_- - \theta_k}{\sigma} \right) - \phi\left( \frac{x_+ - \theta_k}{\sigma} \right).
\]
Observe that for every \( \sigma \),
\[
B_b(\sigma) = \phi\left( \frac{x_- - \theta_b}{\sigma} \right) - \phi\left( \frac{x_+ - \theta_b}{\sigma} \right) \geq \phi\left( \frac{x_- - \theta_g}{\sigma} \right) - \phi\left( \frac{x_+ - \theta_g}{\sigma} \right) = \phi\left( \frac{x_- - \theta_g}{\sigma} \right) \frac{x_- - x_+}{\sigma^2} > \phi\left( \frac{x_+ - \theta_g}{\sigma} \right) \frac{x_+ - x_+}{\sigma^2} = B_g(\sigma),
\]
(40)
where the first inequality follows from \( x_- \geq \theta_b \) and inequality (36) of Lemma 11, the second inequality follows from \( \phi \) single-peaked around zero and \( x_+ - \theta_g < x_+ - \theta_b < 0 \), and the last inequality follows from \( x_- \leq \theta_g \) and (again) inequality (36) of Lemma 11.

But (40) leads to the following contradiction: if the marginal benefit of noise for the bad type strictly exceeds that for the good type at every noise level, then by a simple single-crossing argument, we must have \( \sigma_b > \sigma_g \). But by Observation 1, this contradicts the fact that we are in a bounded replacement equilibrium. 

\[ \blacksquare \]
Proof of Propositions 5, 6 and 7. See the Supplementary Appendix.

References


