J Risk Uncertainty (2007) 35:1–16
DOI 10.1007/s11166-007-9014-9

Conspicuous conservatism in risk choice

Boaz Moselle · François Degeorge · Richard Zeckhauser

Published online: 11 July 2007
© Springer Science + Business Media, LLC 2007

Abstract We analyze the risk levels chosen by agents who have private information regarding their quality, and whose performance will be judged and rewarded by outsiders. Assume that risk choice is observable. Agents will choose risk strategically to enhance their expected reputations. We show that conspicuous conservatism results: agents of different qualities choose levels below those that would be chosen if quality were observable. This happens because bad agents must cloak their identity by choosing the same risk level as good agents, and good agents are more likely to distinguish themselves if they reduce the risk level. Our results contrast starkly with those for the case when risk choice cannot be observed.

Keywords Risk choice · Signaling · Conservatism

JEL Classification D81 · D82 · G30

Many economic activities secure both immediate returns and rewards to reputation. For example, a company that bolsters earnings gets not only those monies, but the reward of a higher price/earnings ratio and hence a higher stock price. Mutual funds that perform well get an inflow of funds. In some cases, a player's reputation reward will swamp any benefits from his immediate return. In the corporate context, there may be a "cash out" event, such as the sale of the whole company. For a division manager, a big promotion or

Part of this research has been carried out within the NCCR FINRISK, a research program supported by the Swiss National Science Foundation. Richard Zeckhauser gratefully acknowledges financial support from the National Science Foundation under grant number IIS-0428868, "Axioms and Algorithms for Reputation Systems." A referee, John Lindsey, and Jay Patel provided helpful comments.

B. Moselle
The Brattle Group, Brussels, Belgium

F. Degeorge
Swiss Finance Institute, University of Lugano, Lugano, Switzerland

R. Zeckhauser (✉)
Kennedy School of Government, Harvard University, 79 Kennedy St, Cambridge, MA 02138, USA
e-mail: richard_zeckhauser@harvard.edu

a strong outside offer would play much the same role. This paper addresses situations where reputations are a heavily weighted consideration.

Asymmetric information drives our analysis. We assume that each player, hereafter agent, knows his quality, but others do not. His reputation will be inferred from his risk choice and the payoff that results. Our central question in this paper is how much risk good and bad agents will choose. Given that there are significant reputation rewards, we argue that even risk-neutral players will choose their risk level strategically. They will sacrifice some expected direct payoff in the hope of burnishing their reputations.

We look at situations where players' risk choices are observable. Transactions in which corporations are sold well illustrate this case. Financial markets can see whether a retail chain attempted to expand rapidly—swifter expansion entails higher risk—or how diversified a Real Estate Investment Trust (REIT) is across metropolitan areas.

The sequence of moves is as follows: 1. The agent chooses his risk level, and that is observable to all. 2. Both the agent and outsiders learn the outcome. 3. The outsiders then contract with the agent based on the inference they draw from what they observe. That is, outsiders update their assessment of the agent using two signals: (1) the lottery outcome, and (2) the level of risk chosen, which serves as a "signal" in the Spence (1974) sense. We show that: First, agents will often pool on their risk choice. Second, when they do, all agents, regardless of type, are induced into conspicuous conservatism. That is, all agents will choose a level of risk below the one that would maximize their expected direct payoff. That is because good agents are in the driver's seat; low risk levels help them distinguish themselves. Bad agents must choose low risk levels as well, lest they reveal their type.

We proceed as follows. "The model" presents the model. "Results" develops our results. "Conclusion" concludes.

1 The model

We assume that agents have private information regarding their quality. Agents care about the lottery outcome because they stand to gain from the reputation they acquire among outsiders, based in part on the lottery that they choose, and in part on the lottery outcome.

Agents choose a level of risk for their activity. We assume the risk choice to be continuous. Formally, each agent chooses from a one-dimensional family of random variables indexed by its variance, V, where $\underline{V} < V < \overline{V}$.

The choice of V is common knowledge. That is, outsiders who are drawing inferences about the agent's type, see V, and update accordingly. By contrast, when the risk choices by agents are unobservable, the situation considered earlier by Degeorge et al. (2004), hereafter DMZ, the results are quite different. The prime concepts in this paper, such as signaling strategies, and pooling and separating equilibria, do not apply. To facilitate contrasts, we follow the features of the DMZ model, with one crucial difference: In our setting risk choice is observable.

