Equilibrium Persuasion*

Tom Cunningham, Inés Moreno de Barreda

March 10, 2015

Abstract

Signaling models are often used to explain the behavior of firms, politicians, and employees as attempts to persuade an observer. However in many models the effort spent on persuasion has no effect on the observer's subsequent action, because in equilibrium the observer rationally adjusts for the effort exerted. We show that when the receiver makes a binary decision, and signals are received with noise, signaling will in fact cause a systematic change in the receiver's choices, and the change will be in favor of the sender's preferred action. This occurs because the receiver will use a threshold rule, the threshold leads to bunching of signals just above the cutoff, and the bunching causes a skew distribution of posteriors, increasing the probability of the receiver taking the sender's preferred action. The paper additionally contributes an analytical solution to a class of noisy signaling models with many applications.

1 Introduction

Signaling and signal-jamming models have been used to explain a wide variety of economic behavior, especially why agents exert effort in the absence of explicit incentives.[1] For example why employees exert effort in the absence of contingent pay (Holmström (1999)); why politicians exert effort when voters are forward-looking (Rogoff and Sibert (1988)); and why firms charge below the monopoly price in the absence of competition (Milgrom and Roberts (1982)).

Although distortion of actions by senders is well-studied, many signaling and signal-jamming models predict no systematic distortion of the signal-*receiver*'s behavior. For example, in Rogoff and Sibert (1988) effort by incumbent politicians does not affect the equilibrium probability of reelection, and in Milgrom and Roberts (1982) effort by monopolists does not affect the probability of entry, compared to the full-information case.

*Cunningham: IIES Stockholm and Caltech, tom.cunningham@iies.su.se. Moreno de Barreda, Department of Economics and St Peter's College, Oxford University, ines.morenodebarreda@economics.ox.ac.uk. We thank Meg Meyer for helpful comments, as well as seminar audiences at Copenhagen, Konstanz, Oxford and Stockholm.

[1] In both classes of model agents wish to influence an observer's belief about some exogenous unobserved variable. In signaling models the agent has private information about the exogenous variable, in signal-jamming models they do not.

Milgrom and Roberts (1982) make a general point: "the observation of the [sender's] actions cannot, in equilibrium, systematically bias the [receiver's] expectations."[2] This holds in the sense that the mean of the posteriors over the unobserved variable must always be equal to the mean prior, independent of the signal structure. However in many cases the mean of the posteriors is not the quantity which is of most interest. In particular, if the receiver faces a binary decision, a change in the signal structure can change the equilibrium probability of each choice.

In this paper we study a noisy signaling model with a binary action and show that, not only is the ~~action~~ of the receiver affected by the sender's signaling, but the direction of the effect is systematic: signaling causes the receiver to choose the sender's preferred action more often, compared to the case where the sender cannot affect the signal.[3] In our motivating examples, signaling will increase the equilibrium probability of incumbents being re-elected, of internal candidates being promoted and of monopolists successfully deterring entry by challengers. The ability to manipulate a signal therefore confers a systematic advantage on signal senders; we refer to this result as "persuasion."

The paper relates to the literature on Bayesian persuasion. In Kamenica and Gentzkow (2011) and subsequent papers,[4] the sender can commit to a signal structure prior to learning their private information, and can take advantage of nonlinearity of payoff in the receiver's beliefs by choosing a skew signal structure. In contrast, in our model, the senders cannot commit, and yet the signal structure that arises endogenously in equilibrium is skewed to their advantage.[5]

Our model assumes a binary decision by the receiver, a uniform distribution of types, a convex additive cost to manipulate the signal, and symmetric additive noise with bounded support.[6] An advantage of assuming noise, besides generating a realistically smooth distribution of outcomes, is the existence of a unique equilibrium with rich testable implications. Models with noise are often regarded as relatively intractable because the sender's strategy is complicated, admitting only an implicit solution. In a similar setup Matthews and Mirman (1983) are only able to establish sufficient conditions for persuasion under specific asymmetric distributions of the noise.

Our advances on Matthews and Mirman (1983) are based on a new way of representing equilibrium. First we show that, although the sender's strategy does not have an analytic representation, its inverse does, i.e. a sender's type can be expressed as a function of the average signal that they send. This allows us to express many quantities of interest directly as integrals over realizations of the random noise. Given the uniform distribution of types

[2] Milgrom and Roberts (1982) do discuss the possibility of bias in a model without noise but say they expect no systematic effect: "the probability that entry actually occurs in equilibrium need not be any lower than it would be in a world of full information... [i]ndeed, the probability of entry in the limit pricing equilibrium may even be higher than with complete information"

[3] The comparison also holds with respect to the case with full information.

[4] There is an extensive recent literature on Bayesian Persuasion. See for example Kamenica and Gentzkow (2012); Gentzkow and Kamenica (2014); Alonso and Câmara (2014)

[5] The distribution of posteriors (or expected types) is 'skewed' in the sense that the median posterior is above the mean posterior.

[6] We assume that the sender's effort in manipulation has no intrinsic value to the receiver.

and the bounded support of noise these integrals admit simple solutions in terms of the cost function and density of noise. We can then get direct expressions for the receiver's beliefs, the probability the receiver takes the preferred action, and the sender and receivers' expected utility, each conditional on the threshold k used by the receiver, which itself has a simple analytic solution.

A rough intuition for our persuasion result comes from the fact that in equilibrium the receiver will use a threshold rule, and the density of signals sent will be characterized by a hill just above the threshold and a corresponding valley just below it. This implies that, conditional on receiving a signal exactly at the threshold, it is relatively more likely that it comes from a sender whose mean signal is above the threshold, i.e. a positive bias. The equilibrium threshold will therefore settle at a lower position than it would without the bunching of signals, and thus it will admit a greater fraction than would be admitted if no manipulation was possible.

We are also able to establish a number of additional results: (1) if the receiver can commit to a threshold *ex ante* then that threshold will be higher than the *ex post* threshold, moreover that threshold will lead to exactly zero equilibrium persuasion; nevertheless (2) the receiver is strictly better off when senders cannot manipulate their signal, even under commitment; and (3) when the cost of manipulation is a power function the senders' expected cost of manipulation exactly equals their *ex ante* benefit, this is in contrast to many signaling models in which the sender is strictly better off when manipulation is not possible (however if the receiver commits to a threshold then the senders are, as in the usual case, better off without manipulation).

