Too Good To Be True?
Retention Rules for Noisy Agents
Francisco Espinosa † Debraj Ray†.

November 2019

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 various extensions, including non-normal signal structures, non-binary types, interacting agents, costly mean-shifting, dynamics with term limits, and principal commitment. We discuss applications to risky portfolio management and political risk-taking, and mention several others.

1. INTRODUCTION

An agent who privately knows his type (good or bad) 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 centered on his type. He can choose to amplify or reduce the precision of this process. But there are 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 and occasionally qualify, 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.”

Because our framework is so stark and minimal, it lends itself easily to extensions and applications. A leading example 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

†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. Author names are in random order, following Ray ‡ Robson (2018)
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. The same is true of 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. Section 7 will discuss these applications.

Our baseline setting is one in which 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. The question is how this translates into sets of signal realizations for which the principal retains. Our initial discussion and Proposition 1 together argue that there are three types of potential equilibria. The first is monotone retention, in which the two agents choose the same level of noise, and the principal retains if the signal is above some threshold. The second is bounded retention, in which 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 first main result is Proposition 2, which singles out just the bounded retention equilibrium when the ambient noise level is sufficiently low. This is the finding that we seek to examine in various extensions, but from a theoretical and methodological perspective, a core contribution is the analysis of a minimal setting where standard single-crossing arguments fail. This failure is built into noisy signaling: spreads dominate means for extreme signal realizations, with the result that replacement and retention become dependent in non-monotone ways on signal realizations. Typically, this makes signaling games hard to analyze, but we believe we provide a simple formulation that offers just enough structure to be rich and yet tractable. While the failure of single-crossing does not fully explain our central finding, the bounded retention equilibrium is fundamentally based on it.

We study several extensions, with the objective of assessing the robustness of our main finding, and also to accommodate various ancillary features. Section 6.1 evaluates whether the ability to commit to a rule changes our result. Proposition 3 answers this question in the negative. This result suggests that the kind of environment we analyze, and the signaling problem we study in that environment, cannot be obviously ameliorated by any form of regulatory intervention, which we identify with the ability to commit.

Section 6.2 extends our baseline to a dynamic version of the model with agent term limits, in which the principal’s outside option from a new agent is endogenously determined. Such dynamics endogenize certain variables which we treated as parameters earlier, such as the prior on agent types in future hires. Proposition 4 shows that this generates our restriction on the ambient noise level as an outcome, thereby strengthening Proposition 2 even further.

Section 6.3 studies situations in which the agent can shift the mean of their signal, presumably at an additional cost, and the principal intrinsically values such effort. The deliberate injection
of ambient noise by the principal can act as a valuable commitment device, by structuring the environment in a way that keeps agent effort high. Section 6.4 replaces the normality restriction by signal structures that satisfy a strong version of the monotone likelihood ratio property. Our results survive (Propositions 5 and 6). Section 6.5 considers multiple agents, each with privately known type. Proposition 7 shows yet again that the main features of the bounded retention equilibrium survive, and that monotone retention cannot be an equilibrium.

Section 6.6 introduces a smooth, convex cost of noise. This is a major extension of the baseline model. Best responses by the agents now acquire a rich and complex structure, and the analysis becomes more nuanced. First, Proposition 8 shows that monotone retention is generically never an equilibrium with or without any restriction on ambient noise. This strengthens the baseline model. On the other hand, bounded replacement now becomes a real possibility, along with the bounded retention equilibria identified earlier. Our main results (Proposition 9 and 10) give conditions under which bounded retention equilibria must exist and also conditions under which no other type of equilibrium exists. These use novel fixed-point arguments that may be of separate interest. Finally, Section 6.7 studies non-binary agent types. The failure of monotone retention regimes is robust to such considerations (Proposition 11).

All in all, it is fair to say that in these extensions and variations, our basic observations survive unscathed. We end the paper by discussing three potential applications; see Section 7.

2. RELATED LITERATURE

While our main results are (to our knowledge) 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 example nothing binds the sender, because talk is cheap. In contrast, as explained above, our chosen communication structures must have mean equal to the true state, and the choice of structure could be costly. It is central to the analysis that each individual chooses a distribution over signals, rather than an announcement, and cannot hide the outcome ex post.

The choice of an information structure is central to Bayesian persuasion; see Kamenica and Gentzkow (2011). But neither sender nor receiver knows agent 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. We return to the question of observability in some detail, first in Section 5.3, and then again in Section 7. These approaches are complementary, and generate their own distinctive features.2

1In 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).

2There are also models of unobserved precision choice with no player types; see, e.g., Penno (1996) on financial reporting.
Dewan and Myatt (2008) examine a model of leadership in which an individual’s clarity in communication is a virtue, in that it attracts attention and thereby generates influence. But clarity also requires lower processing time from the audience, leaving more time for the audience to listen to others. Therefore zero noise is not chosen, because a leader might wish to hold on to an audience for longer, effectively dissuading them from listening to others.

Edmond (2013) also studies the obfuscation of states (say by a dictatorial regime). While such obfuscation occurs through the shifting of the mean signal with the use of a costly action, he also considers the case in which the state is communicated in a deliberately noisy way, with mean unchanged. The noise prevents coordination by receivers against the interests of the regime. Edmond 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.

Harbaugh et al. (2016) study the inclinations of a sender to distort the news about multiple projects, depending on the overall realization of news. By distorting the news from bad projects when the overall news is good, and by exaggerating the news from good projects when the news is bad, the sender effectively adjusts the realized spread of multidimensional news over multiple projects in opposite directions, depending on mean realizations. Such distortions are separate from mean-preserving noisy announcements; moreover, the focus is on realized spread. The results we develop are entirely distinct, but they too take note of a different “too-good-to-be-true” inference problem, whereby a posterior update reverts more strongly towards the prior for certain distributions when an extreme signal is received. Such extremeness (relative to the other components) is therefore eschewed by the sender when the mean news is good.

Hvide (2002) studies tournaments with moral hazard where two risk-neutral agents compete for a prize. The contractible variable is output, which is the result of their effort and a random component. A risk-neutral committee wants to ensure that agents exert high costly effort. If agents can costlessly increase noise in the random component of output (assumed to be normally distributed), rewarding the agent with the highest realization of output will lead to an equilibrium with low effort and high noise. If agents are rewarded depending on who gets closer to some pre-stipulated, finite level of output, a high effort low noise equilibrium is achieved. Less related are Palomino and Prat (2003) and Barron et al. (2017), who also study situations in which agents can inject noise into a moral hazard setting.

Finally, there is a literature on policy uncertainty (see, for example, Shepsle, 1972; Alesina and Cukierman, 1990; Glazer, 1990; Aragones and Neeman, 2000; Aragones and Postlewaite, 2002; Aragones, Palfrey and Postlewaite, 2007), often referred to as “strategic ambiguity.” Candidates offer policy platforms which can be more or less ambiguous, and this ambiguity generates uncertainty about the policies the candidate could implement were she to win the election. (An empirical analysis of strategic ambiguity can be found in Campbell, 1983.) Ambiguity here is

---

3In Palomino and Prat (2003), an agent manages a portfolio for a principal but can hide part of the return, which forces monotonicity of any optimal contract. Barron et al. (2017) study contracts that are immune to risk-taking, thereby forcing concavity of agent payoff with respect to produced output before the noise is added. A similar theme is also present in the endogenous risk-taking model studied in Ray and Robson (2012).
the result of the trade-off faced by the candidate between winning the election and implementing a certain policy (either his ideal policy or the most expedient one).

3. The Model

3.1. A Baseline Setting. An agent works for a principal. The agent can be good \((g)\) or bad \((b)\). He knows his type. The principal doesn’t. She has a prior probability \(q \in (0, 1)\) that the agent is good.\(^4\) 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 presumably 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 in any way 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, \(\textit{that is chosen by the agent}\). That is, the agent cannot shift the mean of his signal (though see Section 6.3), but he can modulate its precision. The principal does not observe \(\sigma_k\), but observes the realization of the signal. She then decides whether to retain or replace the agent.

Our baseline setting assumes that the choice of noise is costless up to a maximum level of precision. That is, we assume that there exists \(\sigma > 0\) such that \(\sigma_k \geq \sigma\) for \(k = g, b\). Any choice smaller than \(\sigma\) is impossible. 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). It can be interpreted as the “ambient”, or “natural” level of noise in the principal-agent interaction. For instance, in the portfolio management example, it is a minimum amount of risk that the money manager cannot hedge against. In the example of political campaigns, \(\sigma\) reflects the existence of minimal disruptions that interfere with the transmission or interpretation of the information from the speaker to the listener. Section 6.6 replaces this baseline specification with a cost function for the choice of noise.

Because \(V\) is the payoff to the principal from non-retention, the variable \(p \in (0, 1)\), defined by

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

is interpretable as an “outside option probability” that leaves the principal indifferent between retaining and replacing. How might this compare with \(q\), the prior probability that the agent is good? A salient benchmark is \(p = q\); we refer to this as the \textit{balanced model}. But there may be systematic departures of \(p\) from \(q\). Notice that \(V\) incorporates the option value of dealing with a new agent, so in a dynamic context, \(p\) should not be smaller than \(q\), and may well be strictly

\(^4\)On multiple types, see Section 6.4.
larger. Call this a model with an \textit{optimistic future}. If, on the other hand, 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 \textit{pessimistic future}. We allow for all three cases for now, though in a simple dynamic extension of our model with term limits, in which \( V \) is endogenous (Section 6.3), we will be able to narrow these alternatives down.

Our model is deliberately minimal, but several extensions and variations are described in Section 6. A core assumption is that the signal structure is (at least partially) unobservable; for more discussion see Section 5.3.

3.2. \textbf{Equilibrium}. As already mentioned, the principal does not observe the agent’s choice of noise, just some realization or signal \( x \) with distribution \( N(\theta_k, \sigma_k^2) \). The principal 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

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

Simple algebra involving the normal density yields the equivalent expression

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

where \( A := \ln \left( \frac{\beta \sigma_b}{\sigma_g} \right) \). The inequality (4) defines a \textit{retention regime}, a zone \( X \) of signals for which the principal will want to retain the agent. An \textit{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,

\[
\sigma_k \in \arg \max_{\sigma \geq 2} \int X \frac{1}{\sigma} \phi \left( \frac{x - \theta_k}{\sigma} \right) \, dx.
\]

Throughout, we work with pure strategy equilibria — where it is understood, of course, that the choice of noise is a pure action. We conjecture, but have not proved, that allowing for mixing will not change the results. The reasoning behind this conjecture is available on request.