The random variable $\widetilde{x}_V$ represents first-period performance (say, the test score, or company earnings). It is distributed normally with variance V and mean $\mu(V, \theta) = \mu(V) + \theta$, where θ indexes the agent's type. We will refer to the θ=0

type agent as bad and the $\theta=\Delta$ type as good, where $\Delta>0$. The prior probability that an agent is of the good type, denoted by p, is common knowledge, as is the mean-variance schedule $\mu(V)$.

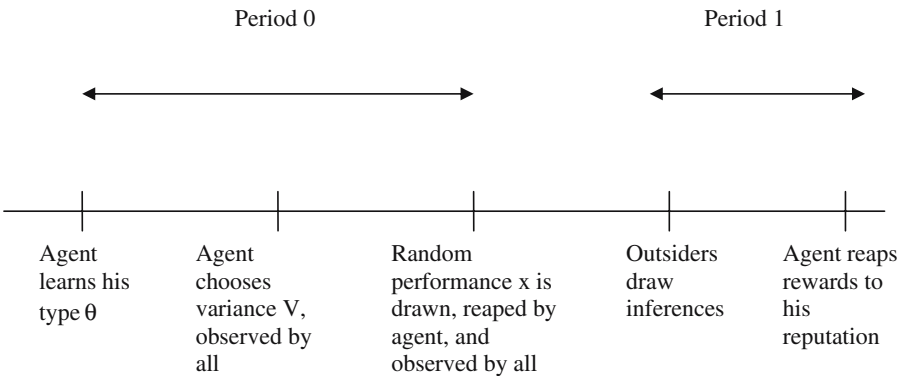
Each agent chooses a point on the mean-variance schedule given by $\mu(V)$ and for given V an agent of type θ has performance

$$\tilde{x} \sim N(\mu(V) + \theta, V).$$

We posit that $\mu(V)$ is single-peaked, with its maximum at internal point V^*, and concave, i.e., the marginal benefit of adding variance decreases throughout, turning negative beyond V^*. In fact, the mere existence of an interior maximum is sufficient for most of our theoretical results.[1] The existence of such an interior maximum follows naturally from the usual assumption that agents have only a finite amount of favorable lotteries.[2] Our theoretical results, apart from Claim 3, do not assume concavity. Single-peakedness of $\mu(V)$ is assumed for the proof of Claim 1 and Claim 3.[3]

It is convenient (though by no means essential) to assume that $\mu(V) \to -\infty$ *as* $V \to \underline{V}$ *or* $V \to \overline{V}$, and we maintain this assumption throughout.

At the beginning of period 0, an agent learns his type, and then chooses his desired variance V. His random performance x is then drawn according to the equation above, reaped by him, and observed by all, and the period ends. At the beginning of period 1, outsiders draw inferences, and the agent reaps the rewards to his reputation.



[1] It is important that $\mu'(0) > 0$. Otherwise, good types would merely choose zero risk, and bad types would have to follow. This condition at 0 is also plausible. Making money in this world inevitably requires risk taking.

[2] Consider for example a company choosing investment projects. A special case occurs when the projects are fixed in size and independent, with some having positive expected returns, others negative returns, and the manager has no budget constraint. In this case $\mu(V)$ will be single-peaked and concave, implying that the manager will allocate funds to every project that offers a positive expected return. Doing so will yield the peak of the $\mu(V)$ curve. If the manager wishes to increase variance, he will invest in the project with the lowest absolute value of the mean-to-variance ratio. Once this opportunity to add variance is exhausted (by the finite size assumption) the manager picks the second lowest mean-to-variance ratio project and the process continues.

[3] In fact, a weaker condition than global single-peakedness of $\mu(V)$ is sufficient for these results to obtain (details are available from the authors).

All agents are assumed to be risk-neutral, so that in a full-information setting agents would choose the mean-maximizing level of variance V^*.[4] We assume risk neutrality to facilitate exposition and because we are confident the same qualitative results would apply—namely agents tilting toward conservatism in risk choice—if agents were risk averse, as is normally assumed. We investigate how asymmetric information leads to departures from this optimum. Choosing a variance less than V^* represents risk-reducing behavior; choosing a variance greater than V^* would increase risk.

We can formulate the last step in the game as the sale by the agent of his capital—human, organizational or intellectual—to a new long-term owner. It will clearly be optimal for the new owner to choose the level of variance V^* that maximizes expected performance (we assume risk-neutrality throughout), so the expected return to an agent of type θ from next period onward is $\sum_{t=1}^{\infty} \frac{\mu(V^*)+\theta}{(1+r)^t} = \frac{\mu(V^*)+\theta}{r}$ where r is the discount rate. Thus, the expected present value (performance plus expected price) of an agent with performance x this period is

$$x + \frac{\mu(V^*)}{r} + \frac{1}{r} E[\theta \,|\text{all available information}],$$

where buyers use Bayesian analysis to compute the expectation term.

Agents choose V to maximize their expected total payoff, which is given by their present value, namely

$$\Pi = \max_V E\left[\widetilde{x} + \frac{\mu(V^*)}{r} + \frac{1}{r} E\left[\theta \,\middle|\, all\ available\ \ information\right] \middle| choice\ of\ V\right].$$

If θ were observable, this problem would be trivial; we would have

$$E[\theta \,|\, all\ available\ \ information] = \theta,$$

and agents would always set $V=V^*$. We consider the case where θ is private information, so the agent is faced with a trade-off in maximizing its expected total payoff: clearly, setting $V=V^*$ maximizes the expected performance; however, deviating from V^* will change the information that flows to outsiders, and has the potential to increase the expected reputation.

2 Results

All proofs are gathered in the Appendix.

In a standard signaling game, e.g., Cho and Kreps (1987), outsiders draw inferences from an agent's choice, e.g., whether he goes to college. Our model adds one stage to a standard signaling game; namely a performance signal x that is produced by a probabilistic process, and that is revealed to all. The agent receives the direct payment x, plus their expected price from an outsider. Total payment is given by the random variable

$$\widetilde{x} + \frac{\mu(V^*)}{r} + \frac{1}{r} E[\theta | V, \widetilde{x}].$$

[4] For a state-of-the art exposition of decision theory under risk aversion, see Gollier (2004a).

We consider only pure strategy equilibria. When both types of agents choose the same *V,* a *pooling equilibrium* results. When agents choose different *V*'s, a *separating equilibrium* emerges. Even in a pooling equilibrium, the current performance is informative about each agent's type. Thus our model involves both signaling and signal-jamming. The choice of variance is the signal in the Spence (1974) sense. The lottery outcome is the performance signal. Its degree of informativeness can be attenuated through a choice of high variance. When risk choice is observable, however, agents cannot surreptitiously manipulate variance, and we shall see that bad agents have a strong incentive to mimic the good agents' choices of risk.

We posit no exogenous difference in costs between types: such differences are the factor that drives ordinary signaling models.[5] Instead, we find that an *endogenous* reputational difference emerges between types. That is, *the good type has less to gain by deviating from the mean-maximizing variance choice than does the bad type,* because, as the good type wishes, his ability will be revealed partly through his performance.

Outsider beliefs play an important role here, as they do in standard signaling games. First, after observing an agent's choice of variance *V*, outsiders update their prior probability *p* that the agent is good to a posterior ξ(p, V) via Bayesian updating (where possible). Second, after observing the agent's performance *x*, outsiders then further update ξ, to a new posterior $\widehat{\xi}(\xi, x)$, using Bayesian updating.

So an agent of type θ=0 or Δ who chooses variance *V* expects a total payoff of

$$\Pi_\theta(V) = \mu(V) + \theta + \frac{1}{r}\left(\mu\left(V^*\right) + \Delta E\left[\widehat{\xi}(\xi(p, V), \widetilde{x})\middle|\theta, V\right]\right).$$

2.1 Pooling equilibria

Claim 1 A necessary and sufficient condition for $\widehat{V}$ to represent a pooling equilibrium is that

$$\frac{\Delta}{r} E\left[\widehat{\xi}(p, x)|0, \widehat{V}\right] \geq \mu(V^*) - \mu\left(\widehat{V}\right)$$

The set of $\widehat{V}$ values leading to a pooling equilibrium includes V and forms a union of closed intervals.*

The left hand side (LHS) of the inequality represents the gain to a bad agent of pooling at $\widehat{V}$ instead of choosing its mean-maximizing variance, *V**, and thereby admitting up front to being bad. The right hand side (RHS) is the cost, in foregone performance, of choosing $\widehat{V}$ rather than the mean-maximizing value. For $\widehat{V}$ to be an equilibrium it must be the case that both types value the returns to pooling more than the cost of pooling, i.e. prefer pooling. It is easy to see that if the bad type prefers $\widehat{V}$ to *V**, so does the good type, since the good type has a higher expected reputation

[5] Cheap talk models (Crawford and Sobel 1982) assume costless signaling. However, they apply to situations in which parties have partially aligned interests. In our setting, the interests of the signalers and the observers diverge.

from pooling but pays the same price for deviating from V^*. If $\widehat{V}$ is equal to V^*, then the right hand side of the inequality is equal to zero while the left hand side is always non-negative, so V^* itself is always a pooling equilibrium.

Welfare and efficiency properties of pooling equilibria Relative to the full information case, signaling through one's choice of risk reduces aggregate welfare. The manipulation of reputations is a negative-sum game among bad and good agents. Consider the "expected reputation"—defined formally below—over the two types of agents. This expectation must equal the prior reputation. Therefore any improvement in the expected reputation of one type of agent will exactly cancel out the corresponding deterioration for the other. However, since agents depart from the performance-maximizing variance V^* in order to enhance their reputations, value in direct payoffs is sacrificed, and the net efficiency effect is negative.[6]

Claim 2 formalizes the comparison of equilibria. Informally, good agents will prefer to pool at lower variance because it enables them to better distinguish themselves from bad agents. Bad agents will prefer to pool at high variance, since noisier performance makes it more likely that bad agents will produce performance common to good agents. Using this result, we show that in general there exists a continuum of Pareto-unranked pooling equilibria, although some such equilibria are Pareto-ranked.[7] In particular, so long as it is increasingly costly to add risk, there will always be a pooling equilibrium at a variance higher than V^*, which is Pareto-dominated by pooling at V^*.

Definition Given a pooling equilibrium $\left(\widehat{V}, \widehat{V}\right)$, the expected reputation in this equilibrium of an agent of type θ is defined as:

$$\begin{aligned} ER\left(\mu, \widehat{V}, \theta\right) &= \Delta E\left[\widehat{\xi}(p, \widetilde{x})\middle|\theta, \widehat{V}\right] \\ &= \int_{-\infty}^{+\infty} \Delta\widehat{\xi}\left(p, x|\mu, \widehat{V}\right) f\left(x|\mu + \theta, \widehat{V}\right) dx. \end{aligned}$$

Thus $ER\left(\mu, \widehat{V}, \theta\right)$ gives the Bayesian estimate of the expected value of θ, conditional on observing the draw, if the agent is in fact of type θ. Here $\widehat{\xi}\left(p, x|\mu, \widehat{V}\right)$ is the posterior probability that the agent is good when p was the prior probability and x was observed, while $f\left(\cdot|\mu, \widehat{V}\right)$ is the density function associated with a normally distributed random variable with mean μ and variance $\widehat{V}$.

Claim 2 Given two pooling equilibria $\left(\widehat{V}_1, \widehat{V}_1\right)$ *and* $\left(\widehat{V}_2, \widehat{V}_2\right)$, if $\widehat{V}_1 < \widehat{V}_2$ then good agents have a higher expected reputation under $\widehat{V}_1$ than under $\widehat{V}_2$ (so bad agents have a lower expected reputation).

[6] If, as before, we write $\Pi_B\left(\widehat{V}\right), \Pi_G\left(\widehat{V}\right)$ for the payoffs of bad and good agents respectively in a pooling equilibrium $\left(\widehat{V}, \widehat{V}\right)$, then the expected total payoff over all agents (normalizing the number of agents to one) satisfies $pE\,\Pi_G\left(\widehat{V}\right) + (1-p)E\,\Pi_B\left(\widehat{V}\right) = \mu\left(\widehat{V}\right) + \frac{1}{r}\mu(V^*) + \left(1 + \frac{1}{r}\right)p\Delta$.

[7] Equilibria are said to be "Pareto-ranked" if one Pareto-dominates the other.

In other words, as the signal becomes noisier, bad agents are better able to hide their type, and good agents less able to reveal theirs. We can make a number of general observations about the payoff profiles generated by different pooling equilibria. Figure 1, which is drawn for the case where good agents comprise one half of the population, illustrates.

Claim 3

(1) For ε small and positive, the pooling equilibrium $\left(V^* + \varepsilon, V^* + \varepsilon\right)$ gives the good agent a strictly lower payoff than does the pooling equilibrium $\left(V^*, V^*\right)$ (point B on Figure 1), and the bad agent a strictly higher payoff. The opposite is true for pooling at$\left(V^* - \varepsilon, V^* - \varepsilon\right)$. In particular, this means that there is a

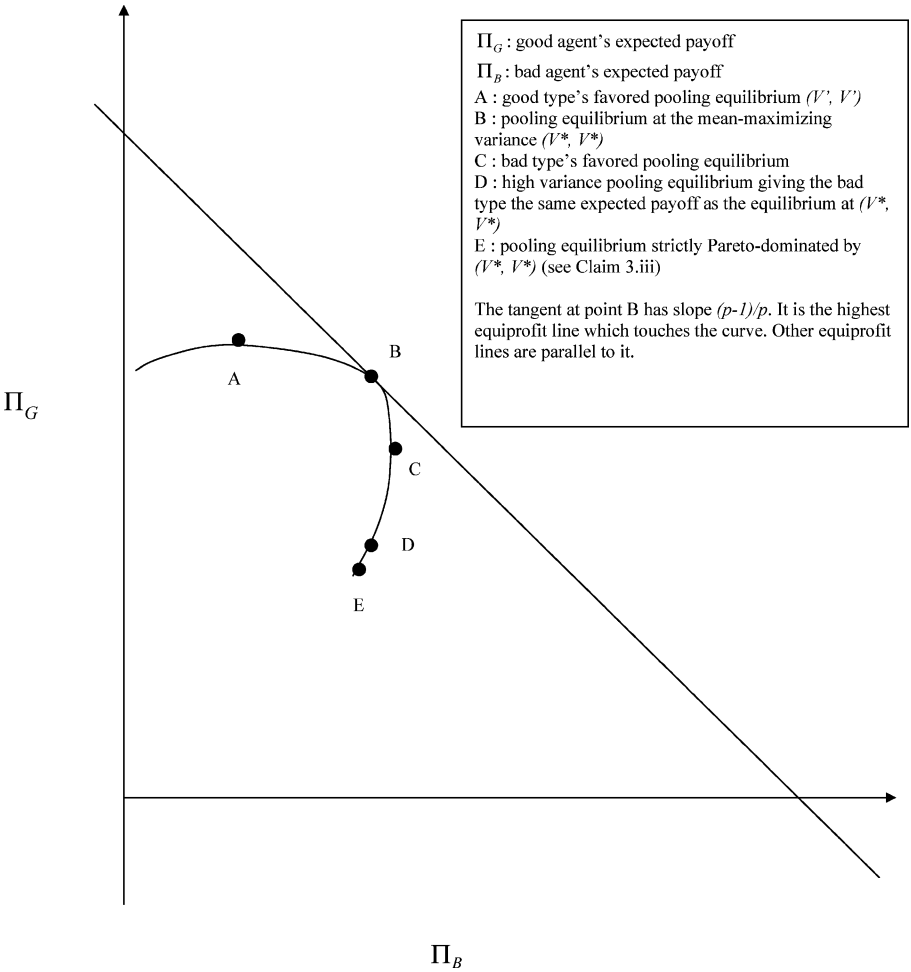


Fig. 1 Welfare and efficiency properties of pooling equilibria

continuum of Pareto-unranked pooling equilibria (the portion of the curve in Figure 1 between points A and C).

(2) If $V^* < \widehat{V}_1 < \widehat{V}_2$, then the good agent strictly prefers $\widehat{V}_1$ to $\widehat{V}_2$. Whenever $\widehat{V}_1 < \widehat{V}_2$ a good agent always gains more, or loses less, than a bad agent from a move to the lower variance equilibrium:

$$\Pi_G\left(\widehat{V}_1\right) - \Pi_G\left(\widehat{V}_2\right) > \Pi_B\left(\widehat{V}_1\right) - \Pi_B\left(\widehat{V}_2\right)$$

or, equivalently, $d\Pi_G \Big/ d\widehat{V} < d\Pi_B \Big/ d\widehat{V}$.

(3) If the marginal cost of taking on variance is increasing ($\mu(v)$ is concave), then there are always pooling equilibria that are weakly (point D on Figure 1) and strongly Pareto-dominated (point E) by pooling at (V^*,V^*) (point B).

(4) There exists $V' < V^*$ such that (V', V') is a pooling equilibrium (point A on Figure 1), and gives the highest payoff to good agents out of all possible equilibria.

2.1.1 Three observations

1. The expected payoff in this game consists of two portions: the immediate payoff and the reputation. Averaged across types the posterior reputation has to be the same as the prior, so that V only affects the expected direct payoff averaged across types, which is maximized at V^*. Hence pooling at V^* maximizes the aggregate expected payoff of all agents. That is, of any pooling equilibrium, it gives the highest value of $p\Pi_G + (1-p)\Pi_B$, so the tangent to the graph of payoff profiles at point B in Figure 1 is given by the equi-payoff line through B, whose equation is $p\Pi_G + (1-p)\Pi_B = (1+r^{-1})\left(\mu\left(V^*\right) + p\Delta\right)$. Other equi-payoff lines are parallel to this one.
2. Figure 1 shows a region of Pareto-ranked low-variance equilibria to the left of A on the curve. However, there is no simple general condition on $\mu(V)$ that guarantees its existence; that is, for low variance equilibria there is no analog to part (3) of Claim 3.
3. The slope of the curve in Figure 1 is given by $d\Pi_G/d\Pi_B = \Pi_G^{'}(V) \Big/ \Pi_B^{'}(V)$, so from part (2) of Claim 3 we can deduce that the curve is flatter than 45° to the left of A, and steeper than 45° to the left and below C.

2.1.2 Equilibrium selection

As is common in signaling models, we have a continuum of equilibria. Alas, standard equilibrium refinements are not effective here. Focal point theory (Schelling 1960) may help the good agents coordinate on a beneficial outcome. Pareto optimality is a salient property, which suggests some location on the frontier between A and C. Point (V', V'), here represented by point A, stands out. It is readily recognizable as the best equilibrium for the good agents. Moreover the cost of deviating from it to a higher variance is less for bad agents than for good, suggesting that deviators will be branded as bad. Given the prominence of (V', V')

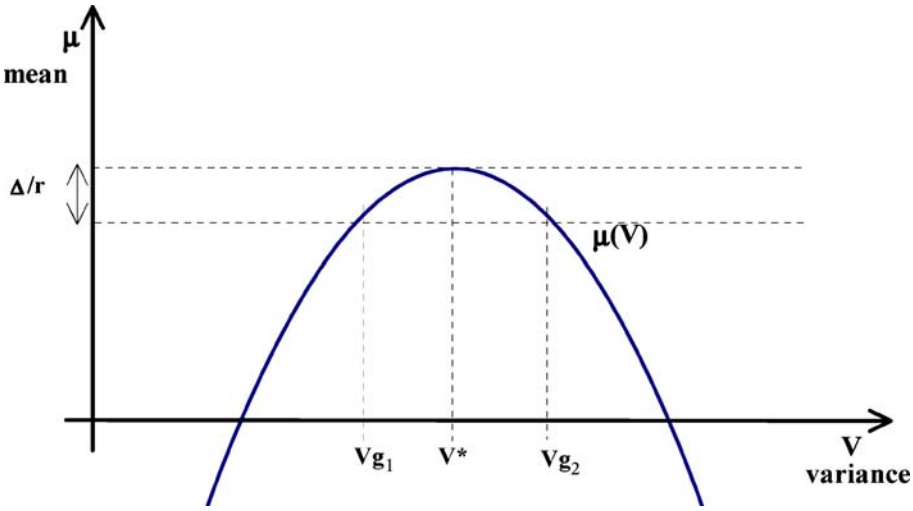


Fig. 2 Separating equilibria

for good agents, they could readily coordinate in choosing it, whether on a tacit or open basis, and bad agents would have no place else to go.[8]

2.2 Separating equilibria

The driving force behind separating equilibria in signaling models is cost differences across types. But our model has no such differences. Although two separating equilibria exist—in both the bad type chooses V*—neither is consistent with intuition or the Banks and Sobel (1987) "divinity" equilibrium refinement, and we rule these equilibria out on these grounds.

Claim 4 This game has exactly two separating equilibria in pure strategies, in which the good type chooses V_g and the bad type chooses V*(see Figure 2). In one separating equilibrium, the good type chooses $V_{g1} < V^*$; in the other, $V_{g2} > V^*$. In each of these equilibria, both types are indifferent between choosing V_g and choosing V* (equivalently, for both types, the incentive compatibility constraint binds). Neither equilibrium is consistent with the Banks–Sobel "divinity" equilibrium refinement (extended to this model in a natural way).

These separating equilibria are intuitively unappealing. First, the good type is indifferent between its own strategy and that of the bad type, even though the market assigns probability 1 to its being of the bad type if it chooses V*. If there were any probability that the market expected the agents to pool at V*, then a deviation to V*

[8] Formally, we are considering a one-play game. Deviations from an equilibrium, however, implicitly invoke repeated play, with deviators accepting a current loss hoping to get to a better equilibrium. At (V', V'), good agents are already at their preferred equilibrium. Bad agents, by contrast, have an incentive to jostle the (V', V') equilibrium, since they prefer any other Pareto-optimal equilibrium. This reinforces the expectation that deviators will be branded as bad, which makes the (V', V') equilibrium more secure.

would be (in an unformalized sense) a dominant strategy for the good type. We therefore rule out these separating equilibria.[9]

2.3 Discussion

Our results suggest that when the level of risk chosen by agents is conspicuous to outsiders, and thus functions as a signal in the Spence (1974) sense, agents will often pool at levels of risk below the performance-maximizing level, since good agents prefer that performance reveal true quality. For a bad agent, increasing noise in the hope of clouding the picture is a pointless venture. That very choice would admit to his low quality.

Strategies that reduce the noise of a signal are well known to induce significant amounts of pooling in various contexts. Thus, students applying to elite colleges that do not require the SAT know that not taking it conveys negative information about their type.

Our conspicuous conservatism argument might also explain the strong resistance of some institutions to change. Consider the French higher education system marked by "grandes écoles," the springboards for most of the French elite. The system was essentially invented in Napoleon's time and it can be argued that it has evolved little since, despite cogent criticisms of its operation. What could account for such stability? In order to enter a "grande école" students must pass a series of examinations that provide a rather precise assessment of ability. Any alternative educational system—such as one that would put less emphasis on testing students upfront and more emphasis on student learning, for example—faces huge hurdles in establishing itself. Students opting for such an alternative system would automatically be labeled as bad. In fact, many students with little chance of entering a "grande école" try anyway, in the small hope of securing the good label.

For bad types, conservatism pays if it is conspicuous. In contrast, given that the agent's choice of risk is unobservable in the situation studied by DMZ, it cannot function there as a Spence-type signal. Outsiders only observe the outcome x of the lottery. This prior work shows that in such contexts, good agents choose low levels of risk, and bad agents choose high levels—provided outsiders have no strong priors about whether agents are good or bad. Good agents are seeking to reduce noise so as to stand out. Bad agents are seeking to increase noise in the hope of producing the results of good agents. For example, a strong student might choose a low-risk strategy, e.g., avoiding guesses on a multiple-choice test that deducts for wrong answers, so as to reduce noise and maximize information flow. A weak student might choose otherwise. A somewhat similar intuition emerges from the work of Tsetlin, Gaba and Winkler (2004). They analyze the strategic choice of risk in multiround contests, and contests with handicaps (but without private information).

[9] How robust are these results to our assumption that Δ, the difference in expected return between good and bad types, is constant across risk levels? If the difference were multiplicative instead of additive, our results would be reinforced; imitation of the good type by the bad would become less costly. In contrast, if the gap in expected performance between types increased as they reduced their variance from V^*, separation would become possible. The good type could go where the bad could not afford to follow.

They find that contestants in a weak position (e.g., low mean, high handicