TOO COOL FOR SCHOOL? SIGNALING AND COUNTERSIGNALING

NICK FELTOVICH, RICK HARBAUGH, AND TED TO

ABSTRACT. In signaling environments ranging from consumption to education, high quality senders often shun the standard signals that should separate them from lower quality senders. We find that allowing for additional, noisy information on sender quality permits equilibria where medium types signal to separate themselves from low types, but high types then choose to not signal or *countersignal*. High types not only save costs by relying on the additional information to stochastically separate them from low types, but countersignaling itself is a signal of confidence which separates high types from medium types. Experimental results confirm that subjects can learn to countersignal.

Journal of Economic Literature Classification Categories: C72, D82, D83.

Date: March 2001.

We thank Chris Avery, Bruno Broseta, William Hamilton, Robin Hanson, John Hardman-Moore, John Kagel, Barry Nalebuff, Al Roth, Karl Schlag, seminar participants at various conferences and departments, two anonymous referees and the Editor Lars Stole. Harbaugh thanks the Yale School of Management for post-doctoral support.

> "For Nash to deviate from convention is not as shocking as you might think. They were all prima donnas. If a mathematician was mediocre he had to toe the line and be conventional. If he was good, anything went."
>
> – Z. Levinson from *A Beautiful Mind*
> (Nasar, 1998, p. 144)

1. Introduction

Following in the tradition of Veblen's (1899) analysis of conspicuous consumption and Akerlof's (1970) model of adverse selection, Spence's (1973a; 1974) signaling model of overeducation showed how seemingly wasteful actions can be valued as evidence of unobservable quality. Signaling models have since been applied to economic phenomena from advertising (Nelson, 1974) to financial structure (Ross, 1977), social phenomena from courtship (Spence, 1973b) to gift exchange (Camerer, 1988), and biological phenomena from a peacock's plumage (Zahavi, 1975) to a tree's autumn foliage (Brown and Hamilton, 1996). These models conclude that in a separating equilibrium "high" types (high in productivity, wealth, fecundity, or some other valued attribute) send a costly signal to differentiate themselves from lower types.

Contrary to this standard implication, high types sometimes avoid the signals that should separate them from lower types, while intermediate types often appear the most anxious to send the "right" signals. The nouveau riche flaunt their wealth, but the old rich scorn such gauche displays. Minor officials prove their status with petty displays of authority, while the truly powerful show their strength through gestures of magnanimity. People of average education show off the studied regularity of their script, but the well–educated often scribble illegibly. Mediocre students answer a teacher's easy questions, but the best students are embarrassed to prove their knowledge of trivial points. Acquaintances show their good intentions by politely ignoring one's flaws, while close friends show intimacy by teasingly highlighting them. People of moderate ability seek formal credentials to impress employers and society, but the talented often downplay their credentials even if they have bothered to obtain them. A person of average reputation defensively refutes accusations against his character, while a highly-respected person finds it demeaning to dignify accusations with a response.

How can high types be so understated in their signals without diminishing their perceived quality? Most signaling models assume that the only information available on types is the signal, implying that high types will be confused with lower types if they do not signal. But in many cases other information is also available. For instance, wealth is inferred not just from conspicuous consumption, but also from information about occupation and family background. This extra information is likely to be noisy in that the sender cannot be sure what the receiver has learned, implying that types of

medium quality may still feel compelled to signal so as to separate themselves from low types. But even noisy information will often be sufficient to adequately separate high types from low types, leaving high types more concerned with separating themselves from medium types. Since medium types are signaling to differentiate themselves from low types, high types may choose to not signal, or "countersignal," to differentiate themselves from medium types.

We investigate such countersignaling behavior formally with a model that incorporates extra, noisy information on type into a signaling game. We find that countersignaling can emerge as part of a standard sequential equilibrium in which all players are forming rational beliefs and are acting rationally given these beliefs. Countersignaling is naturally interpreted as a sign of confidence.[1] While signaling proves the sender is not a low type, it can also reveal the sender's insecurity. Since medium types have good reason to fear that the extra information on type will not differentiate them from low types, they must signal to clearly separate themselves. In contrast, high types can demonstrate by countersignaling that they are confident of not being confused with low types.

The extra information on type in our model can be seen as a second signal following the literature on multidimensional signals (Quinzii and Rochet, 1985; Engers, 1987). This literature is primarily concerned with whether such signals can ensure complete separation when sender type varies in multiple dimensions. We assume that sender type varies in only one dimension and concentrate instead on the opposite problem of how the extra information can encourage partial pooling rather than complete separation.[2] Given the noisy nature of the extra information, it might seem that high types should signal to further emphasize their quality. Instead, we find that the information asymmetry arising from the noisy extra information can give perverse incentives. Pooling with low types can become a signal in itself—a way for high types to show their confidence that the extra information is favorable to them by taking an action that is too risky for medium types.

Because countersignaling serves as a signal of confidence, we show that it is more than just the absence of signaling by types whose high quality is already evident and who wish to save costs. First, when the extra information sufficiently differentiates high types from low types, signaling can actually lower a high sender's payoff. Countersignaling can therefore arise even when signaling is a desirable activity that high types would pursue in a perfect information environment. Second, countersignaling reduces the efficiency of receiver estimates of sender quality. Since countersignaling depends on the existence of additional information on sender quality, eliminating this information can actually increase estimate efficiency. Third, when there

[1]This point was made by Confucius (13:26, *The Analects*): "The Superior Man is self-confident without being arrogant. The inferior man is arrogant and lacks self-confidence."

[2]Hertzendorf (1993) also allows for two signals, one of which is noisy, but considers only two types of senders, precluding the possibility of countersignaling.

is a range of possible signals, high types not only choose a cheaper signal than medium types, but choose the same cheap signal being sent by low types. Only by pooling with low types can high types successfully discourage medium types from mimicking their behavior. Fourth, low signaling costs can paradoxically reduce signaling by encouraging high types to countersignal. In an educational context, an increase in the difficulty of an assignment can therefore "challenge" high-ability students to stop countersignaling and to send the signal of completing it. Finally, standard refinements do not predict a unique equilibrium but allow for multiple equilibria, including mixed strategy equilibria where some high types signal while others countersignal and "counter-countersignaling" equilibria where very high types differentiate themselves from countersignaling high types by signaling.

The idea that signaling-like behavior need not be monotonically increasing in quality has appeared in several areas. Teoh and Hwang (1991) develop a model where firms decide whether to immediately disclose favorable earnings information or wait for the information to be revealed by other sources. Waiting makes higher quality firms look bad at first but eventually separates them from lower quality firms which face more immediate pressure to prove themselves. Bhattacharyya (1998) considers how large a dividend firms should declare and finds that a screening model predicts that, conditioned on earnings, higher quality managers will declare lower dividends since they can use funds more efficiently than lower quality managers. Fremling and Posner (1999) discuss how those with already high status may have lower marginal returns from signaling than those who are not so well-regarded.[3] Hvide (1999) examines a labor market model in which education serves partly to inform workers of their true abilities and finds that only average types will choose to become educated. We differ from these analyses in following a standard signaling model exactly with the sole exception of allowing for the presence of additional information on sender type. This added realism is sufficient to significantly expand the set of equilibria from a standard signaling game, allowing for non-monotonic equilibria which are robust to standard refinements.

Countersignaling theory takes the intuition of signaling and shows how it can lead to quite different behavior than normally supposed, offering insight into phenomena which appear inconsistent with the standard signaling model. Of course, countersignaling is somewhat complex and there remains the issue of whether economic agents are capable of such behavior. To help answer this question we report results of an experimental test conducted in the fall of 1995. The experiments involved two games with three types of senders, high, medium and low quality, and a binary signal. The first game is isomorphic to a standard signaling game and has a unique equilibrium in which high and medium types signal. The second game is identical to the

[3]Our model supports such an argument in that high types benefit less from signaling because they are already partially separated from low types.

first game, except extra, noisy information on types leads to the unique equilibrium involving countersignaling by high types. Experimental results tend to support the theory's predictions. From almost identical initial play in the two games, subject behavior diverged to a large amount of countersignaling by high types in the latter game and almost none in the former game.

2. A Simple Example

Continuing the signaling literature's traditional emphasis on education, consider the following stylized example. A prospective employee who had good grades in high school is considering whether to mention her grades in a job interview. Because grading standards are weak both medium and high productivity employees (Highs and Mediums) are known to have good grades while only low productivity employees (Lows) are known to have poor grades. Since lying about grades involves the chance of getting caught, the signal of mentioning good grades is costly to Lows but free to Mediums and Highs. In addition to this signal, the interviewer will receive from a former boss a recommendation regarding the prospective employee's abilities. Lows expect to receive bad recommendations from their old boss and Highs expect to receive good recommendations, while Mediums receive good or bad recommendations with equal probability.

What should an interviewee do? Without the recommendation, Mediums and Highs should clearly mention their good grades since it costs them nothing and since the grades differentiate them from Lows. With the addition of the extra information as embodied by the recommendation, the situation is less obvious. Consider if the interviewer believes that only Mediums mention their grades. Then if Mediums don't mention their grades they take the chance of either receiving a good recommendation and being thought of as a High or receiving a bad recommendation and being thought of as a Low. If Lows are sufficiently unproductive relative to Mediums and Highs, not mentioning grades is too risky. Highs face a different situation because they expect to receive a good recommendation. Since they need not worry about being perceived as a Low, they face a clear choice between being perceived as a Medium if they mention their grades and a High if they do not. Since receiver beliefs are consistent with sender strategies and sender strategies make sense given receiver beliefs, a countersignaling equilibrium exists in which Highs show off their confidence by not mentioning their grades.[4]

A numerical example may help illuminate this case. Assume that productivity is 400, 700, and 900 for Lows, Mediums, and Highs respectively, and that Lows and Highs are equally prevalent in the population. Given the interviewer's beliefs, Mediums can choose to receive either 700 by mentioning

[4]A partial pooling equilibrium is still possible in which both Mediums and Highs are believed to signal, but if Lows are sufficiently unproductive relative to Mediums and Highs and if the recommendations completely separate Lows and Highs, the equilibrium does not survive the intuitive criterion (Cho and Kreps, 1987). Refinements are discussed further in Section 3.

their grades or $(400 + 900)/2 = 650$ by deviating from the equilibrium and mimicking the Lows and Highs. Meanwhile, Highs are perfectly separated from Lows so they receive 900 by countersignaling versus only 700 by deviating and looking like a Medium. Finally, as long as lying about grades costs Lows at least 300, Lows do not gain from mimicking the Mediums so a countersignaling equilibrium exists.

For simplicity this example assumes the signal of "bragging" about one's grades is free for both Mediums and Highs, but the results do not depend on this assumption. Hence this model would still apply if we were looking not just at the decision to report grades, but at the potentially costly choice of whether to get good grades in the first place. Countersignaling would still be an equilibrium even if signaling cost Mediums as much as 50 and cost Highs as little as *negative 200*, meaning that Highs would prefer to get good grades in a full information environment. Note that countersignaling can break down not just if signaling is too attractive for Highs, but also if signaling is too expensive for Mediums, *e.g.*, the grading standard makes it difficult for Mediums to get good grades. When signaling by Mediums becomes too expensive and they stop signaling in equilibrium, Highs can no longer separate themselves from Mediums by not signaling and must instead signal in order to differentiate themselves. Therefore an increase in signaling costs can actually induce Highs to start signaling.

Regarding the extra information embodied by the former boss's recommendation, the extremely dichotomous information structure simplifies the problem, but noisier information can still support a countersignaling equilibrium. In this example even if Lows receive a good recommendation 25% of the time and Highs receive a bad recommendation 25% of the time, a countersignaling equilibrium still exists. For an interviewee who doesn't mention grades, if a bad recommendation is observed the expected quality of the interviewee is $(3/4)400 + (1/4)900 = 525$, while if a good recommendation is observed the expected quality is $(1/4)400 + (3/4)900 = 775$. Since Mediums expect good and bad recommendations with equal probability, they still expect to receive 650 if they countersignal versus 700 if they signal. Lows expect to receive a bad recommendation 3/4 of the time and to receive a good recommendation 1/4 of the time, so they expect to receive $(3/4)525 + (1/4)775 = 587.5$ by not signaling, giving them even less incentive to deviate than in the previous case. For Highs their quality will be estimated at $(1/4)525 + (3/4)775 = 712.5$ if they countersignal so deviating is unprofitable and the countersignaling equilibrium still stands. Depending on the exact model parameters, even a little bit of extra information can disrupt the standard result that signals are non-decreasing in type.

In this example interviewees faced a simple binary choice of mentioning their grades or not. While signaling decisions are often binary, in many cases a wider range of signals is available, *e.g.*, how expensive a car one buys. In such cases it is less obvious that Highs will be willing to pool with Lows since they have the extra option of breaking off and sending a higher signal that

is not worthwhile for Mediums to mimic. The following section develops the theory for this case, showing that Highs can still choose to countersignal by pooling with Lows. In the final section we return to the simplest case of a binary signal with three types to report on an experimental test of countersignaling.

3. A Theory Of Countersignaling

In this sender–receiver game, we allow for three sources of receiver information. First, there is common knowledge about the distribution of types which incorporates all the background information which both senders and receivers know. For instance, if it is common knowledge that all senders are in a certain age group, then the distribution of types is conditioned on this knowledge. Continuing to assume that there are "Low", "Medium" and "High" types, let the set of types be $Q = \{L, M, H\} \subset \mathbb{R}_+$ where $0 \leq L < M < H$ and where types are distributed according to the probability distribution $f(q)$.

Second, the sender sends a signal s in the set $S \subset \mathbb{R}_+$ which is observed by the receiver. The receiver observes this signal noiselessly, but does not know which type sent the signal. This signal costs the sender $c(s, q)$ where c is increasing and convex in s and decreasing in q. To ensure there is some signal that everyone would be willing to send, we assume $0 \in S$ and c satisfies $c(0, q) = 0$. Further, we assume the standard "single-crossing property" that $c_s(s, q) > c_s(s, q')$ for $q < q'$, *i.e.*, not only is it less costly for higher types to send any given signal than it is for lower types but the marginal cost of that signal is also less. The assumption that the marginals cost of signaling is always positive will be relaxed in Section 3.2.

Finally, and this is the unique aspect of the model, the receiver has extra, noisy information about the sender's type. This information is sent at no cost to the sender and is exogenous in the sense that sender actions cannot at this stage affect it. The sender knows that the receiver has this information, but is unaware of exactly what the receiver knows. We model this information as a noisy exogenous signal, $x \in X$, distributed according to the conditional probability distribution $g(x|q)$. Assume that g has full support over X for any q.[5] The conditional distribution of the exogenous signal is common knowledge but the actual value of x is not known to the sender at the time of sending the *endogenous* signal. In general, the exogenous signal can be thought of as a summary measure of all the other noisy information that the receiver will have about the sender at the time of making the signaling choice. To reduce confusion with the signal, s, we will refer to the noisy exogenous signal, x, as just "extra information."

[5]The assumption of full support simplifies the discussion of out-of-equilibrium beliefs. When the support of g is less than full, extreme information structures can yield unique Intuitive Criterion countersignaling equilibria as in the simple example given in Section 2 and used in the experimental test in Section 4.

The structure of the game is as follows. First, a sender is drawn randomly from the distribution of types. The sender then sends the endogenous signal without knowing what was or will be the realized value of the extra information. Finally, the receiver observes both the extra information and the sender's signal. Given this information and her beliefs about sender signaling strategies, the receiver rewards the sender with the sender's expected quality.[6] This can be thought of as a reduced form of a game where senders are workers and receivers are firms which simultaneously make wage offers.

Regarding the timing of signals, *if all of the available extra information* is embodied in x and is known to both the sender and the receiver prior to sending the endogenous signal then the model reduces to the standard signaling framework where the distribution of types is given by $\hat{f}(q) = g(x|q)f(q)/\sum_{q' \in Q} g(x|q')f(q')$. Our assumption that the sender chooses s without knowing the realized value of x is therefore necessary for a countersignaling equilibrium. We believe this assumption to be innocuous in the sense that regardless of what is known prior to the choice of s, there is always some information that is unknown to the sender. For example, suppose that extra information, x, is observed by the sender and receiver prior to the sender's choice over s, but extra information y is unobserved by the sender. If type is correlated with both x and y, then y plays the exact same role as x in our model.

Except for a brief discussion of mixed strategy equilibria in Section 3.2, we consider only pure strategy Nash equilibria, so a strategy is a mapping between types and signals. Let s_q represent the pure strategy of a sender of type q and let the function $\mu(q|s,x)$ be a probability distribution representing receiver beliefs about which types q send observed signal s and information x. Receiver expectations of sender quality, given receiver beliefs and the observed signals are

$$\sum_{q' \in Q} q' \mu(q'|s,x).$$

Assuming sender risk-neutrality for simplicity, the gross of costs return to type q of sending signal s is the sender's expected perceived quality

$$E_\mu[q'|s,q] = \int_{x \in X} \left(\sum_{q' \in Q} q' \mu(q'|s,x) \right) g(x|q) dx. \tag{1}$$

Definition 1. A pure strategy *Perfect Bayesian Equilibrium* is given by a type contingent strategy profile s_q and receiver beliefs $\mu(q|s,x)$ where

(i) $E_\mu[q'|s_q,q] - c(s_q,q) \geq E_\mu[q'|s',q] - c(s',q)$ for any $s' \in S$ and

[6]The signals are thereby assumed to play a purely informational role, having no effect on the sender's productivity or other valued attributes.

(ii) for any $s \in S$, $\mu(q|s,x)$ is such that if $\{q' \mid s_{q'} = s\} \neq \emptyset$ then

$$\mu(q|s,x) = \frac{g(x|q)f(q)}{\sum_{\{q'|s_{q'}=s\}} g(x|q')f(q')}. \tag{2}$$

Condition (i) requires that agents choose signals as a best response to the receiver's beliefs. Condition (ii) requires that for any information set that can be reached on the equilibrium path, the receiver's beliefs are consistent with Bayes rule and the equilibrium sender strategy.[7]

We follow the convention of calling a perfect Bayesian equilibrium a *signaling equilibrium* if s_q is strictly increasing in the sender's type. Any equilibrium in which s_q is strictly non-monotonic will be called a *countersignaling equilibrium*. Note that in the initial motivating example (Section 2) and the later experimental test (Section 4) we use a binary signal so that the alternative to countersignaling is a *weak signaling equilibrium* in which s_q is weakly increasing in the sender's type and strictly increasing at only one point. Weak signaling equilibria can also survive in the richer signaling space we use in this theory section, but for a clear comparison with the signaling literature we restrict our attention to signaling equilibria and countersignaling equilibria.

As mentioned earlier, we require that the extra information, x, should be in some sense informative. First note that in order for x to have any information content in equilibrium, at least two types must send the same signal. Otherwise, with perfect separation, the extra information plays no role. A sender must believe that if she pools with senders of lower type she will be rewarded more, on average, than them. That is, the sender may do worse than lower types *ex post* once the receiver has observed the available information, but the information is correct on average so that *ex ante* the sender does better in expectation.

To define this notion more precisely, we need to first provide some additional notation. Since, we are only interested in pure strategy equilibria, this assumption will be defined in terms of sets of agents, $\Lambda \subset Q$, who pool together.[8] For any nonempty Λ, let $\bar{q}_\Lambda(q)$ be a sender of type q's gross expected payoff, given that the receiver uses Bayes rule and believes her to be of some type belonging to Λ. That is,

$$\bar{q}_\Lambda(q) = \int_{x \in X} \left(\sum_{q' \in \Lambda} q' \frac{g(x|q')f(q')}{\sum_{q'' \in \Lambda} g(x|q'')f(q'')} \right) g(x|q)dx. \tag{3}$$

[7]Note that if g has less than full support, (ii) would need to be modified to read "...such that if there exists a $q \in \{q' \mid s_{q'} = s\}$ such that $g(x|q) > 0$ then..."

[8]That is, suppose the receiver only knows that the agent belongs to the set Λ with priors based on f and may subsequently adjust those priors based on new information (*i.e.,* x).

The term within the parentheses is the receiver's Bayesian estimate of the sender's quality having observed x. Integrating over all $x \in X$ yields a type-q sender's *ex ante* expected payoff from pooling with the agents in Λ. It is easy to see that if $\Lambda = \{q'\}$ for $q' \in Q$ (*i.e.*, it is a singleton) then $\bar{q}_\Lambda(q) = q'$ for any $q \in Q$. That is, if q' is the only type sending some signal s then upon observing s the receiver must believe that the sender is of type q'.

We will consider the conditional distribution, g, to be *informative* if and only if for any $|\Lambda| \geq 2$ and for any $q, q' \in Q$, whenever $q < q'$ then $\bar{q}_\Lambda(q) < \bar{q}_\Lambda(q')$. A sufficient condition for this to hold is that x and q are affiliated or, equivalently, that $g(x|q)$ satisfies the monotone likelihood ratio property. Note that types sending the same endogenous signal are imperfectly separated by the extra information since g has full support. This implies that $\min \Lambda < \bar{q}_\Lambda(q) < \max \Lambda$ for all $q \in Q$ and $|\Lambda| \geq 2$.

3.1. **Equilibria.** In a signaling equilibrium, $\mathbf{s} = (s_L, s_M, s_H)$, perfect separation implies each type's expected payoff is equal to her quality, $E_\mu[q'|s_q, q] = q$ for $q \in Q$. The distribution of the extra information, $g(x|q)$, therefore plays no role in equilibrium so the standard result for signaling games with cost functions satisfying the single-crossing property still applies.

Proposition 1. *Signaling equilibria always exist.*

Note that the payoffs and signals in a signaling equilibrium are independent of the distribution of the extra information, $g(x|q)$. As we will see, this distribution plays a significant role in the partial pooling that occurs in a countersignaling equilibrium.

With only three types, a countersignaling equilibrium must have Lows and Highs pooling so there are two candidate classes of pure strategy countersignaling equilibria: u-shaped equilibria and hump-shaped equilibria. Since the former class can be ruled out by our informational assumptions,[9] all countersignaling equilibria must be in the latter class. Suppose that senders play strategy $\mathbf{s}^* = (s^*, s^*_M, s^*)$ where $s^* < s^*_M$. Let μ describe beliefs that are Bayes consistent with playing $\mathbf{s}^*$. Then the expected gross payoff to sender q from signal s^*_M is

$$E_\mu[q'|s^*_M, q] = M$$

and the expected gross payoff to sender q from signal s^* is

$$E_\mu[q'|s^*, q] = \int_{x \in X} (\mu(L|s^*, x)L + \mu(H|s^*, x)H)g(x|q)dx.$$

[9]Suppose $\mathbf{s}^* = (s^*, s^*_M, s^*)$ is a countersignaling equilibrium where $s^* > s^*_M$. Since $\mathbf{s}^*$ is an equilibrium then both $\bar{q}_{\{L,H\}}(L) - M \geq c(s^*, L) - c(s^*_M, L)$ and $\bar{q}_{\{L,H\}}(M) - M \leq c(s^*, M) - c(s^*_M, M)$. Furthermore, since the cost function for Mediums is flatter than that for Lows (the single-crossing property), $s^* > s^*_M$ implies that $c(s^*, L) - c(s^*_M, L) > c(s^*, M) - c(s^*_M, M)$. However, this means that $\bar{q}_{\{L,H\}}(L) - M > \bar{q}_{\{L,H\}}(M) - M$, a contradiction since $\bar{q}_{\{L,H\}}(L) < \bar{q}_{\{L,H\}}(M)$.

Since Lows and Highs send the same signal s^* and since μ is Bayes consistent, $E_\mu[q'|s^*, q] = \bar{q}_{\{L,H\}}(q)$. Therefore, by assumption, $E_\mu[q'|s^*, L] < E_\mu[q'|s^*, M] < E_\mu[q'|s^*, H]$.

In a standard signaling model the difference in the gross returns from sending signals s^*_M and s^*, namely $M - E_\mu[q'|s^*, q]$, is unaffected by the sender type because $E_\mu[q'|s^*, q] \equiv E_\mu[q'|s^*]$. Since signaling costs are falling in the quality of the sender, if Mediums found it advantageous to send signal s^*_M then so would Highs. In our framework on the other hand, $M - E_\mu[q'|s^*, q]$ is falling in quality. Since the gains from signaling are lower for Highs than Mediums, Highs may choose to not signal even though signaling costs are also lower.

PROPOSITION 2. *A countersignaling equilibrium exists if $\bar{q}_{\{L,H\}}(L)$ and $\bar{q}_{\{L,H\}}(M)$ are sufficiently small and $\bar{q}_{\{L,H\}}(H)$ is sufficiently large.*

Proof. See Appendix.

In other words, the extra information must be such that Mediums will tend to look like Lows if they do not signal, but Highs will still tend to look like Highs if they do not signal. If Lows and Mediums are insufficiently separated the single-crossing property ensures that Mediums can always find a signal that Lows do not want to mimic but that costs less than $M - \bar{q}_{\{L,M\}}(M)$, so Mediums will signal. And if Highs and Lows are sufficiently separated by the extra information then Highs are better off not signaling than appearing to be Mediums. These restrictions on the extra information are illustrated more concretely in the following Proposition. As long as the distributions $g(x|L)$ and $g(x|M)$ are sufficiently similar and the distributions $g(x|L)$ and $g(x|H)$ are sufficiently dissimilar a countersignaling equilibrium exists.

PROPOSITION 3. *A countersignaling equilibrium exists if $\int_{x\in X} |g(x|L) - g(x|M)|dx$ and $\int_{x\in X} g(x|L)g(x|H)dx$ are sufficiently small.*

Proof. See Appendix.

Figure 1 illustrates the countersignaling equilibrium $(0, s^*_M, 0)$. Level sets represent sender indifference between being various payoff/signal combinations, where utility increases in a northwesterly direction. Following the single-crossing property, the sets are flattest for Highs and steepest for Lows, ensuring that the indifference curves of different types cross only once. The intercepts represent the utility payoffs of each indifference curve. Thus in this equilibrium, Highs get the greatest payoff, $\bar{q}_{\{L,H\}}(H)$, while Lows get the least, $\bar{q}_{\{L,H\}}(L)$. According to level set l, Lows are just indifferent between sending signal s^*_M (pretending to be a Medium) and sending the equilibrium signal of zero (pooling with Highs). That is, s^*_M is the minimum signal that Mediums can send and deter Lows from mimicking them. The level set h represents the utility received by a High from playing according to equilibrium and not signaling. Highs are willing to pool with Lows as long

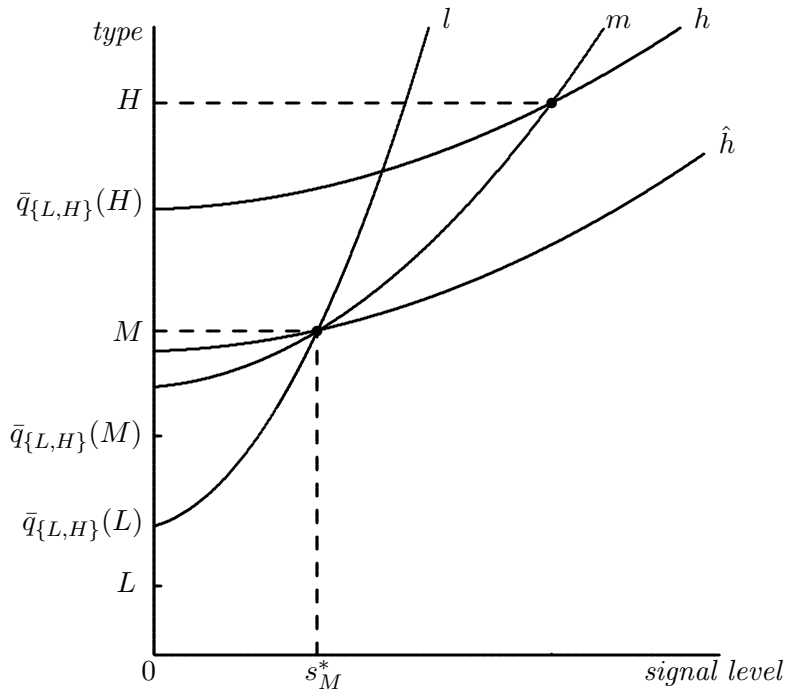


FIGURE 1. Pareto dominant countersignaling equilibrium

as they get a greater payoff than from sending signal s^*_M and pretending to be a Medium, $\bar{q}_{\{L,H\}}(H) > M - c(s^*_M, H))$. This is true in our case since indifference curve h is higher than $\hat{h}$. Finally, Mediums must prefer to send signal s^*_M than to send $s = 0$ and pretending to belong to $\Lambda = \{L, H\}$. This holds since the intercept of m (her equilibrium payoff) is greater than $\bar{q}_{\{L,H\}}(M)$.

As with the standard signaling model, ours is subject to multiple equilibria. A substantial literature has developed in an effort to "refine" away "undesirable" equilibria in signaling models (Banks and Sobel, 1987; Cho and Kreps, 1987; Cho and Sobel, 1990). Contrary to the standard signaling framework, refinements such as the Intuitive Criterion, D1 and D2 are unable to rule out pooling and partial-pooling equilibria.[10] This can be seen from the simplest two-type model. Without extra information Highs are indistinguishable from Lows in a pooling equilibrium so they always have an incentive to break away and send a signal that Lows would never mimic, thereby implying that pooling cannot survive the Intuitive Criterion. With the extra information this is no longer true. Highs are stochastically separated from Lows even as they pool with them, so they are less willing to

[10]Given that the standard refinements cannot eliminate signaling equilibria under the standard model, such refinements will not eliminate signaling equilibria in our augmented framework since such equilibria do not depend on any of the distributional information.

bear the cost of sending the minimum signal that Lows would never mimic. Moreover, the stochastic separation means that Lows gain less from pooling than when there is no extra information. This implies that if we consider the set of signals that Lows could never gain by sending, the minimum of this set is larger than when there is no extra information. Since extra information makes breaking the pooling equilibrium both less rewarding and more costly, a pooling equilibrium can survive the Intuitive Criterion if the stochastic separation is sufficient.

This same logic of pooling by Highs and Lows applies to our three-type case, except that it is even more costly for Highs to send a sufficiently large signal that Mediums would never mimic. As a result, we find that under conditions qualitatively identical to those given in Propositions 2 and 3, countersignaling equilibria continue to exist under the Intuitive Criterion, D1 and D2.[11] In particular, the conditions for the existence of the Pareto dominant countersignaling equilibrium are qualitatively identical. Such countersignaling equilibria might be, in terms of welfare, more appealing than any signaling equilibrium. This is demonstrated formally for the Intuitive Criterion as follows.

PROPOSITION 4. *If the Pareto dominant countersignaling equilibrium survives the Intuitive Criterion it Pareto dominates all signaling equilibria. In particular, every type of sender is strictly better off under the Pareto dominant countersignaling equilibrium.*

Proof. See Appendix.

The argument is roughly as follows. Suppose that the Pareto dominant countersignaling equilibrium does not Pareto dominate the Pareto dominant signaling equilibrium (the Riley equilibrium). Since Lows benefit from pooling with Highs and since Mediums can then send a lower signal to successfully ward off Lows, it must be the Highs who are worse off. However, suppose that Highs deviate and send their equilibrium signal from the Riley equilibrium. With probability 1 they would be thought to be Highs since Mediums would never be willing to send this signal. Since their signaling payoff is greater than their countersignaling payoff, Highs have an incentive to deviate from the countersignaling equilibrium by playing according to the signaling equilibrium. So if the countersignaling equilibrium does not Pareto dominate the Riley equilibrium it must not pass the Intuitive Criterion.

In practice Proposition 4 might overstate the efficiency of countersignaling. We have assumed the receiver is risk neutral, but if the receiver is risk averse or benefits from matching senders to particular jobs based on their

[11]Proofs are available at `http://www.Theo.To/counter/refine.pdf` or from the authors upon request. Note that the value of $\bar{q}_{\{L,H\}}(H)$ as shown in Figure 1 is the smallest such value for which the countersignaling equilibrium depicted survives the Intuitive Criterion, D1 and D2.

quality, the loss in information to the sender in the countersignaling equilibrium might exceed any cost savings to the senders. As discussed below, inefficiencies can also result if signaling is to some extent a desirable activity.

3.2. **Extensions.** In the first three of the following subsections, let $\mathbf{s} = (0, s_M, s_H)$ and $\mathbf{s}^* = (0, s_M^*, 0)$ represent the Riley equilibrium and the Pareto dominant countersignaling equilibrium.

Productive Signaling. While the wasteful nature of signaling is often emphasized in the literature, many forms of signaling are, in moderation, productive or otherwise desirable. For instance, while education may be excessive in a signaling equilibrium, it is often a preferred activity in moderation. When signaling is to some extent desirable, countersignaling by Highs might be inefficient because of insufficient signaling.

To illustrate the point in a simple manner, suppose we relax the condition that signaling costs are strictly increasing in the signal. In particular, assume that costs are initially decreasing for Highs but eventually increasing. For Lows and Mediums the cost structure is unchanged. If signaling costs do not decrease too rapidly for Highs the equilibrium $(0, s_M^*, 0)$ still exists even though in a perfect information environment Highs would choose a strictly positive signal. In other words, in a countersignaling equilibrium there can be too little rather than too much signaling.

Bounded Signals. So far we have assumed that the signaling range has no upper bound. While in many cases this might be quite reasonable, it might be more realistic in other cases to include an upper bound on the highest signal that can be sent, *e.g.*, the best signal that a high school student can send is to get straight A's. Consider if there is some maximum signal $\bar{s}$ so that $S = [0, \bar{s}]$. If $\bar{s} < s_H$ then the signaling equilibrium cannot exist. However, if $\bar{s} \geq s_M^*$ then the countersignaling equilibrium still exists. Thus by eliminating the possibility for a signaling equilibrium, putting an upper limit on the signal can be considered conducive to countersignaling. Highs cannot signal their capabilities relative to Mediums by sending a higher signal, so the only alternatives are countersignaling equilibria and other partial-pooling or pooling equilibria, all of which imply a loss of information to the receiver.

Alternative equilibria. Real world signaling behavior may be more complicated than the pure strategy equilibria we have derived in our simple three–type example. For instance, even in situations conducive to countersignaling, some high types may be observed to be countersignaling while others may be observed to be signaling. Three means of getting more complicated signaling behavior are i) to add more types, ii) to look at mixed strategy equilibria or iii) to add a slightly more complicated information structure. We briefly consider each of these possibilities in turn.

Suppose there is a fourth type, $H^+ \geq H$ for which signaling is completely costless. This modification can yield a "counter-countersignaling" equilibrium where L, M and H types play according to $\mathbf{s}^*$ and type H^+ agents send an arbitrarily large signal. The presence of H^+ types has no effect on equilibrium beliefs over L, M and H types but a sufficiently large signal, $s^*_{H^+}$, will deter imitation by any type even under beliefs which survive the Intuitive Criterion, D1 and D2. Obviously, less extreme cost structures will yield yet other types of counter-countersignaling behavior. For example, it may be that rather than sending a higher signal, there may be equilibria where H^+ types pool with M's.

Returning to the three-type example, now consider the possibility that senders play mixed strategies or that some proportion of each type plays different strategies. In particular, consider the mixed strategy profile where Lows and Mediums and $1-\Delta$ of the Highs play according to $\mathbf{s}^* = (0, s^*_M, 0)$ and the remaining Δ of the Highs send a signal, $\dot{s}^*_H$, at which they get H and are indifferent between sending 0 and $\dot{s}^*_H$. Since fewer high types are now pooling with the low types, this will have the effect of reducing $\bar{q}_{\{L,(1-\Delta)H\}}(q)$ for all $q \in Q$. Provided that $\bar{q}_{\{L,(1-\Delta)H\}}(H) \geq M - c(s^*_M, H)$ and $\bar{q}_{\{L,(1-\Delta)H\}}(M) \leq M - c(s^*_M, M)$, this strategy profile is clearly an equilibrium. Furthermore, if $\bar{q}_{\{L,(1-\Delta)H\}}(H)$ is sufficiently large, $\mathbf{s}^*$ survives the Intuitive Criterion, D1 and D2. Notice that when pure strategy countersignaling equilibria exist, there is in general a continuum of Δ's with equilibria where some Highs signal and some countersignal.

Note that more complicated signaling patterns can also arise when there are multiple sources of extra information. Suppose, as discussed earlier, that there are two sources of extra information, x and y. The former is observed by both the sender and receiver and the latter is observed only by the receiver. In this case, for each $x \in X$, the distribution of types is $\hat{f}(q) = g(x|q)f(q)/\sum_{q' \in Q} g(x|q')f(q')$ and y plays the role of the extra information in our model. In other words, for each realization of x a different game is played so that the same type q might signal or countersignal depending on the realization of x and on the particular equilibria in the game corresponding to that realization of x.[12]

Countersignaling with a continuum of types. When types form a continuum and signals are continuous, non-monotonicities could arise in a variety of forms. We present a simple example in which types of highest and lowest quality send a zero signal while types within an intermediate range send the same signals as in the fully separating Riley equilibrium.[13] In particular, assume that types are distributed uniformly over the unit interval, $Q = [0, 1]$ and that $X = \mathbb{R}$ and $x = q + \epsilon$ where ϵ is a random variable distributed

[12]The role of fully observed information in allowing for different equilibria conditional on the information is analyzed by Spence (1973a) in the context of a regular signaling game.

[13]This example is based on a suggestion by Barry Nalebuff.

normally with zero mean and standard deviation 1/4. Consistent with the single-crossing property, let the cost function be $c(q,s) = s/q^3$.

In a separating equilibrium each type is believed to send a unique signal and the extra information has no impact. Representing this mapping from signals to receiver inferences of type by $\hat{q}(s)$, the return to a signal is then $\hat{q}(s) - s/q^3$. Maximizing with respect to s gives $d\hat{q}(s)/ds = 1/q^3$. In equilibrium, beliefs are consistent with actions so $\hat{q}(s) = q$, implying $q^3 dq = ds$. Integrating gives the family of solutions $s = q^4/4 + K$, each of which is a separating equilibrium. Of these the Riley equilibrium $s = q^4/4$ is the only reasonable solution since type $q = 0$ never benefits from sending a positive signal.

We are interested in a countersignaling equilibrium where there are two types $0 < q_a < q_b < 1$ such that $s_q^* = q^4/4$ for $q \in [q_a, q_b]$ and $s_q^* = 0$ for $q < q_a$ and $q > q_b$. Since the distribution of types is a continuum, let $\mu(q|s,x)$ represent a probability measure in this section rather than a density function. Bayes consistent receiver beliefs for $s = 0$ and x are therefore

$$d\mu(q|0,x) = \frac{\phi(4(q-x))}{\int_0^{q_a} \phi(4(q'-x))dq' + \int_{q_b}^1 \phi(4(q'-x))dq'}$$

where $\phi(\cdot)$ represents the probability density function of a standard normal distribution. For $s \in (0, 1/4]$ we assume $\mu((4s)^{1/4}|s,x) = 1$, just as in the Riley equilibrium. Since not all signals in this range are actually sent in the countersignaling equilibrium, we are thereby assuming that if a signal is observed which would not be sent in the countersignaling equilibrium, the receiver believes it was sent by the type that would have sent it in the Riley equilibrium. Finally, for $s > 1/4$, let $\mu(1|s,x) = 1$. Given these beliefs, the expected gross payoff to sender q from signal $s = 0$ is

$$E_\mu[q'|0,q] = \int_{-\infty}^{\infty} \left(\int_0^{q_a} q' d\mu(q'|0,x)dq' + \int_{q_b}^1 q' d\mu(q'|0,x)dq' \right) \phi(4(x-q))dx.$$

The conditions for the marginal types to be indifferent are

$$E_\mu[q'|0,q] - c(0,q) = q - c(q^4/4, q) \quad \text{for } q = q_a, q_b.$$

Solving numerically, one solution is $q_a \approx 0.521$ and $q_b \approx 0.961$. Further calculations confirm types $q < q_a$ and $q > q_b$ prefer $s = 0$ to $s = q^4/4$ and types $q_a < q < q_b$ prefer $s = q^4/4$ to $s = 0$. The signal level stays at zero for $q < q_a$ but then jumps up to track the Riley equilibrium over the range $[q_a, q_b]$, before falling back to zero for $q > q_b$. Higher types not only save costs by countersignaling, but since $E_\mu(q|0,q_b) \approx 0.728 > q_a$ they are, in expectation, estimated to be of higher quality than many types sending a strictly positive signal.

4. A Test of Countersignaling

To investigate whether agents can countersignal, we ran a simple experiment with two cells, one corresponding to a standard signaling game (the

Player Type (Skill Level)	Number of Type	Cost of Good Grade	% Passing Test S cell	% Passing Test C cell	Productivity
High	4	−25	50	100	900
Medium	8	+25	50	50	700
Low	4	+350	50	0	400

TABLE 1. Characteristics of Player Types

S cell) and the other to a countersignaling game (the C cell). We use a three–type model closely related to the example in Section 2.

4.1. **Description of the Game.** Both the S and C cells share the same basic structure. We motivated the game to the subjects as an education model in which students signal their skill levels to firms. There are three types of student: High, Medium, and Low skill level. They signal in two ways: by grades, which they choose, and by test scores, which are exogenous and have a random component. After students have signaled, they are hired by competitive risk–neutral firms, so each student receives a wage equal to that student's expected productivity, conditional on the student's grade and test score. The role of the firms is suppressed in our experiment; their role is played by computer, rather than by human subjects.[14]

The parameters used in the game are shown in Table 1. The population of students consists of 4 Highs, 8 Mediums, and 4 Lows.[15] Grade is a binary choice, either G (good) or B (bad). A bad grade is costless, while the cost of a good grade varies inversely with skill level and in the case of Highs is negative, i.e., Highs actually receive a direct benefit from signaling. Test scores are also binary, either P (pass) or F (fail). Test scores do not depend directly on grades. In the C cell the probability of passing is increasing in skill level, while in the S cell, the probability is 0.5, irrespective of skill level.

Even though the exogenous signal is still present in the S cell, it is completely uninformative; there is no difference—even probabilistically—in the extra information sent by the different types of student. This game thus reduces to a standard signaling model in which higher–quality types signal

[14]The technique of automating some parts of a game in order to simplify an experimental environment is quite common and has been done in a wide variety of settings, including other signaling games (Cooper et al., 1997a,b), simple markets (Roth et al., 1991), Cournot–Stackelberg duopoly games (Huck et al., 1999), common–value auctions (Garvin and Kagel, 1994; Kagel and Levin, 1986), and asymmetric–information "lemons" markets (Ball et al., 1991). Note that a consequence of our design is that the appropriate equilibrium concept is now Nash equilibrium since all players are choosing simultaneously. Of course, any Nash equilibrium of this simplified game corresponds to a sequential equilibrium of the original game with firms' payoffs appropriately specified.

[15]Since the only decision–making agents will be the students, we use "student" and "player" interchangeably.

to distinguish themselves from lower–quality types; the unique Nash equilibrium outcome of the S cell has all Highs and Mediums choosing a good grade and none of the Lows doing so. In contrast, the extra information in the C cell makes this a countersignaling environment. Lows always fail the exam and Mediums expect to fail the exam half the time, so Mediums can use good grades to differentiate themselves from Lows. Highs need not distinguish themselves from Lows since Highs always pass and Lows always fail. Instead, they are concerned with being mistaken for those Mediums who pass. Since Mediums earn good grades, Highs react by earning bad grades even though they would prefer a good grade in a complete information environment. In the unique equilibrium is that Mediums choose good grades while Highs and Lows choose bad grades.

Our design makes the test of countersignaling difficult in two ways. First, we test for countersignaling when Highs have negative signaling costs, *i.e.*, they receive a bonus for signaling, implying they would always signal in a perfect information environment. Second, we give the games a somewhat normative context in which signaling is described as getting a "good grade" and not signaling as getting a "bad grade."

4.2. **Experimental Procedures.** The experiment consisted of four S sessions and four C sessions. Subjects were mostly undergraduates at the University of Pittsburgh. Sessions lasted for roughly 90 minutes and consisted of one practice round and 12 rounds, except for one S session which, due to time constraints, consisted of one practice round and 10 rounds.[16] Subjects were not told how many rounds would be played, though they probably had some idea of when they were close to the end of the game since they were told the experiment would last no longer than 90 minutes. The experiment was conducted with pen and paper. Instructions were read aloud to subjects, and a table similar to Table 1 (but with only the information concerning the cell they were in) was written at the front of the room. Subjects were given record sheets with spaces to write for each round their skill level, grade, test score, salary (called "gross payoff"), cost of earning a good grade,[17] and net payoff. In each round, each subject drew one of sixteen slips of paper, on which were printed a skill level and test score, and a space for subjects to write their choice of grade. The slips of paper were prepared in advance and were repeatedly folded and sealed so that test scores could not be seen without breaking the seal. Skill levels and test scores were block–randomly assigned to the slips of paper, so that, in the S cell for example, in each round there were exactly 4 Mediums that passed the test, 2 Lows that failed the test, and so on.

[16]Keeping the number of rounds small has the advantages that subjects have time to think about their decisions and that payoffs per decision are relatively high.

[17]Because Highs have a negative "cost" of earning a good grade, we used the phrase "bonus or penalty" in the experiment.

		Grade: G		Grade: B	
Test Score	P	#High:	2	#High:	0
		#Medium:	4	#Medium:	0
		#Low:	0	#Low:	2
		Salary:	734	Salary:	400
	F	#High:	2	#High:	0
		#Medium:	4	#Medium:	0
		#Low:	0	#Low:	2
		Salary:	734	Salary:	400

(a) Signaling Cell

		Grade: G		Grade: B	
Test Score	P	#High:	0	#High:	4
		#Medium:	4	#Medium:	0
		#Low:	0	#Low:	0
		Salary:	700	Salary:	900
	F	#High:	0	#High:	0
		#Medium:	4	#Medium:	0
		#Low:	0	#Low:	4
		Salary:	700	Salary:	400

(b) Countersignaling Cell

TABLE 2. Sample of post-round information given to subjects

In each round, each subject copied her skill level onto her record sheet, chose her grade, and wrote this grade on her record sheet and on the slip of paper. After this was done, a monitor came to the subject's desk and watched her break the seal, revealing her test score. The subject wrote her test score on her record sheet and the monitor collected the slip of paper. When all the slips had been collected, the distribution of skill levels and the salary corresponding to each (grade, test score) pair were posted at the front of the room. Examples of this posted information are given in Tables 2(a) and 2(b). Subjects then wrote their salaries on their record sheets, and calculated and recorded their net payoffs. The next round then began. The posted information remained posted until it was replaced by information from the following round.

At the end of the session, one of the non–practice rounds was randomly selected from those played, and subjects were paid in cash their net payoffs from that round at the exchange rate of 100 points/$1.00, in addition to a $5.00 participation fee. Average earnings were approximately $11.00.

	C Sessions			S Sessions		
	High	Medium	Low	High	Medium	Low
Round 1	15/16	24/32	2/16	16/16	25/32	3/16
	(.938)	(.750)	(.125)	(1.000)	(.781)	(.188)
Rounds 1–3	47/48	77/96	6/48	47/48	74/96	12/48
	(.979)	(.802)	(.125)	(.979)	(.771)	(.250)
Rounds 10–12	27/48	72/96	2/48	38/40	76/80	9/40
	(.562)	(.750)	(.042)	(.950)	(.950)	(.225)
Round 12	7/16	23/32	1/16	11/12	23/24	3/12
	(.438)	(.719)	(.062)	(.917)	(.958)	(.250)
Equilibrium Prediction	.000	1.000	.000	1.000	1.000	.000

TABLE 3. Aggregate Early– and Late–Round Frequency of Signaling

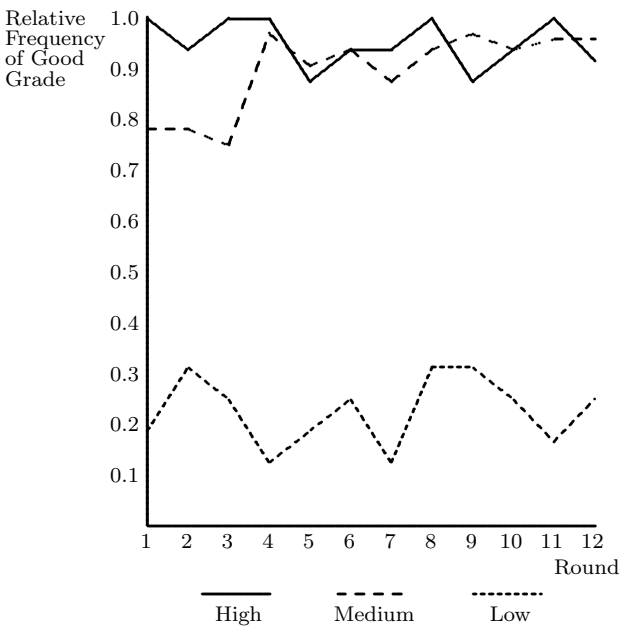
4.3. **Experimental Results.** Figures 2(a) and 2(b) show the relative frequencies of good grades in the two cells.[18] Early–round play is very similar in the two cells, but that play begins to diverge thereafter as Highs choose good grades less and less frequently in the countersignaling cell. Mediums in the signaling cell increase the frequency of choosing a good grade somewhat, with no such increase in the countersignaling cell. Lows also appear to choose good grades less and less frequently in the countersignaling cell, but this difference is slight.

Table 3 reports the frequency with which players in the two cells chose G in early and in late rounds.[19,20] Thus, at least in early rounds, players do not seem to behave in accordance with the equilibrium predictions. However, some aspects of play in later rounds are consistent with the theory's predictions. Recall that the Nash equilibrium predicts Highs should choose G more often in the S cell than in the C cell, while play of Mediums and

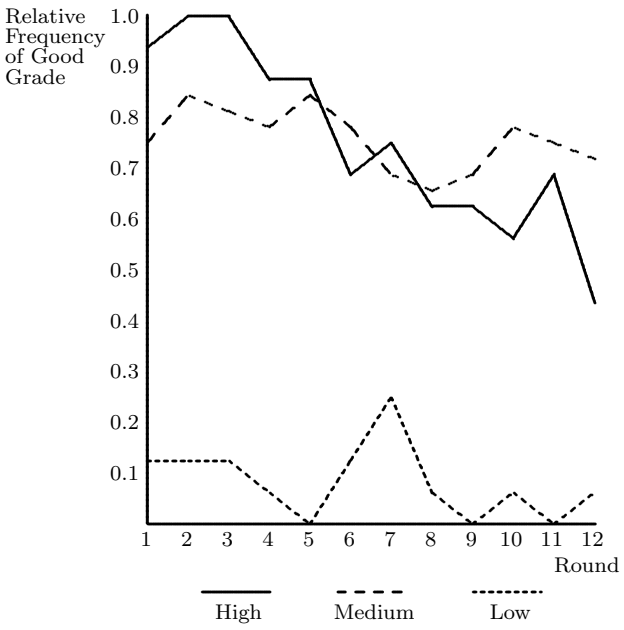
[18]The raw data and instructions from the experiment are available at `http://www.Theo.To/counter/experiment.xls` and `http://www.Theo.To/counter/instructions.pdf` or from the authors upon request.

[19]Play is remarkably similar across the four S sessions, and for the most part is similar across C sessions. The one exception is Highs in the C cell: the frequency of good grades is 58.3%, 68.8%, 79.2%, and 95.8% in the four C sessions, while no other type in either cell has a range of 20 percentage points or more. According to one–tailed permutation tests (Siegel and Castellan, 1988) on the session–level data, there are no significant differences (at even the 10% level) in early–round frequencies of G between the S and C cells.

[20]Permutation tests make no assumptions about the underlying population distributions, unlike the commonly–used Wilcoxon–Mann–Whitney test, which gives broadly similar results in this case, but is inappropriate because it assumes that second– and higher–order moments of the population distributions are the same and can therefore give rejections of the null hypothesis based solely or partly on differences in, for example, variances. See Siegel and Castellan (1988) for a discussion of this issue, as well as more thorough descriptions of the nonparametric statistical tests used in this paper.



(a) Signaling Cell



(b) Countersignaling Cell

FIGURE 2. Plot of Relative Signaling Frequency

Lows should be the same in both cells. In rounds 10–12 Highs are far more likely to play G in the S cell than in the C cell, and this difference is significant (permutation test, $p < .05$) However, Mediums and Lows also play G significantly more often in the S cell than in the C cell (permutation test, $p < .05$).[21]

Because not only Highs but also Mediums choose G more often in the C cell than in the S cell, differences between the play of Highs and that of Mediums in the C cell are smaller than one would hope, even in late rounds of the experiment. (Differences in the S cell are even smaller, but that is exactly the equilibrium prediction.) A chi–square test of the difference between the play of Highs and that of Mediums in round 12, using the individual–level data, gives a p–value of about 0.12, which is suggestive, but not significant at standard levels. A robust rank–order test, using the individual–level data from rounds 10–12, gives a p–value of 0.0643.[22] Using session–level data, rather than individual–level data, yields no significant differences between the play of Mediums and that of Highs in the C cell.

4.4. **Discussion.** The only difference between our S and C cells is that in the S cell the extra information is uninformative, while in the C cell it is informative. There is no difference in early–round play between the cells, but eventually differences emerge in the direction predicted by countersignaling theory, though not always significant at standard levels. The patterns of behavior observed can best be described with the following "learning" story. In the S cell, bad grades are a dominant strategy for Lows; by choosing bad grades, they earn at least 400, while good grades earn them at most 383.33. Indeed, we see that Lows, for the most part, quickly learn to get bad grades. Once most Lows are choosing bad grades, Mediums and Highs do best by earning good grades, and they learn this quickly, too. Play thus moves quickly toward the equilibrium.[23]

[21]Lower choices of G by Mediums and Lows in the C cell are not surprising given the differences between the cells. In the signaling cell choosing G by Mediums and Lows has the advantage of pooling with Highs who choose G frequently. Although the cost of G for Lows is sufficiently high that G is never a best response, the net payoff loss is small. In the countersignaling cell Highs choose G less frequently so the advantage to Mediums of choosing G is smaller. And since Highs and Lows are always distinguished in the countersignaling cell, there is no chance for Lows to pool with Highs by choosing G, increasing the net payoff loss to Lows of choosing G.

[22]In order to implement this test, as well as to eliminate one possible source of dependence among data points, play over each player and type was averaged. Specifically, if the player with ID# 6 was a Medium type in rounds 11 and 12, and played G in 11 and B in 12, she was listed as having chosen G with relative frequency 0.5. If the player with ID# 7 was a Medium type in only round 10, and played B, he was listed as having chosen G with relative frequency 0.

[23]There is a consistent small fraction of Lows choosing good grades, even in the last rounds of the experiment. This may be due to the fact that in equilibrium, the wage of students with bad grades is 400, while that of students with good grades is 766.67. Subtracting the 350 cost of good grades yields 416.67—higher than 400—so a myopic Low

In the C cell, bad grades are again dominant for Lows, and the payoff differential is even larger—now, good grades earn them at most 290. Again, they quickly learn to choose bad grades. Once most of the Lows are choosing bad grades, good grades become a best response for Mediums, and they indeed tend to choose good grades. Once Mediums are choosing good grades, Highs do better by choosing *bad* grades. While the experiment never reaches the point where all Highs choose bad grades, by the final round slightly more than half of them do so, while hardly any Highs in the S cell ever choose bad grades.

One reason for the slow convergence of play in the C cell to the countersignaling equilibrium may be the incentives Highs face when play is not in equilibrium. As shown in Figure 2(b), Highs begin the session by choosing good grades. Given the belief that the other three Highs are going to continue choosing good grades, the fourth High should choose bad grades only if at least seven of the Mediums choose to earn good grades, but the fraction of Mediums choosing good grades in the C cell is rarely this high (only in 18 rounds out of 48). However, as soon as one High is choosing bad grades the incentive for the remaining Highs to also countersignal is stronger.[24]

5. Conclusion

Addition of noisy information on type to a standard signaling model allows for equilibria in which medium types signal to distinguish themselves from low types but high types do not. Such countersignaling by high types can be seen as a sign of confidence. Signaling proves the sender is not a low type but also reveals the sender's insecurity that they would be perceived as such if they did not signal. In contrast, countersignaling indicates the sender's faith that whatever other information the receiver has on the sender will probably be consistent with the sender being of high quality.

Countersignaling captures the intuition that many of the highest quality senders may be understated rather than overstated in their signaling behavior. As a result, countersignaling equilibria can invert a number of the main implications of signaling models. Whereas signaling equilibria can be inefficient because of excessive signaling, countersignaling equilibria may be inefficient because of inadequate signaling. While signaling equilibria can play an informational role in increasing the efficiency of receiver estimates of type, countersignaling equilibria may lower the efficiency of these estimates. And while higher costs tend to reduce signaling in a signaling model, a

might mistakenly choose good grades in the hope of earning a higher payoff. This ignores his own effect on these wages; by choosing good grades, he becomes part of this group, reducing the group's expected productivity. The wage of this group falls to 714.29, so that the L's payoff decreases to 364.29, and his hopes are dashed.

[24]This "positive feedback" might explain the higher variance across sessions in play of the Highs in the C cell.

limited increase in costs can lead to more signaling in a countersignaling model.

The extra information information in a countersignaling model can allow a wide range of pooling and partial-pooling equilibria to survive standard refinements that leave only a unique, separating equilibrium in a signaling model. Both signaling and countersignaling equilibria may coexist, along with mixed strategy equilibria where some high types signal and others countersignal. When there are more than three types, counter-countersignaling equilibria are also possible in which the very highest types signal to separate themselves from countersignaling high types.

Since countersignaling is more complicated behavior than signaling, the question of whether economic agents can countersignal was tested with a two-cell experiment in which extra information on sender type was available. In one cell the extra information was completely uninformative and signaling by medium and high types was the unique Nash equilibrium. In the other cell the extra information was partially informative and the unique Nash equilibrium involved countersignaling by high types even though high types had a negative cost to signaling. The experimental results confirm that adding noisy exogenous information on types to signaling games can affect behavior in directions consistent with the predictions of countersignaling theory. Countersignaling by high types was rare in the signaling cell but was the most common choice by the last period of the countersignaling cell.

Appendix A. Proofs

Recall that $\bar{q}_{\{L,H\}}(q) = E_\mu[q'|s^*, q]$ is the expected gross payoff of a type-q individual whom the receiver believes to belong to $\{L, H\}$. Let $\hat{s}^*_L$ solve $\bar{q}_{\{L,H\}}(L) - c(\hat{s}^*_L, L) = L$. Since $\bar{q}_{\{L,H\}}(L) > L$, $\hat{s}^*_L > 0$. Now for $s^* \geq 0$, let $\tilde{s}^*_M(s^*)$ and $\hat{s}^*_M(s^*)$ solve $\bar{q}_{\{L,H\}}(L) - c(s^*, L) = M - c(\tilde{s}^*_M, L)$ and $M - c(\hat{s}^*_M(s^*), M) = \bar{q}_{\{L,H\}}(M) - c(s^*, M)$. Note that $\tilde{s}^*_M(s^*)$ is the minimum signal required to deter Lows from imitating the Mediums and $\hat{s}^*_L$ and $\hat{s}^*_M(s^*)$ are the maximum signals that Lows and Mediums are willing to send before they would prefer to send some alternative signal.

Lemma 1. *A countersignaling equilibrium exists if and only if there is an s^* and an s^*_M such that $s^* \in [0, \hat{s}^*_L]$, $s^*_M \in [\tilde{s}^*_M(s^*), \hat{s}^*_M(s^*)]$ and $\bar{q}_{\{L,H\}}(H) - c(s^*, H) \geq M - c(s^*_M, H)$.*

Proof. ($\Leftarrow$) For any $x \in X$, let beliefs be given by $\mu(L|s, x) = 1$ whenever $s \notin \{s^*, s^*_M\}$. Let $\mu(L|s^*, x) = g(x|L)f(L)/(g(x|L)f(L) + g(x|H)f(H))$, $\mu(H|s^*, x) = 1 - \mu(L|s^*, x)$ and $\mu(M|s^*_M, x) = 1$. These beliefs clearly satisfy (ii) of the definition of a PBE. Thus we need to show that each agent's best response is to follow their prescribed strategy. Obviously no type has an incentive to choose $s \notin \{s^*, s^*_M\}$ since they would receive a payoff of $L - c(s, q)$ which is no greater than the payoff L they would get by choosing $s = 0$.

Now, since $s^* \leq \hat{s}_L^*$, $E_\mu[q'|s^*, L] - c(s^*, L) \geq \bar{q}_{\{L,H\}}(L) - c(\hat{s}_L^*, q) = L$ it is individually rational for a type-L player to choose s^*. Since, $s_M^* \geq \tilde{s}_M^*(s^*)$, it follows that $E_\mu[q'|s^*, L] - c(s^*, L) = M - c(\tilde{s}_M^*(s^*), L) \geq M - c(s_M^*, L)$ so no L-type individual has an incentive to choose $s = s_M^*$.

Since $s_M^* \leq \hat{s}_M^*(s^*)$, $M - c(s_M^*, M) \geq M - c(\hat{s}_M^*(s^*), M) = \bar{q}_{\{L,H\}}(M) - c(s^*, M)$ no type-M individual has an incentive to choose $s = s^*$.

Finally, in order for Highs to be willing to send signal s^*, they must get at least as much as they would if they imitated the Mediums (*i.e.*, $\bar{q}_{\{L,H\}}(H) - c(s^*, H) \geq M - c(s_M^*, H)$).

($\Rightarrow$) Follows by reversing the previous arguments. ■

Proof of Proposition 2. We already know the interval $[0, \hat{s}_L^*]$ is non-empty so it remains to be shown that if $\bar{q}_{\{L,H\}}(L)$ and $\bar{q}_{\{L,M\}}(M)$ are sufficiently small then $[\tilde{s}_M^*(s^*), \hat{s}_M^*(s^*)]$ is non-empty and if $\bar{q}_{\{L,M\}}(H)$ is sufficiently large then $\bar{q}_{\{L,H\}}(H) - c(s^*, H) \geq M - c(s_M^*, H)$.

First, note that $\tilde{s}_M^*(s^*) > 0$ exists if and only if $\bar{q}_{\{L,H\}}(L) - c(s^*, L) < M$. This is true if and only if $\bar{q}_{\{L,H\}}(L)$ is sufficiently small.

Next, since we are only looking for sufficient conditions, take s^* to be "small." For s^* small, $\tilde{s}_M^*(s^*)$ and $\hat{s}_M^*(s^*)$ satisfy $c(\tilde{s}_M^*(s^*), L) \approx M - [\bar{q}_{\{L,H\}}(L) - c(s^*, L)]$ and $c(\hat{s}_M^*(s^*), M) \approx M - [\bar{q}_{\{L,H\}}(M) - c(s^*, M)]$. If g is such that $\bar{q}_{\{L,H\}}(M)$ is "close" to $\bar{q}_{\{L,H\}}(L)$ then $c(\tilde{s}_M^*(s^*), L) \approx c(\hat{s}_M^*(s^*), M)$. Since $c(s, M) < c(s, L)$ if follows that $\tilde{s}_M^*(s^*) < \hat{s}_M^*(s^*)$. That is, when $\bar{q}_{\{L,H\}}(M)$ is sufficiently small (i.e., close to $\bar{q}_{\{L,H\}}(L)$), the interval $[\tilde{s}_M^*(s^*), \hat{s}_M^*(s^*)]$ is non-empty.

Finally, given some $s_M^* \in [\tilde{s}_M^*(s^*), \hat{s}_M^*(s^*)]$, it is clear that when $\bar{q}_{\{L,H\}}(H)$ is sufficiently close to H, $\bar{q}_{\{L,H\}}(H) - c(s^*, H) \geq M - c(s_M^*, H)$. ■

Proof of Proposition 3. Let $\int_{x \in X} g(x|L)g(x|H)dx < \varepsilon_1$ and $\int_{x \in X} |g(x|M) - g(x|L)|dx < \varepsilon_2$.

From the proof of Proposition 2, first we need to show that $\overline{q}_{\{L,H\}}(L)$ is sufficiently small.

$$\begin{aligned}
L - \overline{q}_{\{L,H\}}(L) &= \int_{x\in X} \left(\frac{g(x|L)f(L)}{g(x|L)f(L)+g(x|H)f(H)} L + \frac{g(x|H)f(H)}{g(x|L)f(L)+g(x|H)f(H)} H \right) g(x|L)dx - L \\
&= \int_{x\in X} \frac{g(x|H)f(H)g(x|L)}{g(x|L)f(L)+g(x|H)f(H)} \left(H-L\right) dx \\
&\leq \frac{1}{f(L)} \int_{x\in X} \frac{\sqrt{g(x|H)f(H)g(x|L)f(L)}}{2} \left(H-L\right) dx \\
&= \frac{1}{2}\sqrt{\frac{f(H)}{f(L)}} \left(H-L\right) \int_{x\in X} \sqrt{g(x|L)g(x|H)} dx \\
&\leq \frac{1}{2}\sqrt{\frac{f(H)}{f(L)}} \left(H-L\right) \sqrt{\int_{x\in X} g(x|L)g(x|H)dx} \\
&< \sqrt{\frac{f(H)}{f(L)}} \frac{(H-L)}{2} \sqrt{\varepsilon}_1
\end{aligned}$$

where the third line follows from the fact that $ab/(a+b) \leq \sqrt{ab}/2$ for $a, b > 0$. Hence $\overline{q}_{\{L,H\}}(L) - L$ can be made arbitrarily small by choice of ε_1.

Next we need to show that $\overline{q}_{\{L,H\}}(M)$ is sufficiently close to $\overline{q}_{\{L,H\}}(L)$.

$$\begin{aligned}
\overline{q}_{\{L,H\}}(M) - \overline{q}_{\{L,H\}}(L) = &\int_{x\in X} \left(\frac{g(x|L)f(L)}{g(x|L)f(L)+g(x|H)f(H)} L + \frac{g(x|H)f(H)}{g(x|L)f(L)+g(x|H)f(H)} H \right) \times \\
&\left(g(x|M) - g(x|L)\right) dx \\
\leq &\int_{x\in X} \left(\frac{g(x|L)f(L)}{g(x|L)f(L)+g(x|H)f(H)} L + \frac{g(x|H)f(H)}{g(x|L)f(L)+g(x|H)f(H)} H \right) \times \\
&|g(x|M) - g(x|L)| dx \\
\leq &H \int_{x\in X} |g(x|M) - g(x|L)| dx \\
< &H\varepsilon_2.
\end{aligned}$$

Hence $\overline{q}_{\{L,H\}}(M) - \overline{q}_{\{L,H\}}(L)$ can be made arbitrarily small by choice of ε_2.

Finally we need to show that $H - \overline{q}_{\{L,H\}}(H)$ is sufficiently small. Following the same logic as in the first step above,

$$H - \overline{q}_{\{L,H\}}(H) < \sqrt{\frac{f(L)}{f(H)}} \frac{(H-L)}{2} \sqrt{\varepsilon}_1.$$

so $H - \overline{q}_{\{L,H\}}(H)$ can be made arbitrarily small by choice of ε_1. ■

Proof of Proposition 4. Let the Pareto dominant signaling equilibrium be $(0, s_M, s_H)$. The Pareto dominant countersignaling equilibrium is that involving the least signaling, or $(0, s_M^*, 0)$ where s_M^* solves $\bar{q}_{\{L,H\}}(L) = M - c(s_M^*, L)$. Since $\bar{q}_{\{L,H\}}(L) > L$, it is clear that Lows are strictly better off.

It is less costly to deter Lows from imitating, so it follows that $s_M^* < s_M$ and therefore Mediums are also strictly better off. So if any type is not strictly better off it must be Highs, *i.e.*, $\bar{q}_{\{L,H\}}(H) \leq H - c(s_H, H)$. Since $(0, s_M, s_H)$ is the Pareto dominant signaling equilibrium, $M - c(s_M, M) = H - c(s_H, M)$. Since $s_M^* < s_M$, $M - c(s_M^*, M) > H - c(s_H, M)$. This implies that beliefs which satisfy the Intuitive Criterion must put probability zero on the event that a signal s in an open neighborhood of s_H was sent by a Medium. Similarly, Lows would never send such a signal. But since $\bar{q}_{\{L,H\}}(H) \leq H - c(s_H, H)$, Highs would have an incentive to deviate from the equilibrium with any signal $s \in N(s_H, \varepsilon)$ where $s < s_H$. Therefore $(0, s_M^*, 0)$ fails the Intuitive Criterion. Finally, receivers are indifferent between any equilibria of the model. ■

References

Akerlof, G. A. (1970), "The Market for 'Lemons:' Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, 84, 488–500.

Ball, S. B., M. H. Bazerman and J. S. Carroll (1991), "An Evaluation of Learning in the Bilateral Winner's Curse," *Organizational Behavior and Human Decision Processes*, 48, 1–22.

Banks, J. S. and J. Sobel (1987), "Equilibrium Selection in Signaling Games," *Econometrica*, 55, 647–662.

Bhattacharyya, N. (1998), "Good Managers Work More and Pay Less Dividends: A Screening Model of Dividend Policy," Mimeo.

Brown, S. and W. Hamilton (1996), "Autumn Colors: A Role for Aphids?" *Proceedings of XXth International Congress of Entomology*, 232.

Camerer, C. (1988), "Gifts as Economic Signals and Social Symbols," *American Journal of Sociology*, 94, S180–214.

Cho, I.-K. and D. M. Kreps (1987), "Signaling games and stable equilibria," *Quarterly Journal of Economics*, 102, 179–221.

Cho, I.-K. and J. Sobel (1990), "Strategic Stability and Uniqueness in Signaling Games," *Journal of Economic Theory*, 50, 381–413.

Cooper, D. J., S. Garvin and J. H. Kagel (1997a), "Adaptive Learning Vs. Equilibrium Refinements in an Entry Limit Pricing Game," *Economic Journal*, 107, 553–575.

Cooper, D. J., S. Garvin and J. H. Kagel (1997b), "Signalling and Adaptive Learning in an Entry Limit Pricing Game," *Rand Journal of Economics*, 28, 662–683.

Engers, M. (1987), "Signalling with Many Signals," *Econometrica*, 55, 663–674.

Fremling, G. M. and R. A. Posner (1999), "Market Signaling of Personal Characteristics," Working paper.

Garvin, S. and J. H. Kagel (1994), "Learning in Common–Value Auctions: Some Initial Observations," *Journal of Economic Behavior and Organization*, 25, 351–372.

Hertzendorf, M. N. (1993), "I'm not a High-Quality Firm – but I Play One on TV," *Rand Journal of Economics*, 24, 236–247.

Huck, S., H. T. Normann and J. Oechssler (1999), "Learning in Cournot Oligopoly: An Experiment," *Economic Journal*, 109, C80–95.

Hvide, H. K. (1999), "The Informational Role of Education in the Allocation of Talent," Norwegian School of Economics working paper.

Kagel, J. H. and D. Levin (1986), "The Winner's Curse and Public Information in Common Value Auctions," *American Economic Review*, 76(5), 894–920.

Muller, A. C. (1997), "Translation of *Analects of Confucius*," web document, `http://www.human.toyogakuen-u.ac.jp/~acmuller/contao/analects.htm`.

Nasar, S. (1998), *A Beautiful Mind: A Biography of John Forbes Nash, Jr.*, Simon and Schuster, New York.

Nelson, P. (1974), "Advertising as Information," *Journal of Political Economy*, 82, 729–754.

Quinzii, M. and J.-C. Rochet (1985), "Multidimensional Signalling," *Journal of Mathematical Economics*, 14, 261–284.

Riley, J. G. (1975), "Competitive Signalling," *Journal of Economic Theory*, 10(2), 174–186.

Ross, S. A. (1977), "The Determination of Financial Structure: The Incentive-Signalling Approach," *Bell Journal of Economics*, 8, 23–40.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara and S. Zamir (1991), "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review*, 81, 1068–1095.

Siegel, S. and N. J. Castellan, Jr. (1988), *Non-Parametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York.

Spence, A. M. (1973a), "Job Market Signaling," *Quarterly Journal of Economics*, 87, 355–374.

Spence, A. M. (1973b), "Time and Communication in Economics and Social Interaction," *Quarterly Journal of Economics*, 87, 651–660.

Spence, A. M. (1974), *Market Signaling*, Harvard University Press.

Teoh, S. H. and C. Y. Hwang (1991), "Nondisclosure and Adverse Disclosure as Signals of Firm Value," *Review of Financial Studies*, 4(2), 283–313.

Veblen, T. (1899), *The Theory of the Leisure Class*, Macmillan, New York.

Zahavi, A. (1975), "Mate Selection – a Selection for a Handicap," *Journal of Theoretical Biology*, 53, 205–214.

Department of Economics, University of Houston, Houston, TX 77204–5882.
E-mail address: nfelt@bayou.uh.edu

Claremont McKenna College and Claremont Graduate University, Claremont, CA 91711.
E-mail address: rharbaugh@claremontmckenna.edu

Bureau of Labor Statistics, Room 3105, 2 Massachusetts Ave., NE, Washington, DC 20212.
E-mail address: To_T@bls.gov