4. \textbf{Retention Regimes}

4.1. \textbf{Trivial Retention Regimes}. Two examples of retention zones are (a) “always retain,” so that \( X = \mathbb{R} \), and (b) “always replace,” that is, \( X = \emptyset \). As far as equilibrium regimes are concerned, these are of little interest. Both generate complete indifference across the two types

---

\(^5\)Let \( V \) be the equilibrium value of restarting an interaction in an infinite horizon setting, normalized by a discount factor \( \delta \). Assume the principal gets utility from the agent in every period, though payoffs cannot be used as signals. Once an agent is replaced, the principal gets \( V \) again. By (1), we have \( V = pU_g + (1 - p) U_b \). However, since "replace the agent no matter what" is a feasible move for the principal at any date, we also have \( V \geq (1 - \delta)[qU_g + (1 - q) U_b] + \delta V \) when our agent is a new hire, which implies that \( p \geq q \).
as to the noise regime. With some cost — however small — to modulate the precision of the noise away from its “natural” ambient level, the agents would best-respond by minimizing such cost, so $\sigma_g = \sigma_b$ in such an equilibrium. But then the expression in (4) must alter sign over different values of $x$, eliminating either regime. (We will explicitly introduce such costs in Section 6.6.)

If we take the zero cost assumption literally, trivial retention regimes are possible. To make them work as equilibria, we would need to take full advantage of agent indifference regarding the choice of $\sigma$, making them choose the right combination of noise levels that would make the principal willing to always take the same decision, irrespective of $x$. Even so, “always retain” necessitates $q > p$, and “always replace” necessitates $p > q$, and in the balanced case with $p = q$ trivial regimes are always eliminated even when the cost of noise is zero.6

All in all, we largely ignore such trivial — and delicately supported — regimes. Not only are they eliminated by cost perturbations, but by other variants as well; see, e.g. Section 6.2.

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 $x^*$.7 See Figure 1. A monotone retention regime arises (and can only arise) when both types transmit with the same noise $\sigma_b = \sigma_g = \sigma$.8 Then (4) reduces to the condition

$$(5) \quad x \geq x^*(\sigma) := \frac{\theta_g + \theta_b}{2} - \frac{\sigma^2}{\theta_g - \theta_b} \ln (\beta),$$

and in particular, the retention zone in a monotone equilibrium must be of the form $X = [x^*, \infty)$. 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. In fact,

---

6Recall that the principal compares two posterior likelihoods to make her decision. But in the balanced case it cannot be that one of these likelihoods is always below the other, irrespective of signal value.

7Whether $x^*$ is included on one side or the other doesn’t matter.

8To see this, recall the retention condition (4), and notice that if $\sigma_g \neq \sigma_b$, then the resulting retention regime is either trivial or non-monotone.
this is the exact interpretation of \( x^*(\sigma) \) in the balanced model with \( p = q \), for then \( \beta = 1 \) and
\[
x^*(\sigma) = \frac{\theta_g + \theta_b}{2},
\]
which is the mid-point between the two means. Notice that \( x^* \) is entirely insensitive to \( \sigma \) in this case. With \( p = q \), the decision to retain is just a matter of comparing two likelihoods, and Panel A of Figure 1 shows that the likelihood for the good type dominates to the right of \( (\theta_g + \theta_b)/2 \). However, 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 the optimistic future setting, we have \( \beta < 1 \), and better performance is required for the principal to retain the current agent; \( x^*(\sigma) \) is higher for each \( \sigma \) as \( \beta \) falls. Panel B of Figure 1 depicts the consequences of an optimistic future, pushing \( x^*(\sigma) \) to the right of the midpoint between \( \theta_b \) and \( \theta_g \), and possibly even to the right of \( \theta_g \).

 Might a monotone regime ever be an equilibrium? Consider any retention threshold between \( \theta_b \) and \( \theta_g \). (In the balanced case, it is given by \( (\theta_g + \theta_b)/2 \) as already discussed.) Faced with such a threshold, it is easy to see that 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.

 Might a monotone retention threshold lie beyond the interval \([\theta_b, \theta_g]\)? Suppose that \( x^* \) lies to the left of \( \theta_b \). In this case, both the bad and the good type want to reduce the noise to its minimum: they both play \( \sigma = \sigma_b \). That is consistent with a monotone retention rule, but there is no guarantee that the resulting best-response threshold will indeed lie below \( \theta_b \). The formal analysis below shows that when \( \sigma \) is small enough, the resulting threshold must instead lie above \( \theta_b \), destroying the putative equilibrium.\(^9\) Finally, no monotone equilibrium retention threshold can lie to the right of \( \theta_g \). For suppose there were; say for instance, Panel B of Figure 1 applies. Then both types will want to add noise: no best response exists.\(^{10}\)

4.3. **Non-Monotone Retention Regimes.** When agents of 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, \( \sigma_b > \sigma_g \). The signal densities cross precisely twice. The two intersections, marked \( x_- \) and \( x_+ \), represent realizations for which there is no Bayes update on the prior. Above \( x_+ \) and below \( x_- \), the likelihood is higher that the type is bad, and the principal replaces. Between \( x_- \) and \( x_+ \), she retains the agent. In Panel B, on the other hand, the bad type transmits at lower noise than the good type, and the retention rule is flipped. Now

\(^{9}\)For small \( \sigma \), there is a high degree of “separation” between the two types. To see this, consider the signal \( x = \theta_b \). At this value, the likelihood of the bad type relative to the good type explodes as \( \sigma \) goes to 0. This makes \( x^* \) shift to the right, until it goes above \( \theta_b \). Now the equilibrium falls apart.

\(^{10}\)If there is some exogenous upper bound on the extent of noise, then it is possible to have a monotone equilibrium with both types choosing noise at the upper bound.
replacement occurs in some bounded interval of signal realizations, but elsewhere the principal actually retains. We make these observations more formal in Proposition 1 below.

The reason that there are just these two possibilities, but no more, is evident from (4). Retention or replacement zones are demarcated by values of the signal that solve a quadratic equation, which has at most two real roots. The absence of distinct real roots is indicative of a trivial “always-retain” or “always-replace” regime. Otherwise, when there are two real roots, one of the two zones of retention or replacement must be a bounded interval.

Now, the quadratic criterion for replacement or retention is a feature of the normal distribution, so we won’t make too much of it. It is perhaps possible that with more general signal distributions, there is alternation between replacement and retention. But the general point is that one of the two decisions must be guided by a bounded zone of signals (see Section 6.4 for more).

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. The zone of retention can then always be thought of as the set of signals going from \(x_-\) to \(x_+\) in a clockwise direction, considering that one can fold the real line on itself in a circle so that the ends \(-\infty\) and \(+\infty\) are identified with each other. Obviously, \(x_+\) and \(x_-\) are the two roots of (4), which means that

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

represents the equalization of weighted likelihoods for the good and bad 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

\[
\beta \frac{1}{\sigma_g^2} \phi' \left( \frac{x - \theta_g}{\sigma_g} \right) > \frac{1}{\sigma_b^2} \phi' \left( \frac{x - \theta_b}{\sigma_b} \right),
\]

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

\[
\beta \phi \left( \frac{x_- - \theta_g}{\sigma_g} \right) \frac{x_- - \theta_g}{\sigma_g^3} - \phi \left( \frac{x_- - \theta_b}{\sigma_b} \right) \frac{x_- - \theta_b}{\sigma_b^3} < 0.
\]
Likewise, the weighted likelihood for the good type must have a lower slope in $x$ relative to that for the bad type, evaluated at $x_+$, so that

$$
\beta \phi \left( \frac{x_+ - \theta_g}{\sigma_g} \right) \frac{x_+ - \theta_g}{\sigma_g^3} - \phi \left( \frac{x_+ - \theta_b}{\sigma_b} \right) \frac{x_+ - \theta_b}{\sigma_b^3} > 0
$$

Begin by using (6) for $x = x_-$ in equation (7) to obtain

$$
(\sigma_b^2 - \sigma_g^2) x_- < \sigma_b^2 \theta_g - \sigma_g^2 \theta_b.
$$

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

$$
(\sigma_b^2 - \sigma_g^2) x_+ > \sigma_b^2 \theta_g - \sigma_g^2 \theta_b.
$$

Combining these two inequalities, we must conclude that

$$
(\sigma_b^2 - \sigma_g^2) (x_+ - x_-) > 0
$$

in any non-monotonic equilibrium. This formalizes the above discussion as:

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

Proposition 1 will recur at different points in this paper. For now, we turn to our baseline result, which isolates a bounded retention regime — with two thresholds — as the plausible candidate for equilibrium.

5. **The Equilibrium Retention Regime is Bounded**

Our main result, that we extend in several directions, is that there is a unique nontrivial equilibrium if the ambient noise $\sigma$ is sufficiently small. In this equilibrium the good type chooses low noise, the bad type chooses high noise, and the principal employs a *bounded* retention zone. She is suspicious of both bad signals and excessively good signals, and replaces the agent in both cases. In short, she follows the maxim: “If it seems too good to be true, it probably is.”

5.1. **Small Ambient Noise.** We begin by describing the bound on $\sigma$. Recall that $\beta = \frac{1 - p}{p} \frac{q}{1 - q}$.

If $\beta \in (0, 1)$ (that is, if $p > q$ so that we have 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:

$$
\frac{\sigma}{\theta_g - \theta_b} < \frac{1}{2} \alpha (\beta)^{-1} \text{ if } 0 < \beta < 1.
$$

But we impose a different condition in case we have a pessimistic future, with $\beta > 1$:

$$
\frac{\sigma}{\theta_g - \theta_b} < \left[ \sqrt{2 \ln (\beta)} \right]^{-1} \text{ if } \beta > 1.
$$
Notice how these conditions become progressively weaker as we converge to the balanced case from either direction (i.e., as $p$ and $q$ get close to each other). At or near the balanced case, no restrictions are imposed at all; both right-hand side terms in (11) and (12) diverge to infinity. We will return to this feature. In particular, 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.

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

**Proposition 2.** (i) A nontrivial equilibrium exists if and only if (11) is satisfied, and when it exists, it is unique.

(ii) If (12) is also satisfied, then 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 particular, in the balanced case or with an optimistic future, (12) trivially holds and the unique nontrivial equilibrium must involve bounded retention.

We relegate a formal proof to the Appendix, but discuss the proposition here. The following observation is useful: when an agent’s type falls in any equilibrium retention zone $X$ — which is to say, when a signal realization equal to his type is an element of that zone — the agent wants to minimize noise in order to maximize his chances of “staying inside” the retention zone. On the other hand, when $\theta_k \notin X$, the actual structure of the retention zone informs the agent’s best response. For instance, in a monotone retention regime, the agent will want to inject infinite amount of noise, shifting signal mass to the tails of the distribution. In a bounded replacement regime, both signal tails belong to the retention zone, so once again infinite noise is the preferred choice. Finally, with bounded retention, there always exists a unique $\sigma_k$ that maximizes type-$k$’s chances of retention. With costs of noise (Section 6.6), these unrealistic observations about infinite noise will need qualification, but as we shall see, with the essential ideas preserved.

With the above in mind, it follows that bounded replacement cannot be an equilibrium outcome. By Proposition 1, such a regime requires $\sigma_g > \sigma_b$, but as argued above, bounded replacement generates infinite noise by every type outside the retention region, or equal noise if there is an upper bound on that variable (see Online Appendix for more details). So this possibility is eliminated, and both types must lie inside the retention region, but then again by the observations in the previous paragraph, this leads to $\sigma_b = \sigma_g = \sigma$ and again contradicts Proposition 1.

We conclude that only monotone and bounded retention can be nontrivial equilibria, an observation we critically re-examine in our extensions. That said, nontrivial equilibria might not always exist. The proposition tells us that this can only happen when $p > q$ and condition (11) fails.

\[\text{11The non-existence of bounded replacement survives more robust arguments which allow for a finite upper bound to the choice of noise. See Online Appendix for more details.}\]
Loosely speaking, ambient noise is so large relative to the distance between types, and the principal so optimistic about future agents, that she can never be convinced to retain the current agent.\textsuperscript{12} In this case, the only outcome is a trivial equilibrium: always replace.\textsuperscript{13}

The remainder of our discussion assumes that (11) holds. Suppose first that \( p \geq q (\beta \leq 1) \); then (12) is irrelevant. In this optimistic-future scenario, the principal is inclined ex-ante to replace the agent. In a monotone equilibrium, such an inclination implies that the (single) threshold \( x^* \) exceeds \( (\theta_g + \theta_b)/2 > \theta_b \). Then the bad type wants to increase noise to an unlimited degree, while the good type could go either way. But in either case, monotone retention is eliminated, and we are therefore left only with bounded retention. And indeed, with (11) satisfied, a bounded retention zone can be shown to exist, and coherence between the principal’s conjecture and the agent’s best responses is possible.

In the pessimistic-future scenario, condition (11) is vacuous, and a nontrivial equilibrium always exists. Now condition (12) must be invoked. If it is satisfied, then \( \sigma \) is small, and the smaller it is, the more sharply is the receiver able to distinguish between good and bad types (for more, see footnote 9). In this case, the two senders choose distinct levels of noise, and the monotone equilibrium is eliminated yet again.\textsuperscript{14}

In summary, we are left with the case of central interest: an asymmetric equilibrium in which the two types of agents choose different noise levels, the bad noise higher than the good. In this equilibrium, 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 trump (pun assuredly unintended) the lower mean in determining relative likelihoods.\textsuperscript{15}

Proposition 2, and its several variants to come, incorporates a minimal setting in which a basic single-crossing property is missing, and yet the resulting model is simple, tractable and quite rich. Under standard assumptions such as the monotone likelihood ratio property, two signal distributions with the same variance but different means will generate higher likelihood ratios in favor of the larger mean as the signal realization gets larger. But if the two distributions differ on account of mean and variance, this single-crossing property fails. For extreme realizations, the likelihood ratio is loaded in favor of the distribution with the larger variance, while

\textsuperscript{12}As we’ve remarked already, this state of affairs is absurd: if both types are let go, where would the optimism regarding a new agent come from in the first place? We formalize this in Section 6.2, when we endogenize \( p \).

\textsuperscript{13}The full argument is more subtle, for even if \( \sigma_g = \sigma \), it is not true that \( \text{any } \sigma_g \) makes the principal want to replace the agent. But every time the principal conjectures some \( \sigma > \sigma \) that makes her want to retain for some bounded interval \([x_-, x_+]\), the best response of the bad type of agent is inconsistent with the principal’s conjecture. For the full argument, the reader is invited to examine the formal proof of the proposition.

\textsuperscript{14}For completeness, we must observe that if (12) fails, then it is possible to have a monotone equilibrium. In it, the retention threshold falls below \( \theta_b \). Faced with that low threshold, \textit{both} types will want to reduce noise to the minimum possible, and this symmetry of choice resurrects the possibility of a monotone regime. This possibility will be eliminated when there is a cost of choosing noise.

\textsuperscript{15}Of course, the principal also distrusts signals that are bad: after all, lower mean and higher variance are particularly synergistic in producing lower signals.
for moderate realizations, the opposite is true. This failure generates a corresponding absence of single-crossing at the level of action choice. Specifically, low types might choose a larger variance in a bid to convince the sender that she is of higher mean. Combining this observation with the earlier argument, we see that the relative likelihood for the low type is actually enhanced under extreme realizations. The argument is delicate because the sender understands the previous sentence, and so dislikes extreme signal realizations. Nevertheless, the low type continues to choose higher noise, so that the equilibrium involves bounded retention.

5.3. A Remark on the Unobservability of Noise. It is important, especially in the context of potential applications, to understand our presumption regarding the non-observability of noise. At one level, it is a non-assumption in the following sense. Suppose that the choice of noise — the variance in our setting — were indeed observable. Then notice that every type must choose the same noise level, otherwise separation would be achieved via the choice itself, and there is absolutely no need to study the signal. The bad type would be out before he had a chance to redeem himself. So, with observable noise that’s costlessly chosen, agent types must choose the same 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).

If risk choices are costly, as they will be in Section 6.6, there is no change in the above argument if the cost function for noise is the same for both types — again, there must be pooling in observed components. But 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, parallel to a Spencian model of costly education. The actual signal realization is of no consequence. We return to this discussion in Section 7.

If observable noise is pooled, it is easy to augment the setting with unobserved noise. Specifically, suppose that the signal is still real-valued, but there are two components of risk: an observed component, as just discussed, and additional variance, or perhaps reductions of variance, that are not observed by the principal. The two components combine to create a single normal signal. With this augmentation, we are fully in our model. The observed components — endogenous though they may be — must be chosen identically by all types. The exact value is immaterial, but with any degree of costly noise, as in Section 6.6, it would simply be set to the minimum level, provided we are willing to refine the equilibrium using reasonable beliefs such as the intuitive criterion. With or without such refinements, the remainder of the analysis can then be conducted with no change in results.

To summarize, then: (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, more generally, is uncorrelated with the agent type; (c) the same results as in our paper would apply to the unobserved components of risk, but (d) observable risk could play a role in separation if the costs of risky choice are correlated with type — in such situations, the actual outcomes of the

---

16For 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.
risky choice would play no additional role. The singular nature of Proposition 2 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

In this Section, we describe several variations on the model. Section 6.1 evaluates whether the principal can benefit from committing ex-ante to retention rules. Section 6.2 analyzes a dynamic version of the model with agent term limits, in which the principal’s outside option from a new agent is endogenously determined. Section 6.3 studies situations in which the agent can shift the mean of their signal, presumably at an additional cost. Section 6.5 considers more than one agent, each with privately known type. Section 6.4 extends the normality assumption to more general signal structures that satisfy a strong version of the monotone likelihood ratio property. Section 6.6 introduces a smooth, convex cost of noise. Section 6.7 studies non-binary agent types.

One could combine these different variations to get yet richer environments. Just as an example, we base Section 6.7 on the costly noise model of Section 6.6, and in the Online Appendix, we do the same for the extension to non-normal distributions. Indeed, the costly noise model could just as easily serve as an alternative benchmark for the paper, which is why we devote some extra attention to it.

6.1. Commitment. To what degree are the baseline results altered if the principal can commit ex ante to a retention zone? It turns out that they are not affected at all. For similar observations in a different context (and for distinct reasons), see Glazer and Rubinstein (2004, 2006) and Hart et al. (2017).

Suppose that the signal realization $x$ is contractible, and that the principal announces an incentive-compatible mechanism that specifies the retention probability for each value of $x$, and for each (declared) type of agent. The agent can then choose one of the rules — possibly revealing his type — and then a noise level. We assume that the rule, given by $r_k(x) \in [0, 1]$ is piece-wise continuous.\(^{17}\) For any type $k$, rule $r$ and chosen noise $\sigma$, define

$$
\rho_k(r, \sigma) := \int_{-\infty}^{\infty} r(x) \frac{1}{\sigma} \phi \left( \frac{x - \theta_k}{\sigma} \right) dx,
$$

which is to be interpreted as the overall retention probability for type $k$ when the retention function is $r$ and he chooses noise $\sigma$. The principal seeks to maximize her surplus

$$
q \rho_g (r_g, \sigma_g) (U_g - V) - (1-q) \rho_b (r_b, \sigma_b) (V - U_b)
$$

by “choosing” $r_k$ and $\sigma_k$ for $k = g, b$, subject to

$$
\sigma_k \in \arg \max_{\sigma \geq \sigma_k} \rho_k (r_k, \sigma)
$$

\(^{17}\)We conjecture that Proposition 3 below is true for all measurable functions.
and

\[ \rho_k(r_k, \sigma_k) \geq \max_{\sigma \geq \sigma_0} \rho_k(r_\ell, \tilde{\sigma}) \]

for each \( k \) and \( \ell \neq k \). The first of these constraints is the familiar choice of noise, and the latter comes from truthful revelation of type. But notice that this latter constraint cannot be slack for type \( b \) at the optimum. If it were, the principal could simply reduce the retention probability \( r_b \)\(^{18}\) — which makes her happier (the expression in (13) goes up), continues to respect (15) for type \( b \), and does no damage to (14) and (15) for type \( g \).

We must conclude, therefore, that (15) binds for type \( b \); that is, \( \rho_b(r_b, \sigma_b) = \rho_b(r_g, \sigma'_b) \), where \( \sigma'_b \) maximizes \( \rho_b(r_g, \tilde{\sigma}) \). Using (13), this further implies that the principal is completely indifferent between type \( b \) reporting his type and facing \( r_b \), or misreporting his type and facing \( r_g \). So, without any loss of generality, the principal may as well offer the agent a single retention function \( r(x) \). That gives rise to a new problem with just one rule, no self-selection constraint (15) for either type, and just payoff maximization (14) for each type. To summarize this new problem, note that by definition of \( p \), \( V - U_b = (U_g - U_b) \, p \) and \( U_g - V = (U_g - U_b) \, (1 - p) \). Using these in (13), the principal equivalently maximizes

\[ \beta \rho_g(r, \sigma_g) - \rho_b(r, \sigma_b) , \]

where \( \beta \) is \( q(1 - p)/p(1 - q) \) as defined earlier, and where for each \( k = g, b \),

\[ \sigma_k \in \arg \max_{\tilde{\sigma}} \rho_k(r, \tilde{\sigma}) . \]

A single retention rule notwithstanding, there is still room for commitment, because the principal can influence the choice of noise. Yet in the context at hand, the principal has no use for it:

**Proposition 3.** Assume condition (11), so that a nontrivial equilibrium exists. Then an optimal contract involves the same retention function (a.e.) and the same values \( \sigma^*_b \) and \( \sigma^*_g \) as in the nontrivial equilibrium of Proposition 2.