We discuss some ~~of the many~~ applications of our model in Section 2. In Section 3 we state the model and prove the existence of a unique equilibrium, defined implicitly. We then show that, by inverting the senders' strategy, the distribution of signals can be characterized analytically as a function of the threshold. In Section 4 we derive an expression for the probability of admission. In Section 5 we solve for the equilibrium threshold, and derives the main result of the paper, equilibrium persuasion. In Section 6 we study the effects of manipulation on the utility of senders and receivers. Finally in Section 7 we summarize and conclude.

2 Applications

We believe that there exist many domains which feature noisy signaling and a binary payoff: elections, hiring, and market entry, among others. Much existing work in those domains has made the simplifying assumption either that the sender has no private information, or that the payoff is linear in the expected type, and then find that no persuasion occurs. We believe that those assumptions have been used, in part, because of a perception that signaling with binary actions is intractable or features multiplicity of equilibria; we therefore believe that our framework, which has a simple unique equilibrium, can be productively applied and extended in each of these literatures.

In political economy effort by politicians is often modeled as signal-jamming (in which

the politician does not know their type). The textbook by Persson and Tabellini (2002) discusses both signaling and signal-jamming models, but uses signal-jamming for most applications, explaining "[w]hich model is more satisfactory? The [signal jamming] model has more clear-cut predictions and makes less-demanding assumptions about the voters' rationality. Moreover, multiplicity of equilibria is an additional problem in the [signaling] model."[7] The model in this paper has a unique equilibrium, clear cut predictions, and (we think) undemanding assumptions about voter rationality.

Applied to signaling by incumbent politicians the model has two clear predictions. First, even if incumbents and challengers are drawn from the same symmetric distribution, the ability of incumbents to signal their types will tend to bias election outcomes, predicting an incumbency advantage in re-elections. Second, voters would be better off if they could commit to a bias against incumbent politicians. This bias could be implemented by a supermajority rule, allowing incumbents re-election only when their vote share exceeds some fraction strictly above 50%. Both predictions are also derived in an earlier paper of ours with two coauthors, Caselli et al. (2013), which worked with a type distribution with 3 symmetric mass points (i.e., a low, intermediate, and high type).[8] That paper discusses the predictions and related literature in more detail.

In industrial organization a series of papers have examined "limit pricing", in which a monopolist lowers their price to dissuade challengers from entering their market. Milgrom and Roberts (1982) consider limit pricing as signaling, and argue that although prices will be lowered, there is no reason to believe that the lower prices will lower the equilibrium rate of entry.[9] Matthews and Mirman (1983) do find equilibrium entry deterrence in a model with noise, however they are able to show this only when the noise has a strictly increasing density function, while the opposite holds when the noise density function is decreasing. That paper does not give an argument for expecting that noise would, in general, have either of these properties. In contrast our result assumes a symmetric noise distribution.

Applied to limit pricing our model implies that (1) entry will be lower than in the case without signaling (i.e., successful deterrence of entry); and (2) entrants would be better off if they could commit to an entry rule which would force them to enter at a higher price than is optimal *ex post* - this could be interpreted as incurring sunk costs before becoming informed.

Finally, the model has applications to the use of tests for sorting students. Many institutions take into account a student's background, in part to offset the variation in difficulty of achievement given their circumstances. However our model shows that such policies will

[7] In the original quotation Persson and Tabellini (2002) use "moral hazard" to refer to what we call "signal jamming", and "adverse selection" to refer to what we call "signaling".

[8] The intuition for the persuasion result is quite straight-forward because the receiver will be indifferent about admission of the intermediate type (it has the same value of the outside option). Therefore the equilibrium threshold will be exactly halfway between the messages sent by the low and high types, as it would be in the case without effort. The intermediate type will exert more effort than that exerted by the high and low types, therefore admission will be boosted above the level in the case without effort.

[9] More precisely, they show that (1) in any separating equilibrium entry will be unaffected; (2) in a pooling equilibrium entry can change, but it could be either higher or lower than in the full-information case; and (3) in any equilibrium entrants' average expectations of profit from entry are unaffected.

not fully offset the disadvantage. Consider two types of student, both apply to university by sitting a standardized exam, but they differ in their ability to manipulate their score, for example because of intensive coaching. If the institution can use a different exam-score threshold for each type then the equilibrium threshold will be higher for advantaged students (those with low cost of effort), to offset their advantage. Nevertheless those students will be more bunched around the threshold, causing a greater skew in posteriors, and so a larger fraction will be admitted. In other words, adjusting the admission threshold *will not* be sufficient to offset an advantage in preparation. A testable prediction of our model is that, if the underlying distribution of quality was the same for each group, then although the marginal admitted student will be the same, the *average* quality of the admitted students will be higher among the disadvantaged group. This is consistent with a fact about university admissions in the UK: although privately educated students are much more likely to be admitted to first-tier universities,[10] they are 5% less likely to be awarded a 1st-class degree once admitted (Machin et al., 2009).

3 Model

For the purposes of exposition we will present the model as an admission problem: the sender (he) is a candidate wishing to join some program, and the receiver (she) is an admissions officer. The sender has ability $t \in \mathbb{R}$, that is also referred to as his *type* and is his private information. The type has cumulative distribution $F(\cdot)$ and density $f(\cdot)$ which we will assume to be uniform over $[-T, T]$.[11]

The receiver faces a binary decision $a \in \{0, 1\}$, where $a = 1$ represents admitting the sender. The payoff to the receiver is given by $v^R(a, t) = at$, i.e, if the receiver admits the sender, she gets a payoff proportional to his ability, whereas if she rejects him then her payoff is normalized to 0. The sender receives a fixed benefit of 1 if admitted and 0 otherwise. Hence, independently of his type, the sender wishes to *persuade* the receiver to admit him. These assumptions imply that exactly half the types would be admitted under full information, however this symmetry in the type distribution is not important to the persuasion result, as long as the distribution of types is uniform in the neighborhood of the marginal type ($t = 0$).

In order to make her admission decision, the receiver observes a noisy signal s of the ability of the sender. You may think of s as the score of an exam the sender has to sit in order to be considered for admission. The sender can manipulate his expected score, m, which we call the *message*, at some cost, $c(m - t)$ where $c(\cdot)$ is a strictly convex function with $c'(0) = 0$. Hence, a sender that does not exert any effort would generate a score which

[10] The three year average success rate for students in the maintained sector at Cambridge is 25.7% compared to 33.6% of privately educated students. These figures at Oxford are 20.7% and 26.7% respectively. (Data taken from Cambridge and Oxford admission offices)

[11] If the distribution of types was not uniform then noise could generate persuasion without manipulation. In general, adding noise causes posteriors to shift towards the prior, i.e. the mean of the distribution, so when the mean of a distribution is not equal to the median then noise can change the fraction of posteriors which are above the mean.

in expectation equals his true ability, but by exerting some effort the candidate can boost his average score.[12]

The receiver then gets the signal $s = m + u$, where u represents noise which is distributed independently of m with mean 0 and density distribution $g(\cdot)$. We assume that $g(\cdot)$ is continuous, differentiable nearly everywhere, symmetric and single-peaked, satisfies the strict monotone likelihood property (MLRP), and has bounded support $[-D, D]$.

A pure message strategy for the sender is a function $m : [-T, T] \to \mathbb{R}$ and a pure action strategy for the receiver is a function $a : \mathbb{R} \to \{0, 1\}$. Given strategy $m(\cdot)$ and a realization of the signal s, the receiver updates her beliefs about the type of the sender. We denote ~~this updated belief by $f(t|s, m(\cdot))$ and assume it is consistent with Bayes' rule where applicable~~. When receiving a signal s which can not be generated by any realization of t and u, given $m(\cdot)$, we make no restrictions on the beliefs.

Definition 1. A pair of strategies (m, a) is a pure ~~strategy~~ equilibrium if:[13]

$$(i) \quad \text{for all } t \in [-T, T] \quad m(t) \in \arg\max_{\tilde{m} \in \mathbb{R}} \int a(\tilde{m} + u) g(u) du - c(\tilde{m} - t)$$

$$(ii) \quad a(s) = 1 \quad \text{if and only if} \quad \int t f(t|s, m(\cdot)) dt \geq 0$$

Before describing the equilibrium we make two additional assumptions that simplify the characterization and illustration of the results. First, we assume that the marginal cost is everywhere steeper than the noise distribution:

$$(A1) \quad \inf c'' > \sup g'.$$

This condition guarantees the continuity of the sender's strategy in equilibrium, and in particular guarantees the uniqueness of the *on path* equilibrium strategies. Note that Assumption (A1) rules out the case without noise.[14]

Second, we assume that the support of the noise distribution is small relative to the support of the type distribution:

$$(A2) \quad D \leq T - \int c^{-1}(g(u)) g(u) du.$$

Assumption (A2) will be sufficient to guarantee that the density of types will be flat in the neighborhood of the equilibrium threshold k^*, i.e., between $k^* - D$ and $k^* + D$. This assumption allows the integrals, which range over the realizations of noise at the threshold, to admit simple closed-form solutions. In the Appendix, we relax this assumption to allow

[12] de Haan et al. (2011) have a noisy signaling model in which the message is multiplicative in the sender's type and which features multiplicity of equilibria.

[13] The action the receiver takes whenever she is indifferent between admitting the sender or not does not affect the equilibrium in a substantial way hence for simplicity we assume that the receiver admits the sender whenever she is indifferent.

[14] If there was no noise then there will exist a continuum of equilibria, corresponding to different thresholds, where a mass of types pool at the threshold, and persuasion will hold in all but one of those equilibria. This is because if some $t > 0$ was not admitted then it must be the unique type sending that signal, and therefore has expected type above zero, contradicting the assumption.

unbounded distributions of noise and we show that the results hold as T goes to infinity.

Proposition 2 (Existence and Uniqueness)**.** *There exists an equilibrium* $(m^*(\cdot), a^*(\cdot))$ *characterized by a threshold* k^**, such that the receiver uses the threshold rule:*

$$a^*(s) = \begin{cases} 1 & if \quad s \geq k^* \\ 0 & if \quad s < k^* \end{cases} \tag{1}$$

and the message strategy is strictly increasing in t and satisfies:

$$g(k^* - m^*(t)) = c'(m^*(t) - t). \tag{2}$$

Moreover, any equilibria is identical to this one except for the receiver's response to signals which never occur in equilibrium.

Although everything is stated in terms of pure strategies, the behaviour along the equilibrium path is identical even when mixed strategies are allowed. This is proved in the Appendix together with Proposition 2

3.1 Inverting the Message Function

Equation (2) in Proposition 2 provides an implicit solution for the optimal response of a sender that faces an admission threshold k. Denote such a best response by $m(t, k)$. Given the sender's behavior we can determine many quantities of interest: the distribution of signals, the probability of admission given a threshold k, the expected type when the receiver observes a signal at the threshold k, the ex-ante utility of the sender, and the ex-ante utility of the receiver:

$$\begin{aligned}
\text{Distribution of Signals:} \quad & J(s,k) = \int_{-T}^{T} G(s - m(t,k)) f(t) dt \\
\text{Admission:} \quad & P[s > k, k] = \int_{-T}^{T} (1 - G(k - m(t,k))) f(t) dt \\
\text{Expectation:} \quad & E[t|s = k, k] = \int_{-T}^{T} t g(k - m(t,k)) f(t) dt \Big/ \int_{-T}^{T} g(k - m(t,k)) f(t) dt \\
\text{Receiver's Ex-ante Utility:} \quad & U_R(k) = \int_{-T}^{T} t(1 - G(k - m(t,k))) f(t) dt \\
\text{Sender's Ex-ante Utility:} \quad & U_S(k) = \int_{-T}^{T} (1 - G(k - m(t,k))) f(t) dt - \int_{-T}^{T} c(m(t,k) - t) f(t) dt
\end{aligned}$$

These expressions are difficult to simplify because the function $m(t, k)$ does not have an analytic expression except for very special cases.[15] However, even if $m(t, k)$ is only defined implicitly, its inverse $t(m, k)$ can be expressed explicitly:

$$t(m, k) = m - r(g(k - m)) \tag{3}$$

where $r(\cdot) \equiv (c')^{-1}(\cdot)$ is the inverse of the marginal cost function. The function $t(m, k)$

[15] Matthews & Mirman (1983) provide an example when the noise density is linearly increasing or decreasing.

represents the type that would best respond by sending message m when facing threshold k.

