Too Good To Be True? Retention Rules for Noisy Agents

Francisco Espinosa ⓡ Debraj Ray[†]

This version: January 2022

Forthcoming, *American Economic Journal: Microeconomics*

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

1. Introduction

If it's too good to be true, it probably is. This is a great motto for the gullible when it comes to financial decisions, because we all know about fly-by-night financial schemes — Charles Ponzi or Bernard Madoff come readily to mind. But the concerns go well beyond deliberate fraud. For instance, consider a fund manager who is eager to carve out a reputation and must therefore take on significant risk. Is his current performance any indicator of future success, or does it signal a strategy of excessive risk-taking, in the hope of masking a lack of competence? After all, even among the high risk-takers — or more accurately, *especially* among them — there must be dramatic winners and losers. A *New York Times* article (March 7, 2014), which we will return to below, had the suggestive title: "When You Evaluate a Fund Manager, Look Beyond Results." Risk-taking is not an illegal activity of course, but falls into a class of situations in which a principal must still be wary of excessively good news coming from the agent.

In a different arena, consider a politician who is eager to make his mark on the national or world stage. He sets up a global summit with a rogue leader — a risky act that might provide for some real progress, such as a temporary cessation of military exercises.

[†]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, participants at the University of Chicago Harris Political Economy Lunch, a Co-Editor and three anonymous referees for useful comments.

This is good news ex post, but was it a wise move ex ante? After all, a politician of unknown competence could try something of high risk, in the hope of a spectacular success — but simultaneously braving the chances of abject failure. With the observed outcome in hand, the median voter would need to infer the extent of risk-taking, and use these as guidelines, say in her voting decision. Specifically, under the knowledge that politicians of varying ability might differ in their risk strategies, the median voter could attach likelihoods of different degrees of risk-taking that led to the observed outcome, thereby drawing inferences about the politician's unknown ability.

Or consider a fledgling organization of unknown ability, say an NGO, seeking funding from donors. The NGO could take on a safe project — say, as a provider of social services in a familiar location, with outcomes that accurately signal its competence. Or it could entertain a risky intervention in a distant setting, but with some chance of attention-grabbing success. A potential donor cannot fully see the ex ante risks in taking on such a project, and only sees the final outcome. Even if the outcome is a good one, the donor's concerns lie with whether the NGO should be funded for *future* activities, and for this purpose the outcome, good though it may happen to be, should also be used to appraise whether an ex ante risk was taken by the NGO, the extent of that risk, and what it might signal about the NGO's type.

One could continue in this vein: a government under pressure might inject noise into official statistics, a researcher might take risky steps to bolster his cv for an upcoming promotion, a lawyer might call a high-risk witness (who could destroy the case or win it), an athlete might engage in doping with spectacular outcomes, and so on. In all these situations, an excellent outcome in the here-and-now is also cause for some caution in evaluating future performance. The model we propose, while admittedly an entirely theoretical exercise, might bear on such cases.

Each of these situations has many specific features at play. But we consider the following common thread. There is an agent (fund manager, politician, NGO, etc) who privately knows his type (good or bad), and who seeks to be retained by a principal (investor, the median voter, funding agency, etc.). The principal wishes to retain a good type, and to remove a bad type. The agent generates a noisy but informative signal of his type. He can choose to amplify or reduce the precision of this process, but there are two restrictions. First, the signal structure is constrained by the type; specifically, the mean of the signal is given by the type. Second, signal realizations cannot be tampered

with ex post.[1] 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 her retention or replacement decision.

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

In our baseline setting, the agent emits a normal signal centered around his type. This centering cannot be changed, but the variance can be freely altered, subject to some ambient lower bound that's "small enough" in a sense to be made precise. The principal sees the outcome, and retains if and only if her posterior on the good type exceeds some threshold. Observation 1 argues that there are three types of potential equilibria. The first is *monotone retention*, in which both types choose the same noise, and the principal retains if the signal is above some threshold. The second is *bounded retention*, in which 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 the bad agent chooses lower noise than the good agent, and the principal replaces the agent for intermediate signal realizations.

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

Our framework is stark and minimal, but easily extendable in several directions. In Section 6, we study some extensions that show the robustness of this observation and accommodate various ancillary features. These include costly noise, a dynamic setting with agent term limits, and non-normal signal structures. In the Supplementary Appendix, we study other extensions: costly mean-shifting, non-binary types, and the possibility of *principals* injecting noise into their assessments. In Vohra ⓡ Espinosa ⓡ Ray (2021), we study an extension to commitment, in which the principal pre-announces mechanisms to assess agents.

In Section 7, we return to two of the applications mentioned above, to conclude the paper.

[1] In Section 6.5, we allow the agent to hide a signal after observing its realization, at some cost.

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.[2] The cheap talk literature beginning with Crawford and Sobel (1982) can be thought of as a leading example of noisy communication. In that literature, nothing binds the sender. In contrast, as explained above, our chosen communication structures have mean equal to the true state, the choice could be costly, and it is crucial that each individual chooses a *distribution* over signals, and cannot alter the realized outcome ex post (though see Section 6.5 on non-disclosure).

The choice of an information structure is central to Kamenica and Gentzkow (2011). But no agent knows the true type ex-ante, and the chosen structure is observed by the receiver. This last feature is shared by Degan and Li (2016), but the type of the agent is privately known, as in our model. In contrast, in our setting, the choice of information structure is not (fully) observed, only the signal, a feature that we share with DeMarzo et al. (2019). We return to the question of observability (and these references), first in Section 6.1, and then again in Section 7.

Dewan and Myatt (2008) study a model of leadership in which an individual's clarity in communication is a virtue, but the leader also wishes to hold on to an audience for longer, to dissuade them from listening to others. Therefore extreme clarity is not chosen. Edmond (2013) studies the obfuscation of states (by a dictatorial regime), but restricts attention in his analysis (by assumption) to receiver-actions that are monotone in the signal realization. In contrast, in our setting, the *non*-monotonicity of receiver actions is a fundamental and robust outcome of the model.

Harbaugh, Maxwell and Shue (2016) study the inclinations of a sender to distort the news about multiple projects, depending on the overall realization of news, which cannot be hidden from the receiver. Such distortions are separate from mean-preserving noisy announcements; the focus is on the realized spread of multidimensional news over multiple projects. While our results are entirely distinct, they too take note of a different "too-good-to-be-true" inference problem. Specifically, the sender will distort

[2]In this brief review we omit discussion of a related but distinct literature with *exogenous* noise, as in the limit pricing game studied by Matthews and Mirman (1983), the choice of mean return by managers of unknown quality who might seek to herd (Zwiebel, 1995), or inference settings when values have exogenous but unknown precision (Subramanyam, 1996).

upwards the spread of the signals when the overall average is bad, and reduce it when the overall average of 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 agent 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 the extent of closeness to some pre-stipulated, finite level of output, a high-effort low-noise equilibrium is achieved.

Palomino and Prat (2003) and Barron et al. (2017) also study situations in which agents can inject noise into a moral hazard setting. In 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).

Makarov and Plantin (2015) study the question of manager behavior and the design of performance contracts (or renewal decisions). Theirs is a dynamic setting with uncertainty in which managers with career concerns use risky gambles to distort their perceived skill temporarily. The design of compensation contracts aims at curtailing excessive risk-taking. Unlike our setting, information is symmetric between investors and managers, and all actions are publicly observable. Money managers are therefore not punished for extremely good performance.

Finally, there is also a literature on policy uncertainty (Shepsle, 1972; Campbell, 1983; Alesina and Cukierman, 1990; Glazer, 1990; Aragones and Neeman, 2000; Aragones and Postlewaite, 2002; Aragones, Palfrey and Postlewaite, 2007), where candidates offer deliberately ambiguous policy platforms that generate uncertainty regarding the policies to be implemented in the event of victory. Our adverse-selection setting is entirely distinct but shares the same feature of endogenous noise.

3. A BASELINE MODEL

3.1. **Setting.** An agent (male) works for a principal (female). The agent can be good (g) or bad (b), with $g > b$. He knows his type. The principal doesn't, but has a prior $q \in (0,1)$ that the agent is good. After 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 has some continuation value $V \in (U_b, U_g)$. The agent gets a payoff equal to 1 if he is retained and 0 otherwise. He therefore prefers to be retained regardless of type, while the principal prefers to retain only the good agent.

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

$$x = \theta_k + \sigma_k \epsilon,$$

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

In our baseline setting the choice of noise is costless but bounded below: $\sigma_k \geq \underline{\sigma}$ for some $\underline{\sigma} > 0$. (We will add costly noise in Section 6.2.) 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 $\underline{\sigma}$ as "small" (see below). Define $p \in (0,1)$ by

$$pU_g + (1-p)\,U_b \equiv V; \tag{1}$$

then p is interpretable as an "outside option probability" that leaves the principal indifferent between retaining and replacing. A salient benchmark is $p = q$ (the *balanced model*). But if V incorporates the option value of retaining an agent in a dynamic context, p could exceed q (see Section 6.3). We call this an *optimistic future*. On the other hand, if there is already positive information about the current agent, then p could be smaller than q; we call this a *pessimistic future*.

[3] For an extension to costly mean-shifting, see the Supplementary Appendix.

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

$$\Pr(k=g|x) = \frac{q\frac{1}{\sigma_g}\phi\left(\frac{x-\theta_g}{\sigma_g}\right)}{q\frac{1}{\sigma_g}\phi\left(\frac{x-\theta_g}{\sigma_g}\right) + (1-q)\frac{1}{\sigma_b}\phi\left(\frac{x-\theta_b}{\sigma_b}\right)} \geq p, \tag{2}$$

where ϕ is the standard normal density. So we have retention if and only if

$$\frac{\frac{1}{\sigma_b}\phi\left(\frac{x-\theta_b}{\sigma_b}\right)}{\frac{1}{\sigma_g}\phi\left(\frac{x-\theta_g}{\sigma_g}\right)} \leq \frac{1-p}{p}\frac{q}{1-q} := \beta. \tag{3}$$

Simple algebra involving the normal density yields the equivalent expression

$$\left(\sigma_g^2 - \sigma_b^2\right)x^2 + 2\left(\sigma_b^2\theta_g - \sigma_g^2\theta_b\right)x + \left(\sigma_g^2\theta_b^2 - \sigma_b^2\theta_g^2 + 2A\sigma_g^2\sigma_b^2\right) \geq 0, \tag{4}$$

where $A := \ln(\beta\sigma_b/\sigma_g)$. Inequality (4) defines a *retention regime*, a zone X of signals for which the principal retains the agent. An *equilibrium* is a configuration (σ_g, σ_b, X) such that given (σ_g, σ_b), X is the set of "retention signals" x which solve (4), and given X, each type k chooses σ_k to maximize the probability of retention; that is

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

3.3. **A Remark on Interpretation.** Principal-agent models are typically concerned with situations in which an agent takes an action that affects the payoff of the principal. That action could *also* influence the principal's retention decision in a dynamic setting; see, for instance, Dutta et al. (1989) for the case of moral hazard, and Banks and Sundaram (1998) for the case of adverse selection.

In our model, the incentives to elicit current effort have been deliberately muted, so as to concentrate on the retention decision alone, and dynamic effects have also been suppressed through the device of an outside option. That said, it may be useful to keep the following structure in mind. Agent-generated signals today are *both* signals and payoff-relevant outcomes, such as current output. The principal's payoff depends on these outcomes or signals, and she is risk-neutral. So the chosen *distribution* of signals — or outcomes — is of no direct consequence to her; only the mean matters for her retention decision. Therefore her retention decision is entirely based on her update following the signal. Our model fits this setting precisely.

This interpretation could be disturbed by the possibility that the agent can shift the mean outcome (relative to their "natural" type) by expending effort. But this additional feature is easy to incorporate, as we show in the Supplementary Appendix. Or there could be dynamic considerations that overturn or unduly complicate the baseline reasoning. We discuss such an extension in Section 6.3. It supports our static arguments, and even simplifies the statement of the results.

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

4. Retention Regimes

4.1. **Trivial Retention Regimes.** Two examples of retention zones are (a) "always retain," so that $X = \mathbb{R}$, and (b) "always replace": $X = \emptyset$. Both generate complete indifference for either agent. With any cost function for noise that is minimized at some common value for both types, we then have $\sigma_g = \sigma_b$, but then (4) must alter sign over x, a contradiction (Section 6.2 shows this explicitly). Even without any cost of noise, these equilibria are eliminated in a dynamic setting (Section 6.3). So in the benchmark model, we ignore such trivial and delicately supported regimes.

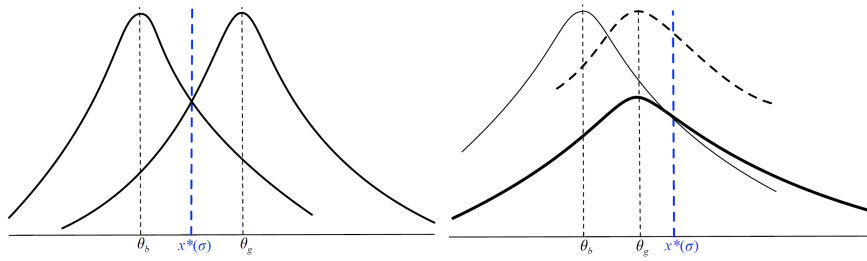
4.2. **Monotone Retention Regimes.** An equilibrium regime is *monotone* if there is a finite threshold x^* such that the principal replaces the agent for signals on one side of x^*, and retains him for signals to the other side of x^*. See Figure 1. A monotone retention regime arises (and can *only*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

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

So $x^*(\sigma)$ is the threshold above which a signal from two possible noisy sources of *equal* variance is more likely to be coming from the higher-mean source. This is the exact interpretation of $x^*(\sigma)$ in the balanced model, for then $\beta = 1$ and

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

[4] If $\sigma_g \neq \sigma_b$, then by condition (4), the resulting retention regime is either trivial or non-monotone.



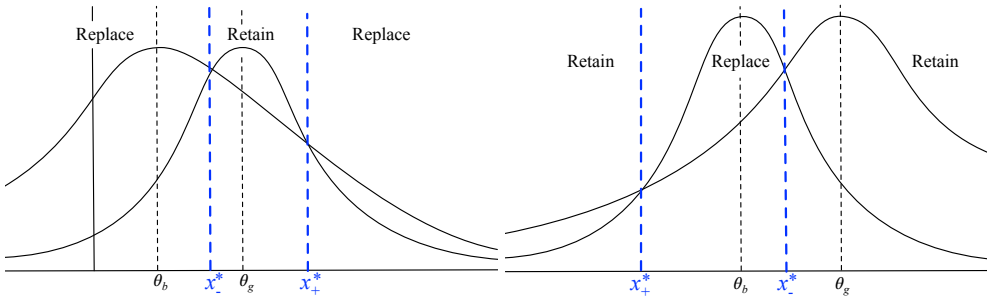
(a) Threshold between types **(b)** Threshold to one side of types

Figure 1. The Symmetric Threshold $x^*(\sigma)$

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

Consider any monotone equilibrium with threshold between θ_b and θ_g (e.g., as in the balanced case). Then the good type will want to minimize noise, while the bad type will want to maximize it. But this 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, relative likelihoods move in favor of the bad type by virtue of a larger choice of variance, and *despite* its lower mean. The analysis in the rest of this section, and in Section 6.2, extends these arguments to *all* single-threshold equilibria, arguing that if the minimal noise level $\underline{\sigma}$ is small enough or if there is a cost of noise, no monotone equilibrium can exist, whether the retention threshold lies between the means of the two types, or to one side of these.

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



(a) Bounded Retention **(b)** Bounded Replacement

Figure 2. Differential Noise and the Retention Decision

replacement occurs. Obviously, x_+ and x_- are the two roots of (4), which means that

$$\beta \frac{1}{\sigma_g} \phi\left(\frac{x-\theta_g}{\sigma_g}\right) = \frac{1}{\sigma_b} \phi\left(\frac{x-\theta_b}{\sigma_b}\right) \tag{6}$$

implies the equalization of weighted likelihoods for both types at $x = x_-, x_+$. Furthermore, the weighted likelihood for the good type must have a higher slope in x relative to that for the bad type, evaluated at x_-, so retention occurs for $x > x_-$. 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 $\phi(z) = (1/\sqrt{2\pi}) \exp\{-z^2/2\}$, $\phi'(z) = -z\phi(z)$, so this is equivalent to:

$$\beta\phi\left(\frac{x_- - \theta_g}{\sigma_g}\right) \frac{x_- - \theta_g}{\sigma_g^3} - \phi\left(\frac{x_- - \theta_b}{\sigma_b}\right) \frac{x_- - \theta_b}{\sigma_b^3} < 0. \tag{7}$$

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

$$\beta\phi\left(\frac{x_+ - \theta_g}{\sigma_g}\right) \frac{x_+ - \theta_g}{\sigma_g^3} - \phi\left(\frac{x_+ - \theta_b}{\sigma_b}\right) \frac{x_+ - \theta_b}{\sigma_b^3} > 0. \tag{8}$$

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

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

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

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

Combining these two inequalities, we must conclude that

$$\left(\sigma_b^2 - \sigma_g^2\right)(x_+ - x_-) > 0 \tag{9}$$

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

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

5. Bounded Retention Equilibrium

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

5.1. No Bounded Replacement.

First, we eliminate bounded replacement equilibria. In such an equilibrium the principal replaces the agent when the signal falls inside $[x_+, x_-]$, with $x_+ < x_-$. Observation 1 tells us that this regime is associated with $\sigma_b < \sigma_g$. In this case, the retention probability for any type goes to 1 as $\sigma_k \to \infty$, and therefore in equilibrium, it can't be that $\sigma_b < \sigma_g$. We remark that this argument is straightforward only because any level of noise above the minimum can be *freely* chosen. We will need to revisit it when the choice of noise is costly.

5.2. Small Ambient Noise.

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

$$\beta \equiv \frac{1}{\alpha(\beta) + \sqrt{1+\alpha(\beta)^2}} \exp\left[-\frac{\alpha(\beta)}{\alpha(\beta) + \sqrt{1+\alpha(\beta)^2}}\right]. \tag{10}$$

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 $\underline{\sigma}$ is small enough so that:

$$\frac{\underline{\sigma}}{\theta_g - \theta_b} < \frac{1}{2}\alpha(\beta)^{-1}. \tag{11}$$

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

$$\frac{\underline{\sigma}}{\theta_g - \theta_b} < \left[\sqrt{2\ln(\beta)}\right]^{-1}. \tag{12}$$

These bounds weaken as we move to the balanced case. *At* the balanced case, no restrictions are imposed at all; both the bounds in (11) and (12) diverge to infinity.

Moreover, in Section 6.3, which studies a dynamic extension with fixed term limits for agents, we show that these restrictions on $\underline{\sigma}$ will be automatically satisfied.

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

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

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

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

The proposition distinguishes between optimistic and pessimistic futures. In the first case, there is a low prior on the ability of our current agent, possibly due to unsatisfactory past performance (not modeled here). So a good type is desperate to reveal himself by reducing noise as far as possible, while the bad type chooses larger noise in the hope that his signals will imitate his good counterpart to the extent possible. So $\sigma_b > \sigma_g$, which serves to uniquely precipitate the bounded retention equilibrium, whenever an equilibrium exists.

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

How permissive is the bound in (12)? One way to view it is to study the perfectly balanced case in which there is neither pessimism nor optimism. Then the right-hand side of (12) is *infinitely* large, and we can unequivocally assert that the unique nontrivial equilibrium involves bounded retention. Now move away from the balanced case by

placing greater faith in the current agent, so that (12) begins to bite. Suppose that the principal's prior on the agent equals $3/4$, while $p = 1/2$. Then (12) implies that the unique equilibrium involves bounded retention as long as the standard deviation of the signal can be brought below approximately 2/3 the difference between the two means.

Two extensions continue to underscore the salience of bounded retention. In Section 6.2, we replace the costless choice of noise by a cost function. That effectively compactifies the space of noises, but the cost function is smooth and not "L-shaped" as in this, our baseline model. Proposition 2 proves that a monotone equilibrium generically cannot exist. This is a more uncompromising prediction than the one of our baseline model, which does admit monotone retention under some circumstances.

Second, in Section 6.3, we describe a dynamic model in which q evolves over time in line with Bayes' Rule. Specifically, we consider an infinite-horizon setting in which each agent faces a two-term limit. It turns out that conditions (11) and (12) *automatically* hold in that exercise, and there is no monotone equilibrium.

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

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

[5]Of course, the principal also distrusts signals that are bad: after all, lower mean and higher variance are particularly synergistic in producing lower signals.

6. Extensions

Section 6.1 remarks on the unobservability of chosen noise. Section 6.2 introduces costly noise. Section 6.3 analyzes a dynamic version with agent term limits. Section 6.4 drops the normality assumption. Section 6.6 briefly describes other extensions, covered in depth in the Supplementary Appendix.

6.1. **Unobservable Noise.** Our result presumes that the choice of agent noise is not observable. Consider the opposite presumption that the noise is observable. Then every type *must* choose the same noise; i.e., separation is impossible via the choice itself — the bad type will deviate by mimicking this choice. So all 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[6]). If risk choices are *costly*, as they will be in Section 6.2, the same argument applies as long as the cost function for noise is the same for both types — again, there must be pooling in observed components.[7]

But any augmentation of this scenario with unobserved noise leads back to our model.[8] The observed component of noise would be chosen to be the same for all types. (If noise is costly, it would be set to the minimum-cost level in any equilibrium refined by intuitive off-path restrictions on beliefs.) The remainder of the analysis dealing with the unobservable component then proceeds with no change. To summarize: (a) *complete* lack of observability is not needed for our results; (b) there will be pooling on the observable components if the choice of risk is costless or uncorrelated with agent

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

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

[8]In an otherwise different setting, DeMarzo et al. (2019) also work with the choice of an unobserved information structure — correctly guessed at in equilibrium. A sender gathers information about the quality of an object by selecting a test, which might return a null result. He can choose to disclose (verifiable) information, but also to suppress test results. The receiver understands this, so there is no value in *observably* choosing a test. But there is some value to choosing an unobserved test — the authors describe this optimal choice.

type; and (c) our results then apply to the unobserved components of risk. The singular nature of Proposition 1 is rooted in the presumption that there is *some* unobserved component of the signal structure, not that the *entire* structure is unobservable.

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

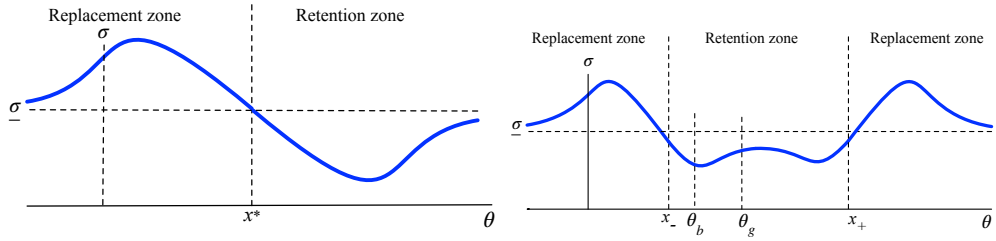
$$\sigma_k \in \arg\max_{\sigma \geq \underline{\sigma}} \left[\int_X \frac{1}{\sigma} \phi\left(\frac{x - \theta_k}{\sigma} \right) dx - c(\sigma) \right].$$

This version of the model presents some new features. First, trivial equilibria never exist. If the retention regime is trivial, both types must choose the lowest cost signal, which in turn makes the signal informative, a contradiction. Second, as we shall see, monotone regimes are generically impossible in equilibrium. But third, the model does opens up the possibility of the existence of bounded *replacement* equilibria, which were easily ruled out in the benchmark model. Now we expand on some of these points, beginning with the agent's best response mapping before moving to a fuller description of equilibrium.

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

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

for each type $k = g, b$. Optimally chosen noise now moves in a subtle and quite complicated way with the location of a player's type. Figure 3, Panel A, illustrates this for a monotone retention threshold. When a player's type is far from the retention



(a) Monotone Retention **(b)** Bounded Retention

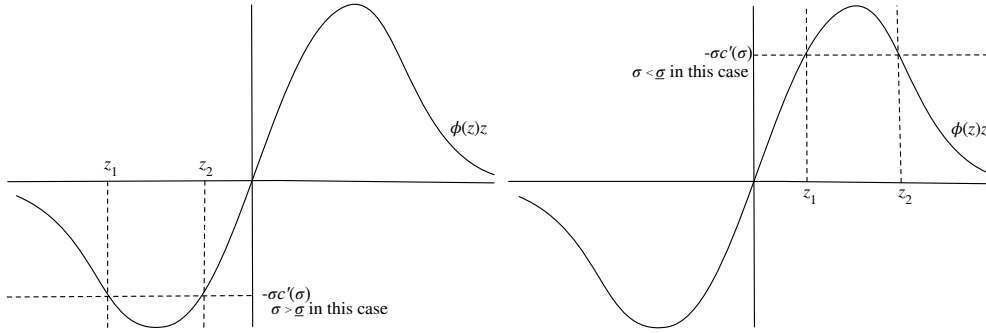
Figure 3. How Choice of Noise Varies With Agent Type

threshold, it takes large noise to generate (with any significant probability) a signal within the retention zone. That's costly, so noise converges to the zero-cost choice $\underline{\sigma}$ as the type moves far from the retention zone. Moving closer to the zone, chosen noise increases, but must decline again: after all, when the type is *on* the edge of the zone, noise makes no difference to the chances of retention, so chosen noise is back to $\underline{\sigma}$ again. As the type moves into the retention zone, noise can only throw him out it, so chosen noise now falls below $\underline{\sigma}$. But the downward movement does not continue forever. Deep in the retention zone, the type is confident of remaining there, and so noise goes up again, converging to $\underline{\sigma}$, this time from below.

With bounded retention zones, the choice function exhibits more non-monotonicities. Panel B of Figure 3 shows that there will generally be five turning points. There is one each for either side of the retention zone, for the same reason as in the earlier discussion. There are three more within the retention zone: noise initially falls as an agent with type close to the edge avoids escape from the zone; then rises in the middle of the zone as the risk of escape falls, then falls again as the risk goes up, and finally rises as we approach the edge.[9] (The noise choice at the edges is below $\underline{\sigma}$, because the retention zone is bounded.)

This behavior is consistent with empirical findings on risk-taking. Genakos and Pagliero (2012) find that risk-taking in weightlifting contests exhibits an inverted-U relationship between risk and rank, with the peak reached around rank 6. Figueiredo et al. (2015) observe that risk-taking by portfolio managers is non-monotonic: managers significantly below a compensation threshold *reduce* risk-taking relative to those who are relatively close. These findings are consistent with our predictions when agents are to

[9]See the Supplementary Appendix, where this discussion is conducted in more detail.



(a) Optimistic Future ($\beta < 1$) **(b)** Pessimistic Future ($\beta > 1$)

Figure 4. Conditions for Monotone Retention

the left of the retention threshold (Panel A in Figure 3). With costly noise, both monotone and bounded retention can generate this type of behavior. That said, monotone retention cannot be an equilibrium outcome of this model, except in degenerate cases.

6.2.2. *No Monotone Retention.*

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

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

$$\phi\left(\frac{x^* - \theta_k}{\sigma_k}\right)\frac{x^* - \theta_k}{\sigma_k^2} - c'(\sigma_k) = 0, \tag{14}$$

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

$$\phi\left(\frac{\sigma}{\Delta}\ln(\beta) + \frac{\Delta}{2\sigma}\right)\left(\frac{\sigma}{\Delta}\ln(\beta) + \frac{\Delta}{2\sigma}\right) = \phi\left(\frac{\sigma}{\Delta}\ln(\beta) - \frac{\Delta}{2\sigma}\right)\left(\frac{\sigma}{\Delta}\ln(\beta) - \frac{\Delta}{2\sigma}\right) = -\sigma c'(\sigma). \tag{15}$$

Equation (15) tells us to study $\phi(z)z$; Figure 4 does so. Denote $\frac{\sigma}{\Delta}\ln(\beta) - \frac{\Delta}{2\sigma}$ by z_1 and $\frac{\sigma}{\Delta}\ln(\beta) + \frac{\Delta}{2\sigma}$ by z_2. Given the shape of $\phi(z)z$, Figure 4 indicates how z_1 and z_2 must be located: they must both have the same sign and the same "height." With an optimistic future ($\ln\beta < 0$), both z_1 and z_2 are negative; see Panel A. With a pessimistic future,

both z_1 and z_2 are positive; see Panel B. In each case, only one value of σ can solve this requirement; i.e., just one value that fits the *first* equality in (15). It is independent of the cost function for noise, and so the second equality generically can not hold.

This contrasts with the case of costless noise, where corner responses are possible at lower bounds of noise, thereby permitting monotone thresholds.

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

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

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

$$c''(\sigma) > \frac{\kappa}{\sigma^2} \text{ for all } \sigma \in [\sigma_*, \sigma^*], \tag{16}$$

where $\kappa \approx 0.6626$, and σ_* and σ^* are two distinct lower and upper bounds on noise that straddle $\underline{\sigma}$, such that $c(\sigma_*) = c(\sigma^*) = 1$.

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

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

Condition T is always satisfied in the balanced case: there, $x^*(\sigma^*) = (\theta_g + \theta_b)/2$. Indeed, [T] can be viewed as a restriction on the extent to which β can depart from 1

[10]With bounded *replacement*, multiple responses are more compelling. An agent located in one of the two retention zones, but close to the replacement zone, could be indifferent between small and large noise.

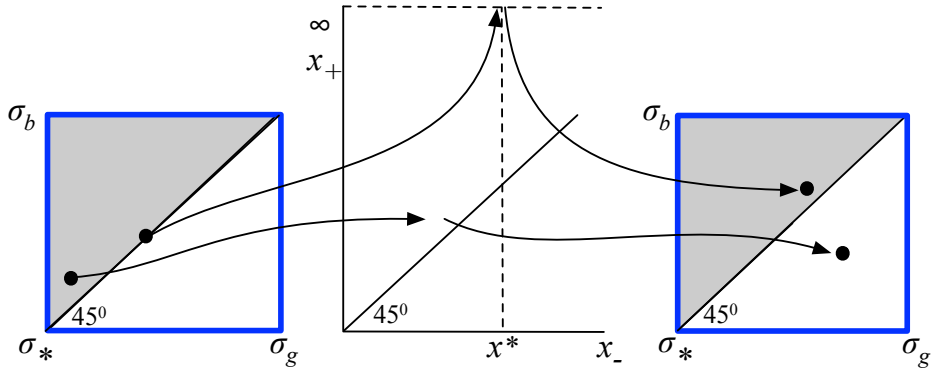


Figure 5. Fixed-Point Mapping to Show Existence of Bounded Retention

on either side of the balanced case. Subtract the formula for $x^*(\sigma^*)$ — see (5) — from θ_b and then θ_g to obtain an equivalent form of [T]:

$$-\frac{(\theta_g-\theta_b)^2}{2\sigma^{*2}} \leq \ln(\beta) \leq \frac{(\theta_g-\theta_b)^2}{2\sigma^{*2}}. \tag{17}$$

Proposition 3. *Under Conditions* U *and* T*, a bounded retention equilibrium exists.*

The proof has economic intuition, so we outline it. The first box in Figure 5 shows the domain of a fixed-point mapping, with agent noise lying between σ_* and σ^*. The mapping is derived as follows: for each (σ_g, σ_b), find the principal's retention decision, shown in the middle graph (x_- and x_+), and then record the best response to that decision, shown by the continuation mapping into the last box, a replica of the first.

The problem is that this fixed point mapping is not well-behaved. For any $\sigma_b < \sigma_g$, the principal best-responds with bounded *replacement*, and the "subsequent" response that completes the mapping is generally discontinuous in (σ_g, σ_b). This problem is endemic. However, to probe the existence of a bounded retention equilibrium, we can start from a smaller domain: the shaded triangle over which $\sigma_b \geq \sigma_g$. On this subdomain, the principal chooses bounded retention (or a monotone threshold), and the subsequent best response by the agents is unique (by Condition U) and continuous. While in general, the mapping could slip out of the smaller domain (see lower pair of arrows in Figure 5), Condition (17) guarantees that this cannot happen.

To see why, study the upper pair of arrows in Figure 5. The first arrow maps a point on the principal diagonal (where $\sigma_b = \sigma_g$) to a monotone retention regime; that is,

(x_-, x_+) is of the form (x^*, ∞). By (17), x^* must lie between θ_b and θ_g. In response, the good type will want to reduce noise as much as possible, while the bad type will want to increase it. Therefore $\sigma_g < \underline{\sigma}$, while the opposite is true of the bad type. But that implies $\sigma_b > \sigma_g$, which takes us back into the starting subdomain from its boundary. A fixed point theorem due to Halpern and Bergman (1968) then completes the argument, establishing the existence of a bounded retention equilibrium when β does not take on "extreme" values.

In summary, when the future is neither too optimistic nor too pessimistic — and certainly when it is balanced — a bounded retention equilibrium must exist. Indeed, under additional conditions, it is the only type of equilibrium. For instance, assume a sizable difference between the two types; specifically, that

$$\theta_g - \theta_b \geq \sigma^*, \tag{18}$$

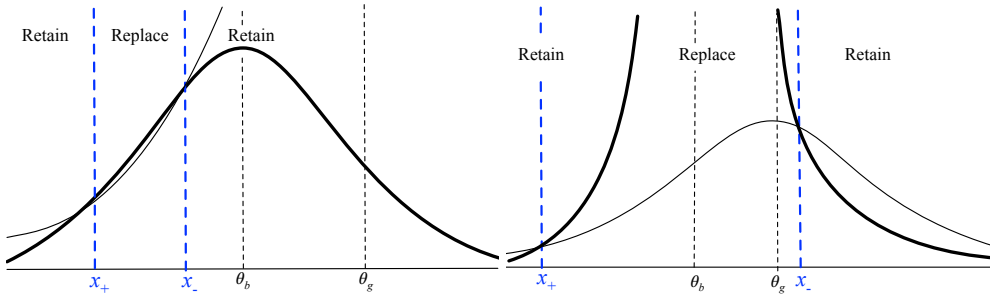
where recall that σ^* is defined by the larger of the two solutions to $c(\sigma) = 1$.

Proposition 4. *Assume* (17) *and* (18)*. Then only bounded retention equilibria exist.*

The argument emphasizes the location of types relative to replacement and retention zones. When the conditions for Proposition 4 fail, Figure 6 shows how bounded replacement might arise. The density for the bad type is the thicker line in both panels. The figure shows that β must be so large or so small (that is, the future is either so optimistic or so pessimistic) that the intersections of the two weighted densities are either on one side of both the mean types, or straddle them both. These are the only two possible kinds of bounded replacement equilibria. The Appendix provides complete numerical examples for each case.[11]

6.3. **Dynamics With Term Limits.** We solve for the "outside option probability" p in a dynamic setting where each agent has a two-period "term limit." We consider

[11]In the first kind of bounded replacement equilibrium, both types are in the retention zone as in Panel A of Figure 6, with $x_+ < x_- < \theta_b < \theta_g$. 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 θ_b and θ_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: $p \gg q$.



(a) $\beta \gg 1$ **(b)** $\beta \ll 1$

Figure 6. Possible Configurations for Bounded Replacement Equilibria

stationary equilibrium, in which every new agent of a given type takes the same action for the same value of p. Given σ_k for each type, and a realization x, the update on q is

$$q(x) := \frac{q\pi_g(x)}{\pi(x)}, \tag{19}$$

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

Start with prior q. At the end of term 1, a signal x is generated, and q is updated to $q(x)$. If V is the lifetime payoff to the principal starting from a fresh agent, the retention zone X is the set of all x for which $(1-\delta)M(q(x)) + \delta V \geq V$, where for any q', $M(q') := q'U_g + (1-q')U_b$ is the expected flow payoff to the principal when her current prior is q'. Let $\Pi_k := \int_X \pi_k(x)dx$ be the type-dependent probability of retention, and $\Pi := q\Pi_g + (1-q)\Pi_b$ the overall probability of retention. Then

$$\begin{aligned} V &= (1-\delta)M(q) + \delta \int_X \left[(1-\delta)M(q(x)) + \delta V\right]\pi(x)dx + \delta \int_{X^c} V\pi(x)dx \\ &= (1-\delta)\left[q(1+\delta\Pi_g)U_g + (1-q)(1+\delta\Pi_b)U_b\right] + \delta\left[1-(1-\delta)\Pi\right]V. \end{aligned}$$

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

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

We can rewrite this expression to obtain a "general equilibrium formula" for β:

$$\beta = \frac{q}{1-q}\frac{1-p}{p} = \frac{1+\delta\Pi_b}{1+\delta\Pi_g}. \tag{20}$$

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

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

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

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

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

The familiar monotone likelihood ratio property (MLRP) guarantees that when two types transmit with the *same* noise, higher signals are increasingly likely to be associated with the higher type; that is $f(z-a)/f(z)$ is increasing in z whenever $a > 0$. We assume a stronger version of this, which is automatically satisfied in the normal case, and guarantees a single, finite threshold for retention when both types use the same noise, no matter how optimistic or pessimistic the principal's prior is regarding agent types:

[12]Suppose $\beta = 1$. Then $p = q$, and we know that in the static model only bounded retention equilibria are possible. But in that situation the principal can *strictly* discriminate in favor of the good type, since there will always exist two distinct real roots to (4). But now $\Pi_g > \Pi_b$, which contradicts our starting point that $\beta = 1$.

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

$$\lim_{z\to\infty}\frac{f(z-a)}{f(z)}=\infty \text{ and } \lim_{z\to-\infty}\frac{f(z-a)}{f(z)}=0. \tag{21}$$

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

$$\beta f\left(-\frac{\theta_g-\theta_b}{\underline{\sigma}(\beta)}\right)\equiv f(0), \tag{22}$$

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

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

(i) *A bounded replacement equilibrium cannot exist.*

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

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

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

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

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

(i) *There exists $\hat{\underline{\sigma}} > 0$ such that a nontrivial equilibrium exists if and only if $\underline{\sigma} \in (0, \hat{\underline{\sigma}})$. When it exists, the equilibrium is unique.*

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

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

6.5. **Signal-Contingent Disclosure.** Consider a variation in which the agent observes the signal realization and can choose whether to disclose it.[13] For concreteness, think of the agent as a supplier who observes the quality of a produced good and then decides whether or not to ship it to a buyer, the principal. There is also an exogenous probability $\zeta > 0$ that the good is not received, independent of the agent's decisions.

Nontransmission of the signal (or good) is costly to the agent, who then pays a penalty $\kappa > 0$. To remove tedious knife-edge cases, assume $\kappa \neq 1$, where recall that 1 is the normalized value of retention for either type of agent. An unsent signal could also be costly to the principal, but that will not affect her retention decision, so we ignore it. The principal's action is $r : \mathbb{R} \cup \{n\} \to \{0, 1\}$, which indicates whether she retains the agent after observing signal $x \in \mathbb{R}$ or no signal (n). The agent choose variances σ as before, as a function of her type, but *also* a disclosure rule $d : \{g, b\} \times \mathbb{R} \to \{0, 1\}$.

Proposition 8. (i) *Suppose that the future is optimistic, so that $q < p$. Then in any sequential equilibrium, each type of agent must intend to transmit every signal, and the principal replaces in the (accidental) event that no signal is received. All other actions by both principal and agent coincide exactly with those made in the benchmark model.*

(ii) *Suppose that the future is pessimistic, with $q > p$. If $\kappa > 1$, then again, each agent type intends to transmit every signal, but now the principal retains in the (accidental) event that no signal is received. All other actions by both principal and agent coincide*

[13] We are grateful to a referee for suggesting this extension.

exactly with those made in the benchmark model. If $\kappa < 1$, the agent is always retained along the equilibrium path, signal or no signal.

(iii) *The balanced case $q = p$, both the scenarios in (i) and (ii) are possible equilibria.*

So the results of our benchmark model can be applied with no change to a setting in which the sender has the power to suppress his signal ex post, provided either that $q < p$, or (even if $q > p$) that the cost of non-disclosure is large. It is only when non-disclosure costs are small *and* $q > p$ that agent replacement is no longer observed in equilibrium, which is reminiscent of the "always retain" trivial equilibrium from the benchmark model. In this case, the retention sets could take many arbitrary forms, going well beyond the retention regimes identified in the Section 4. That is because sequential equilibrium imposes no discipline in the face of an observed signal realization which was never supposed to be sent. Yet, under a simple and natural refinement, the two retention sets can be shown to coincide — see the Supplementary Appendix.

6.6. **Other Extensions.** In the the Supplementary Appendix, we consider additional extensions of our rich and tractable baseline model. We mention them here.

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

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

perfectly identify the agent type *ex post*, the principal gives both types a chance to be retained, thus incentivizing them to exert effort.

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

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

Commitment. The case of principal commitment to retention mechanisms is the subject of a separate paper; see Vohra ⓡ Espinosa ⓡ Ray (2021). We consider a more general mechanism design problem within a setting studied by Ray ⓡ Vohra (2020). A version of our model, but with principal commitment, emerges as a special case of that framework. The main result is that the principal gains nothing from commitment relative to a model with no commitment, such as the one studied in this paper.

7. Applications

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

The potential relevance of our model should be assessed by the following considerations: (a) whether the choice of action by the agent corresponds, at least in part, to obscure or clarify his ability, (b) whether the resulting choice of noise cannot be observed ex ante by the principal, *at least in part*, and (c) whether the outcome, apart from being intrinsically good or bad, serves ex post as an indicator for the extent of risk-taking, thereby leading to some form of inference about the agent's competence. It is important to appreciate the emphasized phrase in part (b). It is only necessary that there be some significant unobserved component to the choice of risk, not that every aspect of that choice be unobserved.

The discussion that follows is only suggestive of some useful directions, and is not intended to serve as a formal empirical investigation of our model.

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. But even though there are persistent differences in managerial skill across funds (Chevalier and Ellison, 1999; Berk and van Binsbergen, 2015), 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).

Moreover, it is well known that underperforming funds inject additional risk into their portfolios in the hope of catching up with the winners (Brown et al., 1996; Chevalier and Ellison, 1997; Koski and Pontiff, 1999; Dasgupta and Prat, 2006) — a strategy colorfully referred to as "gambling for resurrection." For instance, the *New York Times*

article (March 7, 2014) that we refer to in the Introduction, urges the reader to "look beyond [immediate] results" when evaluating a fund manager. It recommends instead that managerial returns are best studied over an entire business cycle, so as to better assess the extent of risk-taking. The article observes that "someone who took very little risk to get to an 8 percent return is better off than someone who made 8 percent but should have made 12 percent given the amount of risk in the investments." Of course, that's easier said than done. 60% percent of the approximately 7500 mutual funds in the United States in 2014 — the year of the article — were launched in the preceding 10-year period; a third in the preceding 5-year period. Often, there is not enough history to make this assessment.

But even if there is history, the assessments are perforce limited. Some smoking guns should be obvious (or perhaps not even these), such as a self-declared large-cap manager who might gamble with small-cap stocks in his portfolio, or a bond fund holding equities. Indeed, investors may not have easy access to direct 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.[14] It is generally hard for investors to infer the level of risk in a given portfolio both ex-ante and ex-post (for other references, see Kritzman, 1987; Sirri and Tufano, 1998; Fama and French, 2010), and ex-post performance must be used for evaluation.

One might respond that it is always possible to trim or cap large positive returns ex ante or ex post, so that the results of our paper do not apply even if risk is unobserved. From the ex ante perspective, a manager could sell call options at judiciously chosen strike prices to offset large positive swings, or simply hold cash. But the point is that he cannot remove or cap his risk *for free*; it will need to happen at the expense of a lower average rate of return. A risk-cutting strategy might statistically reveal that he is unable to generate the same average return as a top-flight manager. In his need

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

to imitate a high type, a low-type manager might willingly load on risk even if it is feasible for him not to do so. The *feasibility* of ex ante risk-capping does not imply its *optimality* under equilibrium interaction — or at least, that is what the model predicts.

From an ex post perspective, it might be argued that a manager can always dispose of excess positive returns that might arouse suspicion. Indeed, there is evidence that fund managers move to safer assets during their portfolio disclosure period to suggest that they are taking on less risk than they are. Haugen and Lakonishok (1988) view this as a component of the so-called "January effect," and argue that both riskier and losing stocks and bonds are undervalued at year-end because they are shed from portfolios during disclosure periods. Musto (1999) studies the specific case of such "window dressing for safety" in the case of retail money funds: allocations to government assets are larger during disclosure periods than in other weeks.[15] In short, money managers attempt to project the image of a safe portfolio during a disclosure period, while loading up on risk at other times of the year. But such attempts do not come for free. The outcomes of many investments are hard to micro-manage, depending as they do on market conditions that are well beyond the manager's control. Some risk-taking choices are irreversible. Moreover, one can look not just at overall behavior but *patterns* of that behavior over time — just as we have been doing here. A protracted attempt to cover up excessive risk-taking may be quite visible to the careful investor.

In summary, some good outcomes could, or should, be viewed as *a priori* suspects for excessive risk taking. As Chevalier and Ellison (1997) observe: "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." While those high returns are a current positive, our investor's goal is to find a manager who will also deliver high expected returns in the future. If the current return is also a signal for excessive risk-taking, the manager's competence could be questioned.

Taken literally, 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 excess current return should eventually increase. In Hvide (2002), the CEO of Skandia Fund Management confesses to the author that the

[15]Though weekly data on fund allocations are obtained in the Money Fund Report newsletter, the subscription price is high enough that most retail investors lack access to this information. Or at least, fund managers appear to bet on that fact.

Fund "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." 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),[16] which appears to go in favor of "monotone retention regimes." Our theory does not rule out such regimes, especially in the mutual fund industry, where 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 the latter differences are "small" (see Fama and French, 2010).

7.2. **Risky Politics.** The unobservability of risk is a salient feature of situations when the observer either does not fully know or cannot judge the full set of consequences (and associated likelihoods) of an observed action. This is true, for instance, of risky political actions. An observer might be able to estimate the risk that political actors of different competencies are likely to be taking, just as an agent computes equilibrium play from her beliefs about opponent strategies, but at the same time not actually *observe* that risk. Perhaps voters can not fully comprehend (or are under-informed about) the implications of a given policy, much as our investor in the previous example.

This last 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. To add to this, there is substantial 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. In effect, and in the language of this paper, a voter may not fully observe the risk of a policy.

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

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 sharper will be the estimate 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).

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 pf necessity. However, when observabilty 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, such a success were to occur — should be treated with a certain degree of reticence by the median voter. It could be a sign of extreme competence. It could also be sign of a desperate move by a largely incompetent individual, which happened to pay off. That outcome, if it occurs, may be good for society. But it may not be a good signal on which to base re-election.

8. SUMMARY

We've studied a model in which an agent who seeks to be retained by a principal might deliberately inject noise into a process that signals his type. Possible equilibrium regimes include monotone retention, in which a principal retains if an agent's signal is high enough, and various non-monotone regimes. Of these, we argue that *bounded retention* is the salient equilibrium regime. In it, different types of agents

choose different degrees of noise, with worse agents behaving more noisily. The resulting equilibrium has a "double-threshold" property: the principal retains the agent if the signal is good, but neither too good nor too bad. We discuss extensions to a variant with costly noise, to a dynamic version with agent term limits, and to non-normal signal structures.

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

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

APPENDIX: MAIN PROOFS

Proof of Proposition 1. The proof of this proposition is long and contains several steps, with many technical details relegated to the Supplementary Appendix. Recall that the discussion in Section 5.1 eliminates all bounded replacement equilibria. With that out of the way, we focus on monotone and bounded retention regimes, and agent responses to them.

Lemma 1. *With bounded retention, $\sigma_b > \sigma_g$, and $X = [x_-, x_+]$, where $\theta_g < \frac{x_- + x_+}{2} < x_+$.*

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

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

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

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

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

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

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

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

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

which establishes (23). ∎

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

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

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

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

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

$$\sigma_k^* = \sqrt{\frac{(x_+ - x_-)(\frac{x_- + x_+}{2} - \theta_k)}{\ln(x_+ - \theta_k) - \ln(x_- - \theta_k)}} \in (x_- - \theta_k, x_+ - \theta_k)\,, \tag{25}$$

and agent k *sets* $\sigma_k = \max\{\underline{\sigma}, \sigma_k^*\}$.

Proof. (i) In the case of monotone retention, the first-order derivative with respect to σ_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 $\sigma_k = \underline{\sigma}$; always positive if $x^* > \theta_k$, so the agent always wants to increase the noise and the problem has no solution; and always equal to 0 if $x^* = \theta_k$, so the agent is indifferent across all choices of σ.

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

$$\Phi\left(\frac{x_+ - \theta_k}{\sigma_k}\right) - \Phi\left(\frac{x_- - \theta_k}{\sigma_k}\right), \tag{26}$$

where Φ is the cdf of the standard normal. The first-order derivative of the objective function with respect to σ_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 (24). 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 $\sigma_k = \underline{\sigma}$.

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

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

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

$$\sigma_k^* = \sqrt{\frac{(x_+ - x_-)\left(\frac{x_- + x_+}{2} - \theta_k\right)}{\ln(x_+ - \theta_k) - \ln(x_- - \theta_k)}}$$

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

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

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

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

Proof. From (4) it is clear that the principal employs a monotone retention regime if and only if both agent types choose the same level of noise, $\sigma_g = \sigma_b = \sigma$. In fact, by Lemma 3(i), $\sigma_g = \sigma_b = \underline{\sigma}$. Otherwise, a non-trivial equilibrium must have bounded retention, in which case $\sigma_g < \sigma_b$ by Observation 1. Suppose, on the contrary, that $\underline{\sigma} < \sigma_g$. Then both choices of noise are interior, and so agent optimality requires

$$\begin{aligned}
\phi\left(\frac{x_- - \theta_b}{\sigma_b}\right)(x_- - \theta_b) &= \phi\left(\frac{x_+ - \theta_b}{\sigma_b}\right)(x_+ - \theta_b), \\
\phi\left(\frac{x_- - \theta_g}{\sigma_g}\right)(x_- - \theta_g) &= \phi\left(\frac{x_+ - \theta_g}{\sigma_g}\right)(x_+ - \theta_g).
\end{aligned}$$

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 Ψ on $(\underline{\sigma}, \infty)$, 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, $\sigma_g = \underline{\sigma}$.) Guided by part (iii) of Lemma 3, our self-map is:

$$\Psi(\sigma) \equiv \max\left\{\sqrt{\frac{[x_+(\sigma) - x_-(\sigma)]\left(\frac{x_-(\sigma)+x_+(\sigma)}{2} - \theta_b\right)}{[\ln(x_+(\sigma) - \theta_b) - \ln(x_-(\sigma) - \theta_b)]}}, \underline{\sigma}\right\}, \tag{27}$$

where for any $\sigma > \underline{\sigma}$,

$$x_-(\sigma) := \frac{\sigma^2\theta_g - \underline{\sigma}^2\theta_b - \sigma\underline{\sigma}R(\sigma)}{\sigma^2 - \underline{\sigma}^2} \text{ and } x_+(\sigma) := \frac{\sigma^2\theta_g - \underline{\sigma}^2\theta_b + \sigma\underline{\sigma}R(\sigma)}{\sigma^2 - \underline{\sigma}^2}, \tag{28}$$

with

$$R(\sigma) := +\sqrt{(\theta_g - \theta_b)^2 + (\sigma^2 - \underline{\sigma}^2)\, 2\ln\left(\beta\frac{\sigma}{\underline{\sigma}}\right)}. \tag{29}$$

To interpret these objects, notice that $x_-(\sigma)$ and $x_+(\sigma)$ are the roots to

$$\beta\frac{1}{\underline{\sigma}}\phi\left(\frac{x - \theta_g}{\underline{\sigma}}\right) = \frac{1}{\sigma}\phi\left(\frac{x - \theta_b}{\sigma}\right), \tag{30}$$

so these bound the retention regime X when the principal expects $(\sigma_b, \sigma_g) = (\sigma, \underline{\sigma})$. (We will verify that these bounds are well-defined.) Given these thresholds, type b reacts as in Lemma 3(iii). So $\Psi(\sigma)$ can be interpreted as b's reaction to a chain that starts with a conjecture about b's action (σ), travels via the principal's thresholds, and culminates in that type's optimal reaction to those thresholds. Hence a fixed point of Ψ 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 $\sigma > \underline{\sigma}$. The following lemma relates this to condition (11).

Lemma 5. *If $\beta \geq 1$, $x_-(\sigma)$ and $x_+(\sigma)$ are well-defined and distinct for $\sigma > \underline{\sigma}$. If $\beta < 1$, $x_-(\sigma)$ and $x_+(\sigma)$ are well-defined and distinct for $\sigma > \underline{\sigma}$ if and only if (11) holds.*

Proof. When $\beta \geq 1$, it is clear that the term within the square root in (29) is strictly positive for all $\sigma > \underline{\sigma}$. In the Supplementary Appendix we show that, when $\beta < 1$, this term is strictly positive for all $\sigma > \underline{\sigma}$ if and only if (11) holds. ∎

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

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

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

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

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

$$\phi\left(\frac{x_+ - \theta_b}{\sigma}\right)(x_+ - \theta_b) = \phi\left(\frac{x_- - \theta_b}{\sigma}\right)(x_- - \theta_b)\,. \tag{31}$$

It follows that

$$\begin{aligned} d_g(\underline{\sigma}) &= \phi\left(\frac{x_- - \theta_g}{\underline{\sigma}}\right)(x_- - \theta_g) - \phi\left(\frac{x_+ - \theta_g}{\underline{\sigma}}\right)(x_+ - \theta_g) \\ &= \frac{\underline{\sigma}}{\beta}\left[\phi\left(\frac{x_- - \theta_b}{\sigma}\right)\frac{x_- - \theta_g}{\sigma} - \phi\left(\frac{x_+ - \theta_b}{\sigma}\right)\frac{x_+ - \theta_g}{\sigma}\right] \\ &= \frac{\underline{\sigma}}{\beta}\frac{x_- - \theta_b}{\sigma}\phi\left(\frac{x_- - \theta_b}{\sigma}\right)\left[\frac{x_- - \theta_g}{x_- - \theta_b} - \frac{x_+ - \theta_g}{x_+ - \theta_b}\right] < 0, \end{aligned}$$

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

With the good type dealt with, we return to the fixed point problem for the bad type. In preparation for the steps ahead, the two retention thresholds $x_-(\sigma)$ and $x_+(\sigma)$ are

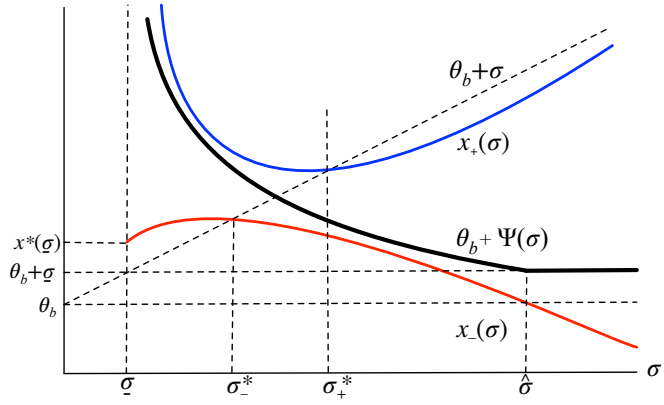


Figure 7. Principal's best responses $x_-(\sigma)$ and $x_+(\sigma)$ and type b's counter-response.

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

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

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

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

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

Proof. See Supplementary Appendix. ∎

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

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

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

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

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

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

$$\phi\left(\frac{x_+(\sigma)-\theta_b}{\Psi(\sigma)}\right)(x_+(\sigma)-\theta_b) = \phi\left(\frac{x_-(\sigma)-\theta_b}{\Psi(\sigma)}\right)(x_-(\sigma)-\theta_b). \tag{32}$$

Equation (32) can be used to compute $\Psi'(\sigma)$. The Supplementary Appendix indicates the steps and shows that this derivative is strictly negative at any fixed point. So $\Psi(\sigma)$ is strictly decreasing at any fixed point, and therefore can have just one fixed point $\sigma^\dagger > \underline{\sigma}$, 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 $\sigma^\dagger > \underline{\sigma}$ 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 $\underline{\sigma}$. Because both types must play a best response, it follows from Lemma 3(i) that

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

or

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

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

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

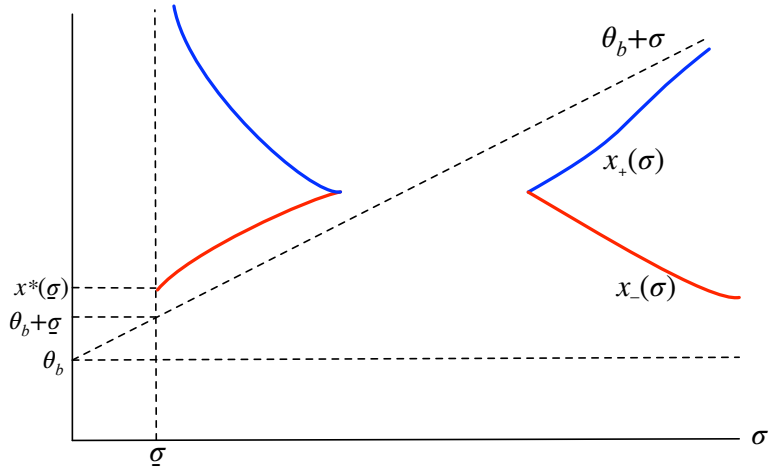


Figure 8. $x_-(\sigma)$ and $x_+(\sigma)$ are not always well-defined if (11) fails.

Proof. If $\beta \geq 1$ and (12) fails, then $x_-(\sigma) < \theta_b$ for all $\sigma > \underline{\sigma}$ 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 $\sigma \geq \underline{\sigma}$. So by Lemma 3 (ii), $\Psi(\sigma) = \underline{\sigma}$ for all $\sigma > \underline{\sigma}$ and has no fixed point with $\sigma > \underline{\sigma}$.

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

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

Lemma 10. *If $\beta < 1$ and (11) fails, a non-trivial equilibrium does not exist.*

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

bounded retention equilibrium is possible. See the Supplementary Appendix for the formal details. ■

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

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

$$\phi(z_1)z_1 = \phi(z_2)z_2 = -\sigma c'(\sigma), \tag{33}$$

where $z_1 = (\sigma/\Delta)\ln(\beta) - (\Delta/2\sigma)$ and $z_2 = (\sigma/\Delta)\ln(\beta) + (\Delta/2\sigma)$. The function $\phi(z)z$ has the shape shown in Figure 4, reaching maxima and minima at $z = 1$ and $z = -1$ respectively, and exhibiting "negative symmetry" around 0. Using (15), this tells us that there are two exclusive possibilities: (i) either $\beta > 1$ and $\sigma < \underline{\sigma}$, or (ii) either $\beta < 1$ and $\sigma > \underline{\sigma}$. 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 σ (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 ∞ as $\sigma \to \infty$. Meanwhile, z_1 is always increasing in σ, 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 σ such that the *first* equality in (33) holds, and it is independent of the cost function c. ■

Proof of Proposition 3. Recall that $\sigma_* < \underline{\sigma}$ and $\sigma^* > \underline{\sigma}$ 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 $\boldsymbol{\sigma} \in \Sigma^+$, 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 Ψ_1. As discussed in the

main text, this map is well-defined when $\sigma_b = \sigma_g$. To check that Ψ_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

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

is strictly positive. But (17) tells us that $\ln(\beta) \geq -[\Delta^2]/2\sigma^{*2}$, and so

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

where the very last inequality uses $\sigma^* \geq \sigma_b > \sigma_g$. So there are distinct roots $x_- < x_+$, and by exactly the same logic as for Observation 1, the zone $[x_-, x_+]$ must involve retention.

Next, for each pair (x_-, x_+) with $x_+ > x_-$ and with x_+ possibly infinite, define (σ_b', σ_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 Ψ_2.

Finally, define a map Ψ with domain Σ^+ and range Σ by $\Psi \coloneqq \Psi_2 \circ \Psi_1$. We claim that Ψ is continuous. We first argue that Ψ_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 $\Psi_1(\sigma_g^n, \sigma_b^n) \to \Psi_1(\sigma_g, \sigma_b)$.

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

Now we turn to the map Ψ_2. As already mentioned, condition [U] guarantees that best-response noise choices are unique, as long as $x_+ > x_-$. They are fully characterized by the first-order condition (13), which we reproduce here for convenience:

$$\phi\left(\frac{x_- - \theta_k}{\sigma_k}\right)\left(\frac{x_- - \theta_k}{\sigma_k}\right) - \phi\left(\frac{x_+ - \theta_k}{\sigma_k}\right)\left(\frac{x_+ - \theta_k}{\sigma_k}\right) = \sigma_k c'\left(\sigma_k\right) \tag{34}$$

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_-, ∞). Let σ_k^n be the best responses for an agent of type k, and let σ_k be the best response at the limit value (x_-, x_+). When $x_+ < \infty$, it is obvious from (34) 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 $\phi(z^n)z^n \to 0$ for any sequence $z^n \to \infty$.

We claim that Ψ is *inward pointing*; that is, for every $(\sigma_g, \sigma_b) \in \Sigma^+$, there exists $a > 0$ such that

$$(\sigma_g, \sigma_b) + a[\Psi(\sigma_g, \sigma_b) - (\sigma_g, \sigma_b)] \in \Sigma^+. \tag{35}$$

First observe that for every $(\sigma_g, \sigma_b) \in \Sigma^+$, we have $(\sigma_*, \sigma_*) \leq \Psi(\sigma_g, \sigma_b) \leq (\sigma^*, \sigma^*)$. Therefore, if $(\sigma_g, \sigma_b) \in \Sigma^+$ with $\sigma_b > \sigma_g$, (35) is easily seen to hold: for $a > 0$ and small, it must be that both components of the vector

$$(\sigma_g, \sigma_b) + a[\Psi(\sigma_g, \sigma_b) - (\sigma_g, \sigma_b)]$$

lie in $[\sigma_*, \sigma^*]$, and the second component is larger than the first. The remaining case is one in which $(\sigma_g, \sigma_b) \in \Sigma^+$ with $\sigma_b = \sigma_g$. In this case, we know from condition (17) that $\Psi_1(\sigma_g, \sigma_b)$ is of the form $(x_-, x_+) = (x^*, \infty)$, where $x^* \in [\theta_b, \theta_g]$. From the first-order conditions that describe each type — see (14) — it is easy to see that $\sigma_k \gtrless \underline{\sigma}$ when $x^* \gtrless \theta_k$. 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 σ_* and σ^*). It follows that for every $a \in (0, 1)$, (35) holds, and the claim is proved.

To summarize, we have: Σ^+ is a nonempty, compact, convex subset of Euclidean space, and Ψ is continuous on Σ^+. In general, however, Ψ will fail to map from from Σ^+ to Σ^+. 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 (σ_g, σ_b), along with the associated bounded retention zone $\Psi_1(\sigma_g, \sigma_b)$, forms an equilibrium. ■

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

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

$$\phi\left(\frac{x_- - \theta_k}{\sigma_k}\right) > \phi\left(\frac{x_+ - \theta_k}{\sigma_k}\right). \tag{36}$$

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

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

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

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

which implies, by single-peakedness and symmetry of ϕ around 0, and $x_+ < x_-$, that

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

which establishes (36). ∎

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

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

$$\left(\sigma_g^2 - \sigma_b^2\right)x^2 + 2\left(\sigma_b^2\theta_g - \sigma_g^2\theta_b\right)x + \left(\sigma_g^2\theta_b^2 - \sigma_b^2\theta_g^2 + 2A\sigma_g^2\sigma_b^2\right) > 0, \tag{37}$$

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

$$\beta < \frac{\sigma_g}{\sigma_b}\exp\frac{\Delta^2}{2\sigma_g^2},$$

but this is guaranteed by the right hand inequality of (17), because $\sigma^* \geq \sigma_g > \sigma_b$. Therefore θ_b lies in the interior of the replacement zone, or put another way, $x_+ < \theta_b < x_-$.

Now putting $x = \theta_g$ in (37) and simplifying, we see that retention is strictly optimal at θ_g if

$$\frac{\Delta^2}{2\sigma_b^2} + \ln(\sigma_b) - \ln(\sigma_g) > -\ln(\beta). \tag{38}$$

The derivative of the left hand side of (38) with respect to σ_b is given by

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

which is strictly negative given (18) and $\sigma_b \leq \sigma^*$, so it follows that the left hand side of (38) is minimized by setting $\sigma_b = \sigma_g = \sigma^*$. To establish (38), then, it is sufficient to have

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

but this is guaranteed by the left hand inequality of (17). Consequently, the principal strictly prefers to retain the agent if she observes $x = \theta_g$. Given $x_+ < \theta_b < x_-$, this can only mean that $x_- < \theta_g$, and the proof is complete. ■

Proof of Proposition 4. In a bounded replacement equilibrium, $\sigma_g > \sigma_b$ and $x_- > x_+$. By Lemma 12, $\theta_g \geq x_- \geq \theta_b > x_+$. Define $B_k(\sigma)$ to be type-k's marginal benefit of noise:

$$B_k(\sigma) \coloneqq \phi\left(\frac{x_- - \theta_k}{\sigma}\right)\frac{x_- - \theta_k}{\sigma^2} - \phi\left(\frac{x_+ - \theta_k}{\sigma}\right)\frac{x_+ - \theta_k}{\sigma^2}. \tag{39}$$

Observe that for *every* σ,

$$\begin{aligned}
B_b(\sigma) &= \phi\left(\frac{x_- - \theta_b}{\sigma}\right)\frac{x_- - \theta_b}{\sigma^2} - \phi\left(\frac{x_+ - \theta_b}{\sigma}\right)\frac{x_+ - \theta_b}{\sigma^2} \\
&\geq \phi\left(\frac{x_+ - \theta_b}{\sigma}\right)\frac{x_- - \theta_b}{\sigma^2} - \phi\left(\frac{x_+ - \theta_b}{\sigma}\right)\frac{x_+ - \theta_b}{\sigma^2} \\
&= \phi\left(\frac{x_+ - \theta_b}{\sigma}\right)\frac{x_- - x_+}{\sigma^2} \\
&> \phi\left(\frac{x_+ - \theta_g}{\sigma}\right)\frac{x_- - x_+}{\sigma^2} \\
&= \phi\left(\frac{x_+ - \theta_g}{\sigma}\right)\frac{x_- - \theta_g}{\sigma^2} - \phi\left(\frac{x_+ - \theta_g}{\sigma}\right)\frac{x_+ - \theta_g}{\sigma^2} \\
&\geq \phi\left(\frac{x_- - \theta_g}{\sigma}\right)\frac{x_- - \theta_g}{\sigma^2} - \phi\left(\frac{x_+ - \theta_g}{\sigma}\right)\frac{x_+ - \theta_g}{\sigma^2} \\
&= B_g(\sigma),
\end{aligned} \tag{40}$$

where the first inequality follows from $x_- \geq \theta_b$ and inequality (36) of Lemma 11, the second inequality follows from ϕ single-peaked around zero and $x_+ - \theta_g < x_+ - \theta_b < 0$, and the last inequality follows from $x_- \leq \theta_g$ and (again) inequality (36) of Lemma 11.

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

Proof of Proposition 8. (i) Suppose that $q < p$. We first claim that $r(n) = 0$. Suppose not, so that $r(n) = 1$. Suppose first that $\kappa > 1$, where 1 is the normalized retention value to the agent. Then neither type will want to willingly hide any signal (it is too costly to do so). But that means that all non-transmitted signals are accidental, and the principal must must retain her prior on the agent in any sequential equilibrium, which is q. But $q < p$, so her optimal action must be to replace the agent, a contradiction. On the other hand, if $r(n) = 1$ and $\kappa < 1$, then there must be universal retention in equilibrium irrespective of signal — for if not, the agent could hide a signal and pick up a surplus of $1 - \kappa$. However, universal retention requires the posterior on the agent to exceed p in all events, which is impossible, given that $q < p$ to begin with.

So the claim is true, and $r(n) = 0$. It follows that either type of agent must send every realized signal, for nondisclosure results in both replacement *and* the penalty κ. Therefore in this case, there is no deliberate withholding of signals. In short, all signals are sent, and we have a perfect embedding of this case into our baseline model, with the additional proviso that the principal replaces in case a signal accidentally fails to arrive. There are no other equilibria with $r(n) = 0$.

(ii) Suppose that $q > p$. We claim that $r(n) = 1$. For suppose on the contrary that $r(n) = 0$. Then no type would ever deliberately hide a signal — this results in a penalty *and* in replacement. Therefore all non-disclosure events must be accidental, forcing the principal to retain her prior in any sequential equilibrium, which is q. But $q > p$, so her equilibrium action must be to set $r(n) = 1$, a contradiction.

(iia) $q > p$ and $\kappa > 1$. Then intentional non-disclosure isn't worth it, given the high cost κ, so both types send all signals. Again, our baseline model describes all

equilibria, with the principal retaining if a signal accidentally fails to arrive. There are no other equilibria.

(iib) $q > p$ and $\kappa < 1$. In this case all signal realizations $\widetilde{x}$ with $r(\widetilde{x}) = 0$ must be hidden. So no agent type can ever replaced in this case: $r(n) = 1$ and $r(\widetilde{x}) = 1$ for all observed signals.

Moreover, observe that both types must transmit exactly the same set of signal realizations. For if only the bad type sends a particular $\widetilde{x}$, then $r(\widetilde{x}) = 0$, a contradiction. And if $\widetilde{x}$ is sent by only the good type, then $r(\widetilde{x}) = 1$, but then the bad type would want to send $\widetilde{x}$ as well, since this saves the cost κ. We can conclude that, in any equilibrium, each signal realization must be either hidden by both types or sent by both types, leading to retention in any case.

(iii) Depending on how we break indifference for the principal, the balanced case can fall into Cases (i) or (ii), and there is no other (pure-strategy) equilibrium. ■

Proof of Propositions 5, 6 and 7. See the Supplementary Appendix.

REFERENCES

ALESINA, A., AND A. CUKIERMAN (1990): "The Politics of Ambiguity," *The Quarterly Journal of Economics*, 105, 829–850.

ALIPRANTIS, C., AND K. BORDER (2006): *Infinite Dimensional Analysis: A Hitchhiker's Guide*: Springer, 3rd edition.

ARAGONES, E., AND Z. NEEMAN (2000): "Strategic Ambiguity in Electoral Competition," *Journal of Theoretical Politics*, 12, 183–204.

ARAGONES, E., T. PALFREY, AND A. POSTLEWAITE (2007): "Political Reputations and Campaign Promises," *Journal of the European Economic Assosiation*, 5, 846–884.

ARAGONES, E., AND A. POSTLEWAITE (2002): "Ambiguity in Election Games," *Review of Economic Design*, 7, 233–255.

ARNOLD, R. D. (1990): *The Logic of Congressional Action*: Yale University Press.

BANKS, J., AND R. SUNDARAM (1998): "Optimal Retention in Agency Problems," *Journal of Economic Theory*, 82, 293–323.

BARRON, D., G. GIORGIADIS, AND J. SWINKELS (2017): "Optimal Contracts with a Risk-Taking Agent," mimeo.

BAUM, M. A., AND S. KERNELL (1999): "Has Cable Ended the Golden Age of Presidential Television?" *American Political Science Review*, 93, 99–114.

BERK, J. B., AND J. H. VAN BINSBERGEN (2015): "Measuring Skill in the Mutual Fund Industry," *Journal of Financial Economics*, 118, 1–20.

BERK, J. B., AND R. C. GREEN (2004): "Mutual Fund Flows and Performance in Rational Markets," *Journal of Political Economy*, 112, 1269–1295.

BOLLEN, N. P. B., AND J. A. BUSSE (2001): "On the Timing Ability of Mutual Fund Managers," *Journal of Finance*, 56, 1075–1094.

BROWN, K. C., W. V. HARLOW, AND L. T. STARKS (1996): "Of Tournaments and Temptations: An Analysis of Managerial Incentives in the Mutual Fund Industry," *The Journal of Finance*, 51, 85–110.

CAMPBELL, J. (1983): "Ambiguity in the Issue Positions of Presidential Candidates: A Causal Analysis," *American Journal of Political Science*, 27, 284–293.

CARHART, M. M. (1997): "On Persistence in Mutual Fund Performance," *Journal of Finance*, 52, 57–82.

_____ (1999): "Are Some Mutual Fund Managers Better Than Others? Cross-Sectional Patterns in Behavior and Performance," *The Journal of Finance*, 54, 875–899.

CHEVALIER, J., AND G. ELLISON (1997): "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, 105, 1167–1200.

COHEN, L., A. FRAZZINI, AND C. MALLOY (2008): "The Small World of Investing: Board Connections and Mutual Fund Returns," *Journal of Political Economy*, 116, 951–979.

COVAL, J. D., AND T. J. MOSKOWITZ (2001): "The Geography of Investment: Informed Trading and Asset Prices," *Journal of Political Economy*, 109, 811–841.

CRAWFORD, V., AND J. SOBEL (1982): "Strategic Information Transmission," *Econometrica*, 50, 1431–1451.

DASGUPTA, A., AND A. PRAT (2006): "Financial Equilibrium with Career Concerns," *Theoretical Economics*, 1, 67–93.

DEGAN, A., AND M. LI (2016): "Persuasion with Costly Precision," mimeo, Department of Economics, Concordia University.

DELLI CARPINI, M. X., AND S. KEETER (1996): *What Americans Know about Politics and Why It Matters*: New Haven, CT: Yale University Press.

DEMARZO, P., I. KREMER, AND A. SKRZYPACZ (2019): "Test Design and Minimum Standards," *American Economic Review*, 109, 2173–2207.

DEWAN, T., AND D. MYATT (2008): "The Qualities of Leadership: Direction, Communication and Obfuscation," *The American Political Science Review*, 102, 352–368.

DUTTA, B., D. RAY, AND K. SENGUPTA (1989): "Repeated Principal-Agent Games with Eviction," in *The Economic Theory of Agrarian Institutions* ed. by Bardhan, P. Oxford: Clarendon Press, Chap. 5, 93–121.

EDMOND, C. (2013): "Information Manipulation, Coordination, and Regime Change," *Review of Economic Studies*, 80, 1422–1458.

FAMA, E. F., AND K. R. FRENCH (2010): "Luck versus Skill in the Cross Section of Mutual Fund Returns," *Journal of Finance*, 65, 1915–1947.

DE FIGUEIREDO, R., E. RAWLEY, AND O. SHELEF (2015): "Bad Bets: Excessive Risk-Taking, Convex Incentives, and Performance,"Technical report, Unpublished manuscript.

FINANCIAL CRISIS INQUIRY COMMISSION (2010): "The Financial Crisis Inquiry Report: Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States,"Technical report, Financial Crisis Inquiry Commission, Washington, DC.

GENAKOS, C., AND M. PAGLIERO (2012): "Interim Rank, Risk Taking, and Performance in Dynamic Tournaments," *Journal of Political Economy*, 120, 782–813.

GLAZER, A. (1990): "The Strategy of Candidate Ambiguity," *The American Political Science Review*, 84, 237–241.

GRUBER, M. J. (1996): "Another Puzzle: The Growth in Actively Managed Mutual Funds," *Journal of Finance*, 51, 783–810.

HALPERN, B. R. (1968): "A General Fixed-Point Theorem," in *Proceedings of the American Mathematical Society Symposium on Non-Linear Functional Analysis*, Chicago.

HALPERN, B. R., AND G. M. BERGMAN (1968): "A Fixed-Point Theorem for Inward and Outward Maps," *Transactions of the American Mathematical Society*, 130, 353–358.

HARBAUGH, R., J. MAXWELL, AND K. SHUE (2016): "Consistent Good News and Inconsistent Bad News," mimeo, Indiana University.

HAUGEN, R., AND J. LAKONISHOK (1988): *The Incredible January Effect : The Stock Market's Unsolved Mystery*: Homewood, IL: Dow Jones-Irwin, 3rd edition.

HENDRICKS, D., J. PATEL, AND R. ZECKHAUSER (1993): "Hot Hands in Mutual Funds: Short-Run Persistence of Relative Performance, 1974-1988," *The Journal of Finance*, 48, 93–130.

HVIDE, H. (2002): "Tournament Rewards and Risk Taking," *Journal of Labor Economics*, 20, 877–898.

IPPOLITO, R. A. (1992): "Consumer Reaction to Measures of Poor Quality: Evidence from the Mutual Fund Industry," *The Journal of Law & Economics*, 35, 45–70.

KACPERCZYK, M., C. SIALM, AND L. ZHENG (2005): "On the Industry Concentration of Actively Managed Equity Mutual Funds," *Journal of Finance*, 60, 1983–2011.

KAMENICA, E., AND M. GENTZKOW (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615.

KOSKI, J. L., AND J. PONTIFF (1999): "How Are Derivatives Used? Evidence from the Mutual Fund Industry," *The Journal of Finance*, 54, 791–816.

KRITZMAN, M. P. (1987): "Incentive Fees: Some Problems and Some Solutions," *Financial Analysts Journal*, 43, 21–26.

MAKAROV, I., AND G. PLANTIN (2015): "Rewarding Trading Skills without Inducing Gambling," *Journal of Finance*, 70, 925–962.

MATTHEWS, A., AND L. MIRMAN (1983): "Equilibrium Limit Pricing: The Effects of Private Information and Stochastic Demand," *Econometrica*, 51, 981–996.

MUSTO, D. (1999): "Investment Decisions Depend on Portfolio Disclosures," *Journal of Finance*, 54, 935–952.

PALOMINO, F., AND A. PRAT (2003): "Risk Taking and Optimal Contracts for Money Managers," *RAND Journal of Economics*, 34, 113–137.

PRIOR, M. (2007): *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections*: New York: Cambridge University Press.

RAY, D., AND A. ROBSON (2012): "Status, Intertemporal Choice, and Risk-Taking," *Econometrica*, 80, 1505–1531.

RAY, D. ⓡ R. VOHRA (2020): "Games of Love and Hate," *Journal of Political Economy*, 128, 1789–1825.

SHEPSLE, K. (1972): "The Strategy of Ambiguity: Uncertainty and Electoral Competition," *American Journal of Political Science*, 66, 555–568.

SHUMWAY, T., M. B. SZEFLER, AND K. YUAN (2011): "The Information Content of Revealed Beliefs in Portfolio Holdings," Unpublished working paper. University of Michigan, Ann Arbor, MI.

SIRRI, E. R., AND P. TUFANO (1998): "Costly Search and Mutual Fund Flows," *Journal of Finance*, 53, 1589–1622.

SOMIN, I. (2013): *Democracy and Political Ignorance: Why Smaller Government Is Smarter*: Stanford, CA: Stanford University Press.

SUBRAMANYAM, K. (1996): "Uncertain Precision and Price Reactions to Information," *The Accounting Review*, 71, 207–219.

TITMAN, S., AND B. TRUEMAN (1986): "Information Quality and the Valuation of New Issues," *Journal of Accounting and Economics*, 8, 159–172.

VOHRA, R. ⓡ F. ESPINOSA ⓡ D. RAY (2021): "A Principal-Agent Relationship with No Advantage to Commitment," *Pure and Applied Functional Analysis*, 6, 1043–1064.

ZHENG, L. (1999): "Is Money Mmart? A Study of Mutual Fund Investors' Fund Selection Ability," *Journal of Finance*, 54, 901–933.

ZWIEBEL, J. (1995): "Corporate Conservatism and Relative Compensation," *Journal of Political Economy*, 103, 1–25.