So the solution to the principal’s problem with commitment is the same as the no-commitment or equilibrium solution, at least as far as the benchmark model is concerned. The intuition is based on the idea that the principal’s payoff can be expressed as a function of the payoffs of the agents. She wants to retain the good type, so these payoffs are perfectly and positively aligned. She wants to remove the bad type, so once again these payoffs are perfectly — though negatively — aligned. From any commitment solution that is not a best response, then, the principal can always profitably move in the direction of her best response by changing the contract by a tiny amount. The consequent responses of the agents will have no first-order effect on the principal’s payoff, by the envelope theorem. This intuition also suggests that the argument may not be robust to other extensions of the model in which the principal’s payoffs cannot be expressed in terms of the agents’ payoffs, but we have not explored these directions more fully.

\[^{18}\text{She can judiciously remove intervals where } r_b(x) > 0 \text{ to drive retention probability continuously from } \rho_b \text{ to } 0.\]
6.2. **Dynamics With Term Limits.** So far we have studied a static setting, but at the same time we’ve suggested that the “outside option probability” $p$ could, in principle, be solved for in a dynamic setting. We study the case in which the agent has a two-period “term limit,” after which he must be replaced. In what follows we study stationary equilibrium, in which every new agent of a given type takes the same action for the same value of $p$. Given the two-term limit, $p$ will be fully pinned down in stationary equilibrium, though of course it is an endogenous variable in this extension. As a consequence, agent actions will be pinned down as well.

For noise $\sigma_k$ for each player of type $k$, and for each realization $x$, the Bayes’ 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 $x$.

We can use this information to calculate the lifetime payoff to the principal at the start of any new interaction. To this end, let $M(q') := q'U_g + (1 - q')U_b$ be the expected payoff to the principal in any period when her prior (for that period) is given by $q'$. This prior equals $q$ for a fresh draw from the pool at any date. At the end of the first term, a signal $x$ is generated, and the prior $q$ is updated to $q(x)$. At this stage, the principal decides whether or not to retain for one more period, after which the term limit kicks in.

If $V$ denotes the normalized lifetime payoff to the principal starting from a fresh agent, we can define a retention zone $X$ as the set of all $x$ for which $(1 - \delta)M(q(x)) + \delta V \geq V$. The lifetime value to the principal can then be expressed as

\[
V = (1 - \delta)M(q) + \delta \int_X [(1 - \delta)M(q(x)) + \delta V] \pi(x)dx + \delta \int_{X^c} V \pi(x)dx
\]

\[
= (1 - \delta) [q(1 + \delta \Pi_g)U_g + (1 - q)(1 + \delta \Pi_b)U_b] + \delta [1 - (1 - \delta)\Pi] V,
\]

where $\Pi_k := \int_X \pi_k(x)dx$ is the type-dependent probability of retention, and $\Pi := q\Pi_g + (1 - q)\Pi_b$ is the overall probability of retention. (The second equality above follows from the definition of $M$ and (18).) 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

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

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

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

Now observe that in any equilibrium, $\Pi_g \geq \Pi_b$. That has to be the case, because the principal can — and will — choose a retention zone that retains the high type at least as often than the low type. Indeed, it is not even possible to have $\beta$ equal to 1 in any equilibrium.\(^{19}\)

\(^{19}\)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 $\beta = 1$.\]
This setup reveals a clear strategy to solve the two-term dynamic extension of our model. For some (provisionally given) value of $\beta$, we obtain the baseline static model. Solve for the equilibrium there. That equilibrium will generate retention probabilities $\Pi_g$ and $\Pi_b$. The circle is closed by the additional condition that $(\beta, \Pi_g, \Pi_b)$ must solve (19).

Formally, we obtain:

**Proposition 4.** When agents can be hired for up to two terms, and the principal always has the option to replace agents with a new draw from a stationary pool, there is a unique equilibrium which has all the properties of the non-trivial equilibrium identified in Proposition 2. In particular, there are no trivial equilibria. Moreover, this unique equilibrium must endogenously display an optimistic future and conditions (11) and (12) do not need to be assumed.

Proposition 4 says that in a dynamic extension with a two-term limit, the equilibrium picks out precisely the two-threshold equilibrium with bounded retention regime, as described in Proposition 2 of the static model. Observe that that equilibrium in the static model does not always exist; after all, $\sigma$ needs to be small enough as described in conditions (11) and (12). Those conditions are automatically satisfied here. So Proposition 4 is not just a mere refinement of the static equilibrium that eliminates all monotone and trivial equilibria. It does that, to be sure, but in addition it guarantees that for any value of $\sigma > 0$, the dynamically determined value of $p$ must adjust itself so that conditions (11) and (12) are automatically met.

6.3. **Mean-Shifting Effort, and Noisy Principals.** We can easily augment the baseline model to include unobserved effort to shift the mean value of one’s type. For instance, suppose that each agent $k$ is endowed with some baseline value (or type) $\theta_k$ (with $\theta_g > \theta_b$). He can augment $\theta_k$ using a cost function $d(\theta_k - \theta_k)$, common to both types, where $d$ defined on $\mathbb{R}_+$ is increasing, strictly convex and differentiable, with $d(0) = 0$. The signal sent is then given by $x_k = \theta_k + \sigma_k \varepsilon_k$. Finally, the principal makes a decision to retain or replace.

Parts of this model run fully parallel to our setting. The principal makes her decisions on the basis of conjectured means and variances chosen by each type, leading to the familiar conditions (6)–(8) for the retention edge-points $x_-$ and $x_+$. Similarly, an agent of type $k$ maximizes the probability of retention net of cost. Whether or not $x_-$ is smaller or larger than $x_+$ (and even when $x_+ = \infty$ as it will be with monotone retention), the agent always maximizes $\Phi ([x_+ - \theta_k]/\sigma_k) - \Phi ([x_- - \theta_k]/\sigma_k) - d(\theta_k - \theta_\kappa)$, this time by choosing both $\sigma_k$ and $\theta_k$. What this extension adds is a first-order condition for $\theta_k$, given by

$$
\frac{1}{\sigma_k} \phi \left( \frac{x_- - \theta_k}{\sigma_k} \right) - \frac{1}{\sigma_k} \phi \left( \frac{x_+ - \theta_k}{\sigma_k} \right) \leq d' (\theta_k - \theta_\kappa)
$$

with equality holding if $\theta_k > \theta_\kappa$. This additional condition can be used to show that the extension fully mimics the original model: we must have $\theta_b < \theta_g$, with other choices of noise and principal decisions just as in our baseline setting; see Online Appendix for details.

This extension is also useful for understanding other aspects of the noisy relationship between principal and agent. For instance, mean-shifting effort for the sake of retention could be directly
valuable to the principal, apart from providing information about type. If neither that effort nor the payoff-relevant “output” from it is contractible, then the principal could want to structure her environment to keep agent effort high. Of particular interest is the case in which the background noise $\sigma$ is close to zero, so that the agents can communicate their types with very high precision.

In general, this limit model has several equilibria, some pooling and some separating. To see the issue that arises, let’s concentrate on a particular parametric configuration in which $\theta_g$ and $\theta_b$ are sufficiently separated from each other so that

$$d (\theta_g - \theta_b) > 1.$$  

In this case it is easy to see that there can be only separating equilibria in zero-ambient-noise limit. In each such equilibrium, the bad type exerts no effort whatsoever. The principal cannot incentivize the agent because there is no noise in the signal. Both types reveal themselves perfectly. There are still many equilibria possible in which the good type is forced to exert effort to raise $\theta_g$ beyond $\theta_g$, simply because the principal’s retention set is some singleton $\{\theta_g\}$ with $\theta_g > \theta_g$. But these equilibria are shored up by the “absurd belief” that observations between $\theta_g$ and $\theta_g$ are attributable to the bad type. These configurations can be eliminated by standard refinements, leaving only the least-cost separating equilibrium in which retention occurs if $x = \theta_g$, and no agent exerts any effort at all. Condition (21) guarantees that the bad type will not want to mimic the good type in this case.

If mean-shifting effort is separately valuable to the principal, this outcome is undesirable to her. The solution will therefore involve the principal adding noise, thereby ensuring that the bad type has some chance of being retained, and so incentivizing him. In any equilibrium of such an extended model in which the principal can move first, the principal will choose $\sigma > 0$, endogenously injecting noise into the system.

6.4. Beyond Normal Signals. Consider a generalization of the benchmark model, in which 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 $\sigma \sigma > 0$, 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:

$$\tilde{f}(x|\sigma) = \frac{1}{\sigma} f \left( \frac{x - \theta_k}{\sigma} \right).$$

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 not only that this is the case, but that the relevant likelihood ratios become unbounded as the signal grows large. Formally, we assume:

\[\text{For other models of relational contracts in which effort provides both current output and information about match quality, see, Kuwalekar and Lipnowski (2018), Kostadinov and Kuwalekar (2018), and Bhaskar (2017).}\]
Strong MRLP. \( f(z - a)/f(z) \) is increasing in \( z \) whenever \( a > 0 \), with

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

In the context of our model, these end-point conditions a single, finite threshold for retention when both types use the same noise, no matter how optimistic or pessimistic the principal’s prior is regarding agent types; that is, for any \( \beta \in (0, \infty) \). The normal density satisfies strong MLRP.

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

\[
\beta f \left( -\frac{\theta_a - \theta_b}{\sigma(\beta)} \right) = f(0),
\]

for all \( \beta > 1 \), and set \( \sigma(\infty) = \infty \) otherwise. This function is well-defined because we place the peak of \( f \) at zero and because \( f(z) \to 0 \) as \( |z| \to \infty \).

The Online Appendix establishes the following two-part proposition:

**Proposition 5.** Consider any density function for the signal such that the strong MLRP condition is satisfied. Then:

(i) A bounded replacement equilibrium cannot exist.

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

The Online Appendix shows that under strong MLRP, whenever the two types of agents choose different levels of noise, the principal optimally responds by employing either bounded retention or bounded replacement zones. Specifically, 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 observation 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 \( \sigma_b < \sigma(\beta) \). 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 5 holds.

The question that remains is whether such an equilibrium exists, and (not as central but still of interest) whether that equilibrium is unique. It turns out — at least for the line of argument that

---

21To verify this, pick any \( z > 0 \) and use the fact that \( f(z - a)/f(z) \) is locally increasing at \( a = z \), along with \( f'(0) = 0 \) by symmetry around 0. A parallel argument works when \( z < 0 \).
we employ — that the interval structure of the retention regime is a helpful property. We obtain this property free of charge for some environments; specifically, in the optimistic future model with \( \theta \geq 1 \). However, this interval structure of the bounded retention regimes may not hold for all possible values of \( \theta \). The following proposition provides conditions under which we obtain this structure.

**Proposition 6.** Consider any signal density function satisfying strong MLRP, and with its single peak at 0. Assume furthermore that either \( \theta \geq 1 \), 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 \( \sigma_g = \tilde{\sigma} \), 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; that is, when \( \beta \leq 1 \), the condition \( \sigma \in (0, \hat{\sigma}) \) automatically holds, and the nontrivial equilibrium must involve bounded retention.

We observe that the convexity condition in the proposition is automatically satisfied in the baseline model with normal \( f \), so this result generalizes the baseline model. That said, it is an open question whether the convexity condition can be dropped free of charge.

### 6.5. Multiple Agents

We’ve assumed that there is a single agent of unknown type. Suppose there are two agents, 1 and 2, who simultaneously signal their types, and the principal must decide which agent to retain. She wants to retain the better agent — or one of them, if she is indifferent. This sort of structure brings us closer to a model of political campaigns.

Assume that it is common knowledge that only one of the two agents is good. The agents know their own types and therefore both types. But they look identical ex ante to the principal, so her prior places equal probability on the two. The communication technology is unchanged:

\[
(25) \quad x_i = \theta_{k(i)} + \sigma_{k(i)} \epsilon_i,
\]

where \( i = 1, 2 \), and \( k(i) \) denotes \( i \)'s type. The errors are independent and identically distributed standard normal random variables. In this game, by symmetry, a strategy for agent \( i \) is a function \( \sigma : g, b \to \mathbb{R}_+ \). As for the principal, a strategy is a function \( r : \mathbb{R}^2 \to \{1, 2\} \), which indicates for every possible pair of signals \((x_1, x_2)\) the agent she wants to retain. After observing \((x_1, x_2)\) the principal retains agent 1 if (and, modulo indifference, only if)

\[
(26) \quad \frac{1}{\sigma_g} \phi \left( \frac{x_1 - \theta_g}{\sigma_g} \right) \geq \frac{1}{\sigma_b} \phi \left( \frac{x_1 - \theta_b}{\sigma_b} \right) \quad \frac{1}{\sigma_g} \phi \left( \frac{x_2 - \theta_g}{\sigma_g} \right) \geq \frac{1}{\sigma_b} \phi \left( \frac{x_2 - \theta_b}{\sigma_b} \right).
\]

In this setting, a monotone equilibrium is defined as one where the principal retains the agent with the higher signal value. Once again, monotonicity can only be achieved if both types of agent play the same \( \sigma \), but that won’t happen.
Proposition 7. If an equilibrium exists, it can only be the case that $\sigma_b > \sigma_g$, and the principal retains agent 1 if and only if $|x_1 - \hat{x}| \leq |x_2 - \hat{x}|$, where $\hat{x} = (\sigma_g^2 \theta_g - \sigma_b^2 \theta_b)/(\sigma_b^2 - \sigma_g^2)$ is the signal value that maximizes the likelihood ratio $\frac{1}{\sigma_g} \phi \left( \frac{x - \theta_g}{\sigma_g} \right)/\frac{1}{\sigma_b} \phi \left( \frac{x - \theta_b}{\sigma_b} \right)$. In particular, monotone equilibria do not exist.

The proof of this proposition is long and involved, and we relegate it to the Online Appendix. Intuitively, when both types choose the same level of noise, the principal retains the one with the higher signal realization. But the bad type then wants to inject additional noise, since the good type has a lot of probability mass around his (higher) mean. At the same time, and for the same reason, the good type wants to decrease noise.

More formally, assume by way of contradiction that an equilibrium features $\sigma_b < \sigma_g$. Then the principal will respond by retaining the agent whose signal is further away from $\hat{x} = (\sigma_g^2 \theta_g - \sigma_b^2 \theta_b)/(\sigma_b^2 - \sigma_g^2)$, which is the value that minimizes the likelihood ratio $\frac{1}{\sigma_g} \phi \left( \frac{x - \theta_g}{\sigma_g} \right)/\frac{1}{\sigma_b} \phi \left( \frac{x - \theta_b}{\sigma_b} \right)$, and it is to the left of $\theta_b$. Then, it turns out that the bad type will want to inject additional noise, which means that this type wishes to deviate.

Proposition 7 bears a broad resemblance to the main result in Hvide (2002), who studies tournaments with moral hazard, when agents can influence both the mean and spread of their output. In equilibrium, there is excessive risk taking. By setting an intermediate value for output and rewarding the agent who gets closer to this threshold, the principal can do better.

6.6. Costly Noise. Consider a richer variant of our model, in which there is a cost to modulating precision: there is a strictly convex cost function $c(\sigma)$, with a minimum at $\bar{\sigma}$, with $c(\bar{\sigma}) = 0$, and $c(0) = c(\infty) = \infty$. That is, deviations from the baseline degree of noise are costly in either direction. The cost of excessive precision is perhaps self-explanatory, but it is also reasonable that high levels of $\sigma$, which involve both very good (and very bad) signals with positive probability, could be very expensive to implement. In effect, we assume that it is costly to disguise one’s true characteristics and intentions in an attempt to generate some chance that the evaluation will be positive. Call this the costly noise model.

An equilibrium is then a configuration $(\sigma_g, \sigma_b, X)$ such that given $(\sigma_g, \sigma_b)$, $x \in X$ solves (4) just as it did before, and given $X$, each type $k$ chooses $\sigma_k$ to maximize the probability of retention, net of noise cost:

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

This version of the model presents many interesting features and some deviations from the benchmark, so we will devote some attention to it. First, as already noted, trivial equilibria are not possible in this variant, with or without parametric restrictions (such as (11) in the benchmark setting). As argued in Section 4.1, if the retention regime is trivial, both types have an incentive to pool on the lowest cost signal, which in turn makes the signal informative, thereby implying that the optimal retention policy must be responsive to the signal, a contradiction. Second, monotone regimes are generically impossible in equilibrium. These are sharp improvements on the baseline model. But third, the model does open up the unexpected possibility of the existence of
bounded replacement equilibria, which were easily ruled out in the benchmark model. Finally, we can describe the best response of an agent facing a monotone or a bounded retention regime in a richer way than in the no-cost model, and such behavior appears to be consistent with empirical findings in the literature on risk-taking. Now we expand on some of these points, beginning with the agent’s best response mapping before moving to a fuller description of equilibrium.

6.6.1. The Agent’s Best Response. We already know that nontrivial equilibria can be either monotone or “bounded.” In each of these cases, type $k$ chooses $\sigma_k$ seeking to maximize $\Phi \left( \frac{x_+ - \theta_k}{\sigma_k} \right) - \Phi \left( \frac{x_- - \theta_k}{\sigma_k} \right) - c(\sigma_k)$. A monotone regime is equivalent to setting $x_+ = \pm \infty$ and $x_- = x^\star$. The necessary first-order conditions are

\begin{equation}
\phi \left( \frac{x_- - \theta_k}{\sigma_k} \right) \left( \frac{x_- - \theta_k}{\sigma_k^2} \right) - \phi \left( \frac{x_+ - \theta_k}{\sigma_k} \right) \left( \frac{x_+ - \theta_k}{\sigma_k^2} \right) = c' (\sigma_k)
\end{equation}

for each type $k = g, b$.

Optimally chosen noise now moves in a subtle and quite complicated way as a player’s type moves relative to the retention zone. Figure 3, Panel A, illustrates this for a monotone retention threshold. When a player’s type is outside the retention zone and far away from the threshold, it takes a large amount of noise to create a significant probability that a signal will be generated 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, noise increases, but reaches a maximum when the type is still some distance away. The easiest way to understand this is to think of what happens when the type is on the edge of the zone, at which point noise makes no difference to the chances of retention, so that the noise level is back to $\sigma$ again. Now continue the process by moving the type into the retention zone. Moving closer to the zone, noise increases, but reaches a maximum when the type is still some distance away. The easiest way to understand this is to think of what happens when the type is on the edge of the zone, at which point noise makes no difference to the chances of retention, so that the noise level is back to $\sigma$ again. Now continue the process by moving the type into the retention zone. In this case, noise can throw the player out of the zone, so she seeks to lower it. Her optimum choice therefore 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 again to $\sigma$, but this time from below.

With bounded retention zones, the choice function exhibits even more non-monotonicities.\(^{22}\) Panel B of Figure 3 shows that there will generally be five turning points. There is one each for

\(^{22}\)Formal details are available on request from the authors.
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
dges is below \( \sigma \), because the retention zone is bounded.)

This behavior of our agents in the costly noise model is consistent with empirical findings on
risk-taking behavior. First, consider a scenario where there is an exogenous monotone retention
threshold. Genakos and Pagliero (2012) finds that risk-taking in weightlifting contests exhibits
an inverted–U relationship between risk taking and rank, with the peak reached around rank 6.
Figueiredo et al. (2015) finds that risk-taking by portfolio managers is non-monotonic: managers
who are significantly below a compensation threshold reduce risk-taking relative to those who
are relatively close. These findings are consistent with our model predictions when agents are to
the left of the retention threshold (Panel A in Figure 3).

Chevalier and Ellison (1997) study risk-taking by mutual funds, and remark that: “The one
clear regularity in the data that is somewhat puzzling . . . is that higher excess returns are clearly
correlated with larger risk increases.” With costly noise, both monotone and bounded retention
can generate this type of behavior by the agents, both inside and outside of the retention zone.

To be sure, there are some technical complications that we did not emphasize in our discussion
above. The single-peakedness of the noise distribution generates a non-convexity in the agent’s
optimization problem, which raises the possibility that an agent’s choice could be multi-valued.
For monotone or bounded retention regimes, such multivaluedness is more a technical nuisance
than a feature of any economic import,\(^23\) 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.

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 Appendix shows that
a sufficient condition for [U] is

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

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

6.6.2. Monotone Retention is (Almost) Never an Equilibrium. In the costly noise model, an
equilibrium can involve monotone retention; see Online Appendix for a specific example. But
the example isn’t robust: in “almost all” cases, there is none:

**Proposition 8.** Generically, a monotone equilibrium can not exist in the costly noise model.
Specifically, given model parameters, there is at most one common value of \( \sigma \) that both players

\(^{23}\)For bounded replacement regimes, the possibility of multiple solutions is more natural. For instance, an agent
located in one of the two retention zones to the side, but close to the replacement zone, could be indifferent between
a small and a large choice of noise.
must choose in any monotone equilibrium, and this value is pinned down independently of the cost function for noise choice.

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 (27) becomes

\[
\phi \left( \frac{\sigma g}{\sigma_k} \right) \frac{\sigma g}{\sigma_k} - c' (\sigma_k) = 0,
\]

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 \( x^* \in (\theta_b, \theta_g) \), so once again, if the threshold lies strictly between the two types, the incentives for each type push in opposite directions away from \( \sigma \), and at least one of them will wish to deviate from \( \sigma \).

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

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

Equation (30) tells us that we will need to study the function \( \phi(z)z \); 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 relative to each other: they must both have the same sign and generate 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, \( \ln \beta > 0 \), so \( z_1 \) and \( z_2 \) are both 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 (30). It is entirely independent of the cost function for noise, and so the second equality cannot generically hold. (The Appendix formalizes the argument.)

6.6.3. Bounded Retention and Replacement Equilibria. With monotonicity out of the way, we are left with equilibria in which the two types choose different noise levels, and the principal
employs bounded intervals where retention or replacement occurs. We wish to examine the robustness of the bounded retention equilibrium, which plays a central role in our baseline model.

Throughout, we maintain Condition U, which guarantees unique best responses in the choice of noise (when retention is bounded), and is a technical restriction of little economic import. But we will now impose a substantive restriction. Recall the definitions of $\sigma_*$ and $\sigma^*$ from our discussion of Condition U, equation (28). These are lower and upper bounds on noise that straddle $\sigma$, with $c(\sigma_*) = c(\sigma^*) = 1$. Because agent payoff from retention is normalized to 1, and that from replacement to 0, no agent would ever transmit noise outside $[\sigma_*, \sigma^*]$. Now imagine that both agents transmit common noise equal to the upper limit $\sigma^*$. We know already that the principal would respond by choosing a single threshold $x^*(\sigma^*)$ for retention, described by equation (5).

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

The assumption states that when both types employ common noise $\sigma^*$, the weighted relative likelihood for the type being good or bad must flip sign at some signal value between $\theta_b$ and $\theta_g$. Condition T is automatically satisfied in the balanced case with $\beta = 1$, because in that case, as already observed, $x^*(\sigma^*) = (\theta_g + \theta_b)/2$. Moreover, we can write this condition as a set of restrictions on the extent to which $\beta$ can depart from 1 on either side of the balanced case. That is, we want the future to be neither too optimistic nor too pessimistic. Do this by subtracting the formula for $x^*(\sigma^*)$ — see (5) — from $\theta_b$ and then $\theta_g$ to obtain an equivalent form of [T]:

$$-\frac{(\theta_g - \theta_b)^2}{2\sigma^2} \leq \ln(\beta) \leq \frac{(\theta_g - \theta_b)^2}{2\sigma^2}. \tag{31}$$

Proposition 9. Under Conditions U and T, there is an equilibrium with bounded retention.

The proof carries some intuition, so we loosely outline it here. Begin by searching for any equilibrium via a fixed-point mapping. The very first box in Figure 5 delineates the domain of that mapping. No agent will choose noise below $\sigma_*$ or above $\sigma^*$, so we have a compact domain. The image of this mapping is derived as follows: for each $(\sigma_g, \sigma_b)$, find the retention decision of
the principal, shown in the middle graph (where \( x_- \) and \( x_+ \) are chosen), 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. A fixed point of this mapping will yield an equilibrium.

The problem is that this fixed point mapping is not well-behaved. For any point \((\sigma_g, \sigma_b)\) in the domain with \(\sigma_b < \sigma_g\), the planner will best-respond with bounded replacement, and the “subsequent” response that completes the mapping is generally not continuous in \((\sigma_g, \sigma_b)\). This discontinuity problem is endemic. Given that the retention region (under bounded replacement) is made out of separated zones, the choice of two or more noise levels that maximize retention probabilities is generally unavoidable. With that multiplicity in place, discontinuities in the fixed-point mapping are unavoidable. The simplest fixed-point approach is a dead end.

However, given our specific interest in the existence of a bounded retention equilibrium, we want to start from an even smaller domain, which is the shaded triangle in the left box, over which \(\sigma_b \geq \sigma_g\). This subdomain is better-behaved — the principal chooses bounded retention (or a monotone threshold) as a best response, and the best response by the agents to each such retention policy is unique (by Condition U) and therefore continuous. But now the problem is different: it may well be that the mapping slips out of the smaller domain. In general, this slippage cannot be controlled. In Panel B of Figure 3, introduced earlier, we have a bounded retention zone that could arise from some “starting” \((\sigma_g, \sigma_b)\) with \(\sigma_b > \sigma_g\). And yet in response, type \(g\) chooses larger noise as illustrated, which propels the system out of the triangle. This is mirrored by the lower pair of arrows in Figure 5.

At the same time, the mapping on the smaller domain has an interesting property. On the boundary between the two subdomains, the mapping “points inwards” whenever (31) holds. Look at the upper pair of arrows in Figure 5. The first arrow in the pair maps a point on the principal diagonal of the square (where \(\sigma_b = \sigma_g\) to a monotone retention regime; that is, \((x_-, x_+)\) is of the form \((x^*, \infty)\). By our restriction on \(\beta\) in condition (31), \(x^*\) must lie between \(\theta_b\) and \(\theta_g\). So the good type wants to reduce noise to remain within the retention zone, while the bad type wants to increase it. That means that the good type must choose noise \(\sigma_b < \sigma^*\), while the opposite is true of the bad type. But that implies a best response with \(\sigma_b > \sigma_g\), which takes us back into the starting subdomain from its boundary. (It also implies, in passing, that under condition (31), a monotone equilibrium cannot exist, whether generically or otherwise.) A fixed point theorem due to Halpern (1968) and 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, we have shown that 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

\[
\theta_g - \theta_b \geq \sigma^*,
\]

where recall that \(\sigma^*\) is defined by the larger of the two solutions to \(c(\sigma) = 1\).
Proposition 10. Assume that Condition (31) used in Proposition 9 holds, and so does (32). Then only bounded retention equilibria exist.

While the Appendix contains a formal proof, it is easy enough to illustrate the main argument. Consider the same fixed point mapping used to establish the existence of a bounded retention equilibrium. The first component of this mapping takes noise choices \((\sigma_g, \sigma_b) \in [\sigma_s, \sigma^*]^2\) to best responses by the principal of the form \((x_-, x_+)\). These responses, as already noted, could involve bounded retention \((x_- < x_+)\), bounded replacement \((x_- > x_+)\) or monotone regimes \((x_+ = \infty)\). In all these cases, conditions (31) and (32) can be used to show that the bad type must lie outside the retention zone, while the good type lies in it. Now consider the second component of the fixed point mapping in which the agents react to these retention and replacement zones. The Appendix formally shows that in all such situations, the bad type exerts more noise in a quest to land inside the retention zone, while the good type attempts to reduce noise so as not to wander out of it. In short, \(\sigma_b > \sigma_g\). But now we’ve established that starting from any \((\sigma_g, \sigma_b) \in [\sigma_s, \sigma^*]^2\), the mapping points into the shaded triangle of Figure 5 in which \(\sigma_b > \sigma_g\). Consequently, every equilibrium must have \(\sigma_b > \sigma_g\), which — as we know already from Proposition 1 — must involve bounded retention.\(^{24}\)

The heart of this argument concerns the location of types relative to replacement and retention zones. Figure 6 illustrates the exceptions when the conditions for Proposition 10 fail, so raising the possibility of bounded replacement equilibria. 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 intersection points of the two weighted densities are either on one side of both the mean types, or straddle them both.\(^{25}\) These are the only two possible

\(^{24}\)In particular, the careful reader will have noticed that under the additional restriction imposed by (32), the Halpern-Bergman theorem no longer needs to be invoked to prove Proposition 9; Brouwer will suffice.

\(^{25}\)This argument shows, in particular, that Panel B of Figure 2 — which we put forward as a possible candidate for a bounded replacement equilibrium — cannot ever be a full equilibrium satisfying both best response conditions.
kinds of bounded replacement equilibria. For completeness, the Appendix provides examples for each of them; for more discussion of the examples, see this footnote.26

6.7. Multiple Types. We extend Proposition 8 to many types, in the costly noise model. We can do so at a level of generality that nests the two-type case, but it is expositionally easiest to assume that there is a prior on types given by some density \( q(\theta) \) on \( \mathbb{R} \). Let \( \mathcal{Q} \) be the space of all such densities and give it any reasonable topology; for concreteness, think of \( \mathcal{Q} \) as a subset of the space of all probability measures on \( \mathbb{R} \) with the topology of weak convergence. A subset \( \mathcal{Q}^0 \) of \( \mathcal{Q} \) is degenerate (relative to \( \mathcal{Q} \)) if its complement \( \mathcal{Q} - \mathcal{Q}^0 \) is (relatively) open and dense in \( \mathcal{Q} \).

Given \( q \in \mathcal{Q} \), each agent of type \( \theta \) chooses noise \( \sigma(\theta) \) as in the model of Section 6.6. Following the choice of noise, a signal is generated. The principal obtains payoff \( u(\theta) \) from type \( \theta \), where \( u \) is some nondecreasing, bounded, continuous function. There is some given continuation payoff — \( V \) — from replacing an agent, which reasonably lies somewhere in between the retention utilities: \( \lim_{\theta \to -\infty} u(\theta) < V < \lim_{\theta \to \infty} u(\theta) \). We also make the generic assumption that \( u(\theta) \) is not locally flat exactly at \( V \). As before, the principal maximizes expected payoff by deciding whether or not to retain the agent after each signal realization, and agents do their best to get retained, with the cost of noise factored in.

**Proposition 11.** Fix all the parameters of the model except for the type distribution. Then, under Condition \( U \), an equilibrium with a monotone retention regime can exist only for a degenerate subset of density functions over types.

See the Online Appendix for a formal proof.

7. Applications

Our theory separates three distinct features: the action (or the choice of risk), the realization of the signal, and the subsequent inference and decision by the principal. A central implication of the model is that the realizations may be "good" — even in the sense of generating high payoffs for the principal today — while they could simultaneously serve as a cautionary indicator for a great deal of risk-taking by the agent, which may well generate a negative inference in agent ability. 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.

In any potential application, the 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

---

26In the first kind of bounded replacement equilibrium, both types are embedded in the retention zone as in Panel A of Figure 6, with \( x_+ < x_- < \theta_b < \theta_g \). 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 \( x_+ < \theta_b < \theta_g < x_- \), 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: \( p \gg q \).
obscure or clarify his ability, (b) whether it is reasonable to suppose that the resulting choice of noise cannot be understood 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 Section 6.3, 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.

In what follows we describe two potential applications in some detail.

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.27 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 2008.28 It is

---

27See 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

28The 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
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.\footnote{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 5.3, 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.}

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 (Ippolito, 1992; Chevalier and Ellison, 1997; Sirri and Tufano, 1998),\footnote{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).} 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.”

question, understand and manage evolving risks within a system essential to the well-being of the American public” (Commission, 2010, emphasis added).
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 non-linear, 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. 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 6.3. 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.

7.2. **Risky Politics.** As a general point, 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 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.

---

31 The use of risky gambles by managers with career concerns is studied in a dynamic setting by Makarov and Plantin (2015).
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 a 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.

7.3. Leaks and Media Reporting. Imagine an individual or a coalition of individuals, including their supporters in the media, who are leaking information or reporting news stories in order to discredit a government administration. Think of particular leaks as signal realizations, none of which can be fully confirmed and need to be assessed by the public, which serves as the principal.

Here, “retention” would mean that the public is won over by the credibility of the leaker, and rewards the leaker by willingly publishing and consuming more stories from the source(s) in question, or by moving against the administration which the leaker seeks to discredit. “Replacement” means that the public turns to other news sources, or increases its support for the government under the belief that the leaker is not credible.

We could think of a type as a description of the attributes of the current administration. A good type is an executive aligned with the preferences of the median voter, whereas a bad type is the opposite. The leakers want the public to believe that the administration is of a bad type with sufficiently high probability, regardless of the administration’s true type (perhaps for partisan reasons). The overall quality of the leaks will depend on the access that the leaker has to inside information. The ambient level of noise, $\sigma$, corresponds to the rate at which the news generating process operates. One can think of the choice of which information to release, or pursue in a media report, as an unobserved choice of risk, every time the public is not fully aware of the existence of this group of leaks. The leakers can choose to release all the inside information, or to pursue a more secretive strategy, which in the language of our model means to choose a higher variance, to the extent that higher variance is associated with less information provision.
The assumption that the type is fixed here means that the leakers cannot create (fake) news about the current government.

In this scenario, our model suggests that, when the government is of a bad type, the leakers will pursue the most possible permissible policy, and release all the information they have access to. When the government is good, however, the leakers will employ a more secretive strategy. In turn, the public will observe the leaks and will choose to “buy” them and update their beliefs downwards if the news is bad, but not too bad. In this case, they follow the maxim, “if it seems too bad to be true, it probably is.”

The model can also be applied to the case of choosing a policy on leaks. Our bad and good leaders are now in office, and they (as always) want the public to believe they are of a good type with sufficiently high probability, so they can stay in office longer.

The executive is full of bureaucrats who have random access to inside information. Assume that, regardless of the type of the government, there always exists a group of public employees who can release these pieces of inside information — exogenously — at the same rate at which they are generated. Think of particular leaks as signal realizations, none of which can be fully confirmed and need to be assessed by the public, which serves as the principal. The current administration can choose to plug the leak, which is here interpreted as choosing a higher variance, to the extent that higher variance is associated with less information provision. The assumption that the type is fixed here means that the administration cannot create good (fake) news about itself.

Our model suggests that the good type will not put any effort into plugging the leaks. After all, every piece of information that is revealed to the voters is most likely to be good and help the government to stay in office. The bad government, in turn, will employ a more secretive strategy, and each voter will update their beliefs downwards when the news are bad, and when the news are too good.

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 non-normal signal structures, non-binary agent types, multiple agents each with privately known types, situations in which the agent can shift the means of their signals at an additional cost, a dynamic version of the model with agent term limits, and environments in which the principal can ex ante commit to retention rules.
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 in some detail — 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 2. The proof of this proposition is long and contains several steps, with many technical details relegated to the Online Appendix.

First, we eliminate bounded replacement equilibria. In such an equilibrium the principal replaces the agent when the signal falls inside \([x_+, x_-]\), with \(x_+ < x_-\), and Proposition 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 it is therefore clear that \(\sigma_b < \sigma_g\) can never hold.

With bounded replacement out of the way, we study feasible monotone and bounded retention regimes, and agent responses to them.

Lemma 1. Under 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), it is easy to check that

\[
\frac{x_+ + x_-}{2} = \frac{\sigma_b^2 \theta_g - \sigma_g^2 \theta_b}{\sigma_b^2 - \sigma_g^2}.
\]
So if \( \sigma_b > \sigma_g \) then \( x_+ > \frac{x_++x_-}{2} > \theta_g \).  

**Lemma 2.** In any bounded retention equilibrium with thresholds \( x_- \) and \( x_+ \),

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

for each type \( k \).

**Proof.** In a bounded retention equilibrium, \( \sigma_b > \sigma_g \) 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 (33). \( \blacksquare \)

**Lemma 3.** (i) If \( X = [x^*, \infty) \) and \( \theta_k > x^* \), the agent chooses \( \sigma_k = \bar{g} \); 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 \( \sigma_k = \bar{g} \).

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

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

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

\[
\sigma^*_k = \sqrt{\frac{x_+ - x_-)(\frac{x_+ + x_-}{2} - \theta_k)}{\ln(x_+ - \theta_k) - \ln(x_- - \theta_k)}} \in (x_- - \theta_k, x_+ - \theta_k),
\]

and agent \( k \) sets \( \sigma_k = \max\{\bar{g}, \sigma^*_k\} \).

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

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

It is always negative if \( x^* < \theta_k \), so \( \sigma_k = \bar{g} \); 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 \( \sigma_k \geq \bar{g} \), to maximize

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

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

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

where \( d_k \) is defined in (34). By Lemma 1, \( x_+ > \theta_g \geq \theta_k \) for any \( k \). If in addition, \( x_- \leq \theta_k \), then the sign of the derivative is always negative, so \( \sigma_k = \bar{\sigma} \).

(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

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

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 (36), given by setting the term within curly brackets equal to zero, which yields:

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

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

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

**Lemma 4.** In any non-trivial equilibrium, \( \sigma_g = \bar{\sigma} \).

**Proof.** If an equilibrium is monotone, then \( \sigma_g = \sigma_o = \bar{\sigma} \) by Lemma 3(i). Otherwise, a non-trivial equilibrium must have bounded retention, in which case \( \sigma_g < \sigma_o \) by Proposition 1. Suppose, on the contrary, that \( \bar{\sigma} < \sigma_g \). Then both choices of noise are interior, and so agent optimality requires

\[
\begin{align*}
\phi \left( \frac{x_+ - \theta_b}{\sigma_b} \right) (x_- - \theta_b) &= \phi \left( \frac{x_+ - \theta_b}{\sigma_b} \right) (x_- - \theta_b), \\
\phi \left( \frac{x_+ - \theta_g}{\sigma_g} \right) (x_- - \theta_g) &= \phi \left( \frac{x_+ - \theta_g}{\sigma_g} \right) (x_- - \theta_g).
\end{align*}
\]
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:
\[ (37) \quad \Psi(\sigma) \equiv \max \left\{ \frac{[x_+ (\sigma) - x_- (\sigma)](x_+ (\sigma) + x_- (\sigma) - \theta_b)}{[\ln(x_+ (\sigma) - \theta_b) - \ln(x_- (\sigma) - \theta_b)] \sigma} \right\}, \]
where for any \( \sigma > \sigma \),
\[ (38) \quad x_- (\sigma) := \frac{\sigma^2 \theta_g - \sigma^2 \theta_b - \sigma \sigma R(\sigma)}{\sigma^2 - \sigma^2} \quad \text{and} \quad x_+ (\sigma) := \frac{\sigma^2 \theta_g - \sigma^2 \theta_b + \sigma \sigma R(\sigma)}{\sigma^2 - \sigma^2}, \]
with
\[ (39) \quad R(\sigma) := +\sqrt{(\theta_g - \theta_b)^2 + (\sigma^2 - \sigma^2) 2 \ln \left( \frac{\beta \sigma}{\sigma} \right)}. \]

To interpret these objects, notice that \( x_- (\sigma) \) and \( x_+ (\sigma) \) are the roots to
\[ (40) \quad \beta \frac{1}{\sigma} \phi \left( \frac{x - \theta_g}{\sigma} \right) = \frac{1}{\sigma} \phi \left( \frac{x - \theta_b}{\sigma} \right), \]
so these are the bounds of the principal’s retention regime \( X \) when she 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 the bad type’s reaction to a chain that starts with a conjecture about that type’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 \( \sigma > \sigma \). The following lemma relates this to condition (11).

**Lemma 5.** \( x_- (\sigma) \) and \( x_+ (\sigma) \) are well-defined and distinct for \( \sigma > \sigma \) if and only if (11) holds.

**Proof.** The proof proceeds by showing that condition (11) is equivalent to the requirement that the term within the square root on the right-hand side of (39) is strictly positive. We relegate the details to the Online Appendix.

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 \((40)\), then the good type optimally chooses \( \sigma_g = \sigma \).

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

Otherwise, \( x_-(\sigma) > \theta_g \). 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 \((34)\). 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 \((34)\) that

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

It follows that

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

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

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

where the second equality follows from \((40)\), the third equality from \((41)\), 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 highest and lowest 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 \( \theta_b \) 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.” 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 \((11)\). 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 (12) fails, then $x_-(\sigma) < \theta_b$ for all $\sigma > \sigma_\dagger$.

Proof. See Online Appendix.

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

**Lemma 8.** If (11) and (12) hold, there is a unique non-trivial equilibrium. It has bounded retention.

Proof. By Lemma 7 (i), $\lim_{\sigma \to \sigma_\dagger} x_- (\sigma) = x^+ (\sigma)$. Inspect the definition of $x^+ (\sigma)$ in (5) and note that if (12) holds, then $x^+ (\sigma) > \theta_b$. Also by Lemma 7 (i), $\lim_{\sigma \to \sigma_\dagger} x_+ (\sigma) = \infty$. Using this information in (37), we see that $\lim_{\sigma \to \sigma_\dagger} \Psi (\sigma) = \infty$.

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 $\sigma > \sigma_\dagger$, 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 (37) must bind. It follows that $\Psi (\sigma)$ solves $d_b (\Psi (\sigma)) = 0$, where $d_b$ is defined in (34), so that

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

Equation (42) can be used to compute $\Psi' (\sigma)$. This is routine but tedious; the Online 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 $\sigma_\dagger > \sigma$, 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 $\sigma^\dagger > \sigma$ 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). So only bounded retention equilibria can exist.

Lemma 9. If (12) fails (in which case (11) automatically holds), there is a unique non-trivial equilibrium. It has monotone retention.

Proof. If (12) fails (and (11) holds), then $x_-(\sigma) < \theta_b$ for all $\sigma > \sigma$ 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 $\sigma \geq \sigma$. So by Lemma 3 (ii), $\Psi(\sigma) = \sigma$ for all $\sigma > \sigma$ and has no fixed point with $\sigma > \sigma$.

We separately verify that there is an equilibrium with monotone retention and both types choosing $\sigma$. If $(\sigma_b, \sigma_g) = (\sigma, \sigma)$, then the planner uses the monotone retention strategy with threshold $x^*(\sigma)$. Because (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 (11) fails, a non-trivial equilibrium does not exist.
Proof. The details of this argument center around establishing the validity of Figure 8. When (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 Online Appendix for the formal details.

We can now complete the proof of Proposition 2. Part (i): If (11) holds, Lemmas 8 and 9 imply the existence of a unique nontrivial equilibrium. If (11) fails, Lemma 10 says a nontrivial equilibrium does not exist. Part (ii): The existence of the bounded retention equilibrium is provided by Lemma 8. The equilibrium strategies are given by Lemmas 1, 4 and 3 (iii). Part (iii) is immediate.

Proof of Proposition 3. First note that at an optimum, the surplus to the principal must be strictly positive. The reason is that the optimal payoff must be at least as high as the equilibrium payoff in Proposition 2, which is strictly positive.\(^{32}\) Now let \( r \) be an optimal contract, and let \( \sigma_b \) and \( \sigma_g \) be the accompanying choices of noise. Given these latter values, let \( x_1 \) and \( x_2 \) solve the quadratic equation (4). These roots must be distinct (one of them could be infinite), otherwise the principal’s optimal payoff cannot be strictly positive.\(^{33}\) The principal’s best-response retention strategy, \( \hat{r} \), is either \( \hat{r}(x) = 1 \) on \([x_1, x_2]\) and zero elsewhere, or \( \hat{r}(x) = 0 \) on \([x_1, x_2]\) and 1 elsewhere.

We claim that \( r = \hat{r} \) a.e. Suppose not. For \( \epsilon > 0 \) and small, define a new retention function by \( r(x, \epsilon) \equiv \epsilon \hat{r}(x) + (1 - \epsilon) r(x) \) for every \( x \). Because \( \hat{r}(x) = 1 \) if and only if the inequality in (3) holds, and because equality in (3) holds only at the roots \( x_1 \) and \( x_2 \), this new retention function must yield strictly higher payoff to the principal as \( \epsilon \) increases, assuming that \( \sigma_b \) and \( \sigma_g \) are held constant. Moreover, this increase is first order in \( \epsilon \), evaluated at \( \epsilon = 0 \). At the same time, by the envelope theorem, the payoff change to either type of agent is second-order in \( \epsilon \) at \( \epsilon = 0 \), because \( \sigma_b \) and \( \sigma_g \) are optimal choices at \( r \). That means the net gain to the principal is strictly positive, when evaluated at \( \epsilon = 0 \), even taking the reactions of the agent types into account. We therefore have a contradiction to optimality.

Proof of Proposition 8. Recall (30); this is the equation that \( \sigma \) must satisfy if it commonly chosen by both types:

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

where \( z_1 = (\sigma/\Delta) \ln (\beta) - (\Delta/2\sigma) \) and \( z_2 = (\sigma/\Delta) \ln (\beta) + (\Delta/2\sigma) \). The function \( \phi (z) \) has the shape shown in Figure 4, reaching maxima and minima at \( z = 1 \) and \( z = -1 \) respectively, and

---

\(^{32}\)The principal can always choose the equilibrium retention strategy of Proposition 2 as a feasible option here, along with the corresponding agent choices in that equilibrium. Because the principal strictly prefers to retain within \( X \), her equilibrium surplus — and therefore her optimal payoff in the current problem — must be strictly positive.

\(^{33}\)After all, the optimal payoff to the principal is no lower than the payoff under her optimal response to \((\sigma_b, \sigma_g)\).

If the roots were not real and distinct, then the principal would always either retain or replace, neither of which can generate a strictly positive payoff.
exhibiting “negative symmetry” around 0. Using (30), this tells us that there are two exclusive possibilities: (i) either \( \beta > 1 \) and \( \sigma < \bar{\sigma} \), or (ii) either \( \beta < 1 \) and \( \sigma > \bar{\sigma} \). 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 (43) holds, and it is independent of the cost function \( c \).

**Proof of Proposition 9.** Recall that \( \sigma_* < \bar{\sigma} \) and \( \sigma^+ > \bar{\sigma} \) are the two solutions to \( c(\sigma) = 1 \). Let \( \Sigma := [\sigma_*, \sigma^+]^2 \), and define

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

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

\[
\Delta^2 + (\sigma_b^2 - \sigma_g^2) 2 \ln \left( \frac{\beta \sigma_b}{\sigma_g} \right)
\]

is strictly positive. But (31) tells us that \( \ln(\beta) \geq -\frac{\Delta^2}{2\sigma^*} \), and so

\[
\Delta^2 + (\sigma_b^2 - \sigma_g^2) 2 \ln \left( \frac{\beta \sigma_b}{\sigma_g} \right) = \Delta^2 + (\sigma_b^2 - \sigma_g^2) 2 \ln \left( \frac{\beta \sigma_b}{\sigma_g} \right) \\
\geq \Delta^2 + (\sigma_b^2 - \sigma_g^2) 2 \ln (\beta) \\
\geq \Delta^2 \left[ 1 - \frac{\sigma_b^2 - \sigma_g^2}{\sigma^*} \right] > 0,
\]

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 Proposition 1, the zone \([x_-, x^+]\) must involve retention.

Next, for each pair \((x_-, x^+)\) with \( x^+ > x_- \) and with \( x^+ \) possibly infinite, define \((\sigma'_b, \sigma'_g)\) 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 \( (\sigma_g^n, \sigma_b^n) \to (\sigma_g, \sigma_b) \) with \( \sigma_b > \sigma_g \), then \( \Psi_1(\sigma_g, \sigma_b) = (x_-, x^+) \) with \( x_- < x^+ < \infty \), and it is obvious that \( \Psi_1(\sigma_g^n, \sigma_b^n) \to \Psi_1(\sigma_g, \sigma_b) \).
of the quadratic condition (4) (the roots of which yield \( x_- \) and \( x_+ \)) reveals that \( \Psi_1(\sigma^b, \sigma^b) = (x_-^n, x_+^n) \) 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 (27), which we reproduce here for convenience:

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

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 \( \sigma_k^n \) 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 (44) that \( \sigma_k^n \to \sigma_k \). In the latter case, the fact that \( \sigma_k^n \to \sigma_k \) 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 \((\sigma_s, \sigma_s) \leq \Psi(\sigma_g, \sigma_b) \leq (\sigma^*, \sigma^*)\). Therefore, if \( (\sigma_g, \sigma_b) \in \Sigma^+ \) with \( \sigma_b > \sigma_g \), (45) 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 \([\sigma_s, \sigma^*]\), and the second component is larger than the first. The remaining case is one in which \( (\sigma_g, \sigma_g) \in \Sigma^+ \) with \( \sigma_b = \sigma_g \). In this case, we know from condition (31) that \( \Psi_1(\sigma_g, \sigma_b) \) is of the form \((x_-, x_+) \equiv (x^*, \infty)\), where \( x^* \in [\theta_k, \theta_g] \). From the first-order conditions that describe each type — see (29) — 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 \( \sigma_s \) and \( \sigma^* \)). It follows that for every \( a \in (0, 1) \), (45) 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 10, we first state the following two results.

**Lemma 11.** In any bounded replacement equilibrium with thresholds \( x_- \) and \( x_+ \),

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

for each type \( k \).
Proof. When $\sigma_b \neq \sigma_g$, and $x_-$ and $x_+$ are both finite and given by (4), we have that

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

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, 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 (46). 

Lemma 12. Under (31) and (32), $x_+ < \theta_b < x_- < \theta_g$ in a bounded replacement equilibrium.

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

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

(47)

where $A = \ln (\beta \sigma_b / \sigma_g)$, and replacement is strictly optimal if the opposite strict inequality holds. Putting $x = \theta_b$ in (47) and simplifying, we see that replacement is strictly optimal at $\theta_b$ if

$$\beta < \frac{\sigma^2_g}{\sigma^2_b} \exp \frac{\Delta^2}{2 \sigma^2_g},$$

but this is guaranteed by the right hand inequality of (31), 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 (47) and simplifying, we see that retention is strictly optimal at $\theta_g$ if

$$\frac{\Delta^2}{2 \sigma^2_g} + \ln (\sigma_b) - \ln (\sigma_g) > - \ln (\beta).$$

(48)

The derivative of the left hand side of (48) with respect to $\sigma_b$ is given by

$$\frac{1}{\sigma_b} \left(1 - \frac{\Delta^2}{\sigma^2_b}\right),$$

which is strictly negative given (32) and $\sigma_b \leq \sigma^*$, so it follows that the left hand side of (48) is minimized by setting $\sigma_b = \sigma_g = \sigma^*$. To establish (48), then, it is sufficient to have

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

but this is guaranteed by the left hand inequality of (31). 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.

$\blacksquare$
Proof of Proposition 10. Suppose that a bounded replacement equilibrium exists. Then we have \( \sigma_g > \sigma_b \) and \( x_- > x_+ \). By Lemma 12, we have \( \theta_g \geq x_- \geq \theta_b > x_+ \).

Define \( B_k(\sigma) \) to be type-\( k \)'s marginal benefit of noise:

\[
B_k(\sigma) := \phi\left(\frac{x_+ - \theta_k}{\sigma}\right) \frac{x_+ - \theta_k}{\sigma^2} - \phi\left(\frac{x_- - \theta_k}{\sigma}\right) \frac{x_- - \theta_k}{\sigma^2}.
\]

That this is indeed the marginal benefit can be seen easily by recalling (27), which sets this expression equal to marginal cost. Observe that for every \( \sigma \),

\[
\begin{align*}
B_b(\sigma) & = \phi\left(\frac{x_- - \theta_b}{\sigma}\right) \frac{x_- - \theta_b}{\sigma^2} - \phi\left(\frac{x_+ - \theta_b}{\sigma}\right) \frac{x_+ - \theta_b}{\sigma^2} \\
& \geq \phi\left(\frac{x_- - \theta_b}{\sigma}\right) \frac{x_- - \theta_b}{\sigma^2} - \phi\left(\frac{x_+ - \theta_b}{\sigma}\right) \frac{x_+ - \theta_b}{\sigma^2} \\
& = \phi\left(\frac{x_+ - \theta_b}{\sigma}\right) \frac{x_+ - x_+}{\sigma^2} \\
& > \phi\left(\frac{x_+ - \theta_g}{\sigma}\right) \frac{x_+ - x_+}{\sigma^2} \\
& = \phi\left(\frac{x_+ - \theta_g}{\sigma}\right) \frac{x_+ - \theta_g}{\sigma^2} - \phi\left(\frac{x_- - \theta_g}{\sigma}\right) \frac{x_- - \theta_g}{\sigma^2} \\
& \geq \phi\left(\frac{x_- - \theta_g}{\sigma}\right) \frac{x_- - \theta_g}{\sigma^2} - \phi\left(\frac{x_+ - \theta_g}{\sigma}\right) \frac{x_+ - \theta_g}{\sigma^2} \\
& = B_g(\sigma),
\end{align*}
\]

where the first weak inequality follows from \( x_- \geq \theta_b \) and inequality (46) of Lemma 11, the strict inequality follows from \( \phi \) single-peaked around zero and \( x_+ - \theta_g < x_+ - \theta_b < 0 \), and the last weak inequality follows from \( x_- \leq \theta_g \) and (again) inequality (46) of Lemma 11.

But (50) 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 Proposition 1, this contradicts the fact that we are in a bounded replacement equilibrium.

Proof of Propositions 4, 5, 6, 7, and 11. See the Online Appendix.

References


Degan, A. and M. Li (2016) “Persuasion with Costly Precision,” mimeo, Department of Economics, Concordia University.