In particular, using the inverse message function we can compute the distribution of messages when the sender best responds to a threshold k.

Lemma 3. *For a given threshold k, with $k < T - D$, the cumulative distribution of messages $H(\cdot, k)$ and density $h(\cdot, k)$ are given by:*

$$H(m,k) = \begin{cases} \frac{1}{2T}(m - r(g(k-m)) + T) & \text{if } |m| \leq T \\ 0 & \text{if } m < -T \\ 1 & \text{if } m > T \end{cases}$$

$$h(m,k) = \begin{cases} \frac{1}{2T}(1 + r'(g(k-m))g'(k-m)) & \text{if } |m| \leq T \\ 0 & \text{if } |m| > T \end{cases}$$

Moreover the density of the signals is antisymmetric around k for all m such that $|m - k| < T - |k|$.

Proof. For $|m| \leq T$,

$$\begin{aligned} H(m,k) &= Prob(\tilde{m} \leq m) \\ &= Prob(\tilde{t} \leq t(m)) \\ &= Prob(\tilde{t} \leq m - r(g(k-m))) = F(m - r(g(k-m))) = \tfrac{1}{2T}(m - r(g(k-m)) + T) \end{aligned}$$

The density is derived by differentiating $H(m,k)$. The anti-symmetry of $h(\cdot,k)$ around k is shown below. Consider $m = k + d$ with $|d| < T - |k|$, then,

$$\begin{aligned} h(k+d,k) - \tfrac{1}{2T} &= \tfrac{1}{2T}\left(1 + r'(g(-d))g'(-d)\right) - \tfrac{1}{2T} \\ &= \tfrac{1}{2T}\left(1 - r'(g(d))g'(d)\right) - \tfrac{1}{2T} \\ &= -\left(h(k-d,k) - \tfrac{1}{2T}\right), \end{aligned}$$

where the second equality follows by the symmetry of $g(\cdot)$ and anti-symmetry of $g'(\cdot)$. □

The distribution of the messages is illustrated in Figure 1. Observe that there is bunching of messages above the threshold with the corresponding reduction in the density of the messages below the threshold. Intuitively, those candidates with natural ability (t) below but close to the threshold would exert a lot of effort to boost their score just above the threshold. This implies that whenever the receiver gets a signal equal to the threshold, the signal is relatively more likely to be generated by a message above the threshold.

Since we are able to express the density of the messages depending exclusively on the primitives of the model, the can treat the messages as an exogenous variable and the types as endogenously derived using equation (3).

Finally, the signals are related to the messages through the realization of the noise u, therefore, given a signal s, and a realization of the noise u we can compute which type

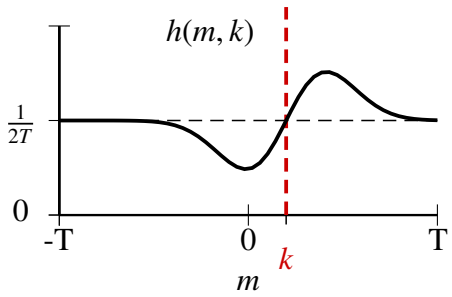


Figure 1: The density of messages, $h(m,k)$, when the senders best respond to threshold k

would have generated such a signal. This is given by:[16]

$$t^k(s,u) = s - u - r(g(k-s-u)) \tag{4}$$

In particular, for the special (but important) case in which the signal observed by the receiver coincides with the threshold:

$$t^k(k,u) = k - u - r(g(u)) \tag{5}$$

Equation 5 is particularly helpful because it allows us to express all the quantities of interest as integrals over the exogenous noise u.

Figure 2 illustrates the change in the representation of the model.

4 Admission

In what follows we assume that the receiver is using a threshold rule in which all signals above k are admitted, and that the senders are best responding to that rule. We first compute the probability of admission given a threshold k.

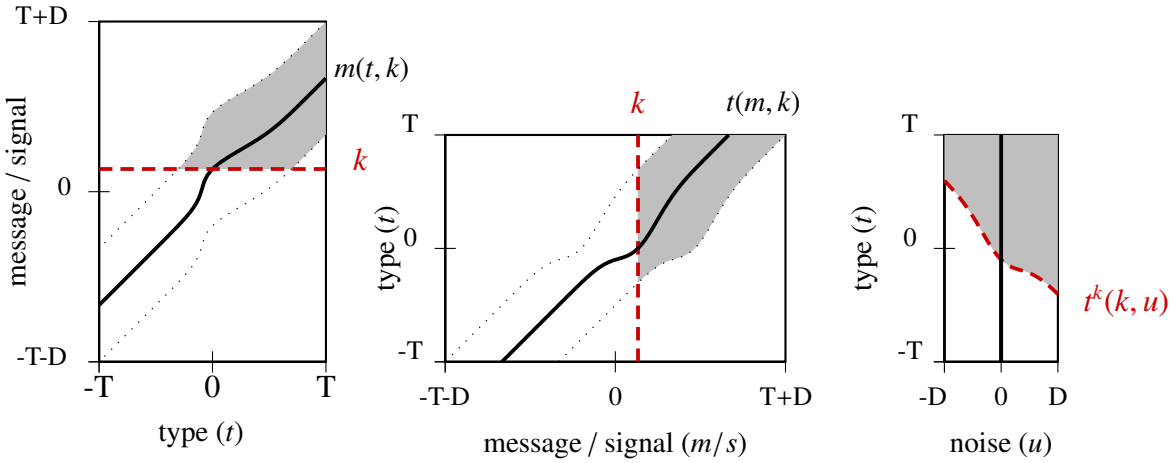
Proposition 4. *For any threshold k with $|k| \leq T - D$, the probability of admission is:*[17]

$$Admission(k) = \frac{1}{2} - \frac{k}{2T} + \frac{1}{2T}\int r(g(u))g(u)du$$

The first term, $\frac{1}{2}$, represents the fraction of types above zero which the receiver would admit under full information. The second term represents the loss from setting a higher threshold: for each unit the threshold increases above 0, a density $\frac{1}{2T}$ of types are rejected. The third term is of most interest: it represents additional admission due to agents' exertion of effort to exceed the threshold. If effort was infinitely costly, then $r(\cdot)$ would be zero and this term would disappear. Notice that this third term is independent of the threshold, k.

[16] Note that the density of signals can be derived by the density of messages and expressed as well in terms of the primitives of the model.

[17] The condition $|k| \leq T - D$ is not restrictive in equilibrium since by Assumption (A2) $D \leq \frac{T}{2}$. Therefore any signal $s > T - D$ can only be sent by senders with strictly positive types, and should always be admitted.



(a) In type-message space: The solid line represents the message (m) as a function of type (t), with dotted lines above and below indicating the bounds of the noise. The dashed line is the threshold k, and therefore the shaded area indicates the signals which are admitted. Note that the intermediate types exert the most effort (i.e,. $m - t$ is largest for intermediate types).

(b) In message-type space: This is simply the previous graph with axes flipped. With this representation, the function $t(m)$ deviates from the 45-degree line by an analytical quantity, $r(g(k-m))$.

(c) In noise-type space: This graph can be thought of as a horizontal compression of the prior graph between the two dotted lines. Each point corresponds to a signal s, and the dashed locus represents all the values of u and t which give rise to the threshold signal k. The equation of this curve is $t^k(k,u) = k - u - r(g(u))$.

Figure 2: Transformation of the representation from t-m space to u-t space.

Proof. Admission corresponds to the probability of observing a signal above the threshold. Given a realization of the noise u, this corresponds to the probability of a message sufficiently high to achieve a signal above the threshold, i.e. $m \geq k - u$:

$$\begin{aligned} Admission(k) &= 1 - \int H(k-u,k)g(u)du \\ &= 1 - \int F(k-u-r(g(u)))g(u)du \\ &= 1 - \tfrac{1}{2T}\int (k-u-r(g(u))+T)g(u)du \\ &= \tfrac{1}{2} - \tfrac{k}{2T} + \tfrac{1}{2T}\int r(g(u))g(u)du \end{aligned}$$

Where the third equality follows because $|k-u| \leq T$ for any realization of the noise, given that $|k| < T - D$. The last step follows because $\int ug(u)du = 0$ by the symmetry of $g(\cdot)$. □

An important benchmark is the threshold which admits exactly one half of all the senders, k^0. This is the fraction that would be admitted either under full information (because then only the types with $t > 0$ would be admitted), or if there was no manipulation (in which case $m = t$ and the equilibrium threshold would be $k^* = 0$). Proposition 4 allows us to characterise the threshold k^0.

Corollary 5. *The threshold k^0 at which exactly $\frac{1}{2}$ of the senders are admitted is:*

$$k^0 = \int r(g(u))g(u)du > 0. \tag{6}$$

The quantity $\int r(g(u))g(u)du$ is a measure of the precision of the noise, normalised by the cost of manipulating a message. For a quadratic cost function ($c(m-t) = \frac{c}{2}(m-t)^2$) the expression becomes simply $\frac{1}{c}\int g(u)^2 du$, proportional to the expected density of the noise, $E[g(u)]$.[18] Intuitively, a higher dispersion of noise causes a lower average effort, and therefore less admission for any given k, and so a lower k^0 is necessary to achieve admission of $\frac{1}{2}$. Likewise a higher cost c causes less manipulation and so less entry.

5 Equilibrium

We can now compute, for any threshold k, the expected type of a sender when the receiver observes a signal that is exactly equal to the threshold. Similarly to the admission function, the expected type is linear in k.

Proposition 6. *Given a threshold k, with $|k| < T - D$, the expected type of the sender when the signal observed is equal to the threshold can be expressed in either of these two forms:*

$$\begin{aligned} E[t|s=k,k] &= k - \int r(g(u))g(u)du - \int ug(u)r'(g(u))g'(u)du \\ E[t|s=k,k] &= k + \int ug'(u)r(g(u))du \end{aligned} \tag{7}$$

and the unique equilibrium threshold is given by:

$$k^* = -\int ug'(u)r(g(u))du > 0 \tag{8}$$

Proof. The expectation can be expressed as:

$$E[t|s=k] = \frac{\int t(m,k)g(k-m)h(m,k)dm}{\int g(k-m)h(m,k)dm}$$

We solve for the numerator and denominator separately. Substituting for $t(m,k)$ and $h(m,k)$ and changing the variable of integration to $u = k - m$:

$$\begin{aligned} \int t(m,k)g(k-m)h(m,k)dm &= \int (k-u-r(g(u)))\, g(u)\tfrac{1}{2T}(1+r'(g(u))g'(u))du \\ &= \tfrac{1}{2T}\left(k - \int r(g(u))g(u)du - \int ug(u)r'(g(u))g'(u)du\right) \end{aligned}$$

For the first equality we replace $h(k-u,k)$ taking into account that $|k-u| \leq T$ for all $|u| \leq D$. For the second equality we use the fact that $\int g(u)g'(u)du = 0$. The denominator represents

[18] This quantity can be interpreted as a measure of the concentration of a distribution; it is sometimes called the Rényi entropy (Rényi, 1961) and is the continuous counterpart of the Herfindahl index of a distribution. For a Gaussian distribution it is inversely proportional to the standard deviation ($\frac{1}{2\sqrt{\pi}}\frac{1}{\sigma}$).

the density of signals at k, which because of the anti-symmetry of $h(\cdot, k)$ is just $\frac{1}{2T}$.

$$\begin{aligned}\int g(k-m)h(m,k)dm &= \int g(u)\tfrac{1}{2T}(1+r'(g(u))g'(u))du \\ &= \tfrac{1}{2T}\end{aligned}$$

This yields the first expression:

$$E(t \mid s=k) = k - \int r(g(u))g(u)du - \int ug(u)r'(g(u))g'(u)du$$

And the second expression comes from integrating by parts $\int ug(u)r'(g(u))g'(u)du$:

$$\begin{aligned}\int ug(u)r'(g(u))g'(u)du &= [ug(u)r(g(u))]_{-\infty}^{+\infty} - \int r(g(u))(g(u)+ug'(u))du \\ &= -\int r(g(u))g(u)du - \int ug'(u)r(g(u))du.\end{aligned}$$

Thus:

$$E[t \mid s=k] = k + \int ug'(u)r(g(u))du.$$

Finally, by the symmetry of $g(\cdot)$, $ug'(u) < 0$ for all u and hence $k^* > 0$. □

This proposition allows us to derive the main result in the paper: allowing the senders to manipulate their messages leads to persuasion even though the receiver is fully aware of the manipulation by the senders.

Theorem 7. *At the equilibrium threshold there is persuasion, i.e. $k^* < k^0$, and the persuasion is given by:*

$$\begin{aligned}Persuasion &= Admission(k^*) - \tfrac{1}{2} \\ &= -\tfrac{1}{2T}\int ug(u)r'(g(u))g'(u)du > 0.\end{aligned} \tag{9}$$

Proof. Note that we can write $k^* = k^0 + \int ug(u)r'(u)g'(u)du$, and hence it is enough to see that $\int ug(u)r'(g(u))g'(u)du < 0$, which follows as $ug'(u) < 0, g(u) > 0, r'(g(u)) > 0$ for all u, given the symmetry of g, and the convexity of the cost function. The expression can be computed by substituting k^* in the admission function. □Equation (9) together with Assumption (A1) allow us to put an upper bound on persuasion. More precisely, Assumption (A1) implies that $r'(g(u)) < \frac{1}{|g'(u)|}$ for all u. Hence:

$$-\frac{1}{2T}\int ug(u)r'(g(u))g'(u)du < \frac{1}{2T}\int |u|g(u)du$$

Below we provide an example for the case of quadratic cost and triangular noise distribution. We also compute a tight bound on persuasion for the particular case of quadratic cost functions.

Example: Quadratic Cost-Triangular Noise. Suppose that cost is quadratic ($c(e) = \frac{c}{2}e^2$) and that the noise has a symmetric triangular density between $-D$ and D, i.e.:

$$g(u) = \begin{cases} \frac{1}{D} + \frac{u}{D^2} & \text{if } -D < u \leq 0 \\ \frac{1}{D} - \frac{u}{D^2} & \text{if } 0 < u \leq D \\ 0 & \text{otherwise} \end{cases}$$

Therefore, $\int r(g(u))g(u)du = \frac{1}{c}\int g^2(u)du = \frac{2}{3cD}$. The probability of admission given a threshold k is:

$$Admission(k) = \frac{1}{2} - \frac{k}{2T} + \frac{1}{3cDT}$$

and the expectation of the sender's type given the receiver observes a signal at the threshold becomes:

$$E[t|s = k, k] = k - \frac{1}{3cD}$$

These two functions imply that the threshold at which there is no persuasion is $k^0 = \frac{2}{3cD}$, and the equilibrium threshold is $k^* = \frac{1}{3cD}$. This leads to a persuasion in equilibrium of $\frac{1}{6cDT}$. For example, for $T = \frac{1}{2}, D = \frac{1}{4}, c = 16$, which satisfy Assumptions (A1) and (A2), we have that $k^0 = \frac{1}{6}, k^* = \frac{1}{12}$ and the persuasion in equilibrium is of $\frac{1}{12} = 8.3\%$

In fact an observation helps us to bound persuasion in the case of quadratic cost functions. Given quadratic cost, the persuasion at equilibrium can be written as

$$Persuasion = \frac{1}{4cT}\int g^2(u)du$$

Moreover, given Assumption (A1) which bounds the slope of the density of the noise and the assumption that the density is symmetric and single-peaked, the triangular distribution is the density that maximises the persuasion. Hence a bound for the persuasion is found by solving:

$$\begin{aligned} \max_{c,D,T} \quad & \frac{1}{6cDT} \\ st. \quad & c \geq \frac{1}{D^2} \qquad (A1) \\ & T \geq D + \frac{2}{3cD} \quad (A2) \end{aligned}$$

which gives a bound of exactly 10%[19]. This bound is achieved when for example $T = \frac{5}{3}$ and $D = c = 1$.

6 Comparing Different Regimes

In this section we derive solutions for the expected utility of the sender and receiver. We then compare, in terms of ex-ante utilities, the model to two alternative specifications. First, we allow the receiver to commit to a threshold *ex-ante* that may not be optimal *ex post*. We find that the optimal commitment threshold k^C is above k^*, meaning that a higher threshold improves the separation of types. From the sender's point of view the higher threshold

[19] In order to solve for this problem, it is enough to realize that the two inequalities have to be binding.

clearly makes them worse off. Surprisingly we find that k^C is exactly equal to k^0, implying that there is zero net persuasion when the receiver can commit.

Second, we study the case in which the senders cannot manipulate the signal, for example if candidates were given a surprise exam for which they could not prepare, or if a monopolist was not aware of potential challengers and therefore was setting monopolistic prices. We find that receivers strictly prefer this regime, even to the case in which they can commit to a threshold. For senders the outcome is ambiguous unless the cost of manipulation is a power function, in which case the costs and benefits of manipulation exactly cancel out. This equality highlights the quantitative importance of persuasion: it implies that ~~that~~ wherever we see significant effort being expended on signaling (e.g., money on reelection campaigns, time studying, significant limit pricing), there exists an equally large net persuasion that compensates those efforts.

We first state the expected utility for the sender and receiver:

Proposition 8. *For any threshold $|k| < T - D$ the receiver's expected utility is:*

$$U_R(k) = \frac{T^2}{2} - \frac{1}{2}\int (k - r(g(u)))^2 g(u)du + \frac{1}{2}\int u^2 g(u)du,$$

and the sender's expected utility is:

$$U_S(k) = \frac{1}{2T}\int (r(g(u))g(u)du - k) - \frac{1}{2T}\int c(r(g(u)))du.$$

Proof. Given a threshold k and a realization of the noise u, the types sending signals above the threshold k are $t \in [t(k-u,k), T]$. Therefore the expected utility of the receiver given k and u is:

$$\int_{t(k-u,k)}^{T} tdt = \frac{1}{2}(T^2 - t(k-u,k)^2),$$

and the *ex ante* expected benefit of the receiver will be:

$$\begin{aligned} U_R(k) &= \tfrac{T^2}{2} - \tfrac{1}{2}\int t(k-u,k)^2 g(u)du \\ &= \tfrac{T^2}{2} - \tfrac{1}{2}\left[\int (k - r(g(u)))^2 g(u)du + \int u^2 g(u)du\right] \end{aligned}$$

Turning to the sender's expected utility, the first term is the expected benefit, which is simply $Admission(k)$. The second term is the cost, which can be expressed as:

$$\begin{aligned} \text{Cost} &= \int c(e(t))f(t)dt = \tfrac{1}{2T}\int c(m(t) - t)dt \\ &= \tfrac{1}{2T}\int c(r(g(k-m)))(1 + r'(g(k-m))g'(k-m))dm \\ &= \tfrac{1}{2T}\int c(r(g(u)))(1 + r'(g(u))g'(u)du \\ &= \tfrac{1}{2T}\int c(r(g(u)))du \end{aligned}$$

□

Corollary 9. *The threshold chosen under commitment (k^C) achieves admission of exactly $\frac{1}{2}$, and is greater than the no-commitment equilibrium threshold (k^*), which in turn is greater*

than the no-manipulation threshold (k^{NM}), i.e.:

$$k^0 = k^C \geq k^* \geq k^{NM} = 0$$

Proof. For the receiver, the marginal benefit of raising the threshold is:

$$\frac{\partial W}{\partial k} = -\int (k - r(g(u)))g(u)du$$

and hence the optimal commitment threshold is $k^C = \int r(g(u))g(u)du$. This is equal to k^0, the threshold which admits half of the signals, and therefore admission will be equal to admission in the case without effort. □

Corollary 10. *The receiver is better off with no manipulation ($U_R^{NM}(k^{NM})$) than with the commitment threshold, which is better in turn than the no-commitment threshold, i.e.:*

$$U_R^{NM}(k^{NM}) \geq U_R^M(k^C) \geq U_R^M(k^*).$$

Proof. When no manipulation is possible ($c'(\cdot) = \infty$, $r(\cdot) = 0$), then $k = r(g(u)) = 0$ for all u, and so the intermediate term will disappear from U_R, therefore delivering higher expected utility than the equilibrium with manipulation, with or without commitment. □

Corollary 11. *The sender is worse off under commitment, i.e.*

$$U_S(k^*) < U_S(k^C),$$

In addition, if the sender's cost is a power function then they will be indifferent between having no manipulation and facing the ex post *threshold k^*, i.e.*

$$U_S(k^*) = U_S^{NM}(0).$$

Proof. If cost is a power function, with $c(e) = \frac{1}{1+\alpha}ce^{1+\alpha}$, this implies the following:

$$c'(e) = ce^{\alpha}, \quad r(g) = \left(\frac{g}{c}\right)^{\frac{1}{\alpha}}, \quad r'(g) = \frac{1}{\alpha}c^{\frac{-1}{\alpha}}g^{\frac{1-\alpha}{\alpha}}, \quad c(r(g)) = \frac{1}{1+\alpha}c^{-\frac{1}{\alpha}}g^{\frac{1+\alpha}{\alpha}}$$

Substituting these expression in (for the sake of presentation we have omitted the arguments of the functions):

$$\begin{aligned}
\int ur'g'g + \int c &= \int u\frac{1}{\alpha}c^{\frac{-1}{\alpha}}g^{\frac{1-\alpha}{\alpha}}g'g + \int \frac{1}{1+\alpha}c^{-\frac{1}{\alpha}}g^{\frac{1+\alpha}{\alpha}} \\
&= \frac{1}{\alpha}c^{\frac{-1}{\alpha}}\int ug^{\frac{1}{\alpha}}g' + \frac{1}{1+\alpha}c^{-\frac{1}{\alpha}}\int g^{\frac{1+\alpha}{\alpha}} \\
&= \frac{1}{\alpha}c^{\frac{-1}{\alpha}}\left[u\frac{\alpha}{1+\alpha}g^{\frac{1+\alpha}{\alpha}}\right]_{-\infty}^{\infty} - \frac{1}{\alpha}c^{\frac{-1}{\alpha}}\frac{\alpha}{1+\alpha}\int g^{\frac{1+\alpha}{\alpha}} + \frac{1}{1+\alpha}c^{-\frac{1}{\alpha}}\int g^{\frac{1+\alpha}{\alpha}} \\
&= 0
\end{aligned}$$

Where the second-last step uses integration by parts on the first term. □

7 Conclusion

This paper has demonstrated that a precise characterization can be given to the equilibrium of a noisy signaling model with additive effort and a binary decision by the receiver. What's more, a variety of non-trivial predictions about equilibrium behavior arise: the anti-symmetric distribution of signals, the existence of persuasion, and the desire for a higher threshold under commitment by receivers, among others.

There are a number of interesting extensions to this work, including (1) asking whether the results hold for more general cost functions $c(m,t)$; (2) applying the model to multidimensional signaling, a.k.a. multitasking.

Appendix

Proof of Proposition 2

We prove a more general result than that given in the body of the paper, and then show that proposition 2 follows from this.

A mixed *message strategy* for the sender is denoted by $\sigma : [-T, T] \to \Delta(\mathbb{R})$, where $\sigma(t)$ is a probability distribution over signals, and a mixed *action strategy* for the receiver is denoted by $p : \mathbb{R} \to [0, 1]$, where $p(s)$ represents the probability of choosing action $a = 1$ given signal s. We assume that both σ and p are measurable functions.

Definition 12. A pair of strategies (σ, p) is an equilibrium if:

(i) for all $\tilde{m}$ in the support of $\sigma(t)$ $\quad \tilde{m} \in \arg\max_{m\in\mathbb{R}} \int p(m+u)g(u)du - c(m-t)$

(ii) $\quad p(s) \in \arg\max_{p\in[0,1]} p \int tf(t|s)dt$

We say that the sender's mixed strategy σ is non-decreasing (increasing) in his type if for any $t' > t$, if $m' \in \text{Support}(\sigma(t'))$ and $m \in \text{Support}(\sigma(t))$, we have $m' \geq (>)m$.
Finally, given a strategy σ, we denote by S^σ the set of signals that might arise given strategy σ:

$S^\sigma = \{s \in \mathbb{R} \mid s = m + u,\ m \in \text{Support}(\sigma(t)) \text{ for some } t \in [-T, T], u \in [-D, D]\}$.

Lemma 13. *In any equilibrium the sender's strategy is increasing in his type and the receiver uses a threshold rule on the equilibrium path, i.e. for some $k \in \mathbb{R}$, and $s \in S^\sigma$*

$$p(s) = \begin{cases} 1, & s > k \\ 0, & s < k \end{cases}$$

Proof. Given an equilibrium (σ, p), all senders face a common benefit from sending a message m:

$$\pi(m) = \int p(m+u)g(u)du$$

We first show that σ will be non decreasing in the sender's type. Consider $t' > t$ and denote $m' \in \text{Support}(\sigma(t'))$ and $m \in \text{Support}(\sigma(t))$, optimality implies that:

$$\begin{aligned} \pi(m) - c(m-t) &\geq \pi(m') - c(m'-t) \\ \pi(m') - c(m'-t') &\geq \pi(m) - c(m-t'), \end{aligned}$$

implying:

$$c(m'-t) - c(m-t) \geq \pi(m') - \pi(m) \geq c(m'-t') - c(m-t').$$

Because c is strictly convex this implies that $m' \geq m$.

Second, the marginal benefit function must be bounded because by substitution:

$$\begin{aligned}\pi(m) &= \int_{-D+m}^{D+m} p(u)g(u-m)du \\ \pi'(m) &= -\int_{-D+m}^{D+m} p(u)g'(u-m)du,\end{aligned}$$

where we are using the fact that g is continuous, so $g(D) = g(-D) = 0$. The last expression is bounded, because $p(u)$ is bounded between 0 and 1, and $g(0)$ is bounded. This implies that the support of $\sigma(t)$ and the support of $\sigma(t')$ cannot overlap, because every m elicits a finite marginal benefit, and the marginal cost strictly falls in t, thus no value of m can be optimal for two different values of t.

Finally, the strict monotonicity of messages sent by different types of senders implies that $E[t|s]$ must be strictly increasing on S^σ, due to g satisfying the MLRP. Therefore the restriction of p to S^σ must be a threshold rule: for all $s \in S^\sigma$, $p(s) = 1$ if $s > k$ and $p(s) = 0$ if $s < k$. Note that the threshold must be in the interior of S^σ, by the law of iterated expectations (i.e., there must exist signals with expected types both above and below zero). □

Lemma 14. *If (σ, p) is an equilibrium, then $(\sigma, \bar{p})$ is also an equilibrium, where $\bar{p}(s) = p(s)$ for all $s \in S^\sigma$ and $\bar{p}$ is a threshold rule over the whole $\mathbb{R}$.*

Proof. First note that by Assumption (A1), the message satisfies $|m(t) - t| < D$, hence S^σ must be convex. Denote by $\underline{m} = \inf\{\text{Support}(\sigma(-T))\}$ and $\overline{m} = \sup\{\text{Support}(\sigma(T))\}$, i.e. $\underline{m}$ and $\overline{m}$ are the extreme messages sent in equilibrium. Finally, let $\pi(m)$ and $\bar{\pi}(m)$ be the associated expected benefit when message m is sent given strategies p and $\bar{p}$ respectively. By construction π and $\bar{\pi}$ coincide on $[\underline{m}, \overline{m}]$. Consider an out of equilibrium message m, such that $m \notin [\underline{m}, \overline{m}]$. If $m < \underline{m}$, $\bar{\pi}(m) \leq \pi(m)$ since $\bar{p}(s) \leq p(s)$ for all $s \leq \inf S^\sigma$. Hence such a message m cannot become a profitable deviation.
Consider now $m > \overline{m}$. In this case $\bar{\pi}(m) \geq \pi(m)$ since $\bar{p}(s) \geq p(s)$ for all $s \geq \sup S^\sigma$. However, the marginal benefit under $\bar{\pi}$ is $\bar{\pi}'(m) = g(k-m)$. Hence by Assumption (A1), the slope of the marginal benefit is always smaller than the slope of the marginal cost, so there cannot be additional crossing of marginal cost and marginal benefit, and the optimality of the equilibrium message remains. □

Proof. (Extended Proposition 2)
By Lemma 13 and Lemma 14 we can restrict the receiver's strategy to a threshold rule over the whole real line i.e., $p(s) = 1$ for any $s > k$ and $p(s) = 0$ for any $s < k$. The sender's utility function becomes:

$$\pi(m) - c(m-t) = \int_{k-m}^{\infty} g(u)du - c(m-t).$$

The sender's first-order and second-order conditions will therefore be:

$$g(k - m) = c'(m - t), \tag{10}$$

$$-g'(k - m) - c''(m - t) < 0, \tag{11}$$

The second-order condition holds because of Assumption (A1), implying that $m(t)$ will be a continuous increasing function. The behaviour in Equation (10) determines uniquely the equilibrium threshold k, where

$$E(t|s = k, k) = 0. \tag{12}$$

Equations (10) and (12), determine uniquely the behaviour of both players in equilibrium. By Lemma 14, any equilibrium strategies coincide with these ones on the equilibrium path.

□

References

Alonso, R. and Câmara, O. (2014). Persuading voters.

Caselli, F., Cunningham, T., Morelli, M. and Moreno de Barreda, I. (2013). The incumbency effects of signalling, *Economica* .

de Haan, T., Offerman, T. and Sloof, R. (2011). Noisy signaling: theory and experiment, *Games and Economic Behavior* **73**(2): 402–428.

Gentzkow, M. and Kamenica, E. (2014). Costly persuasion, *The American Economic Review* **104**(5): 457–462.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective, *The Review of Economic Studies* **66**(1): 169–182.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion, *American Economic Review* **101**(6).

Kamenica, E. and Gentzkow, M. (2012). Competition in persuasion.

Machin, S., Murphy, R. and Soobedar, Z. (2009). Differences in labour market gains from higher education participation, *Research commissions by the National Equality Panel* .

Matthews, S. A. and Mirman, L. J. (1983). Equilibrium Limit Pricing: The Effects of Private Information and Stochastic Demand, *Econometrica* **51 (4)**: 981–996.

Milgrom, P. and Roberts, J. (1982). Limit pricing and entry under incomplete information: An equilibrium analysis, *Econometrica: Journal of the Econometric Society* pp. 443–459.

Persson, T. and Tabellini, G. E. (2002). *Political economics: explaining economic policy*, MIT press.

Rényi, A. (1961). On Measures of Entropy and Information, *Proc. 4th Berkeley Symp. Math. Statist. and Prob.*, Vol. 1, Univ. Calif. Press, Berkeley, pp. 547–561.

Rogoff, K. and Sibert, A. (1988). Elections and macroeconomic policy cycles, *The Review of Economic Studies* **55**(1): 1–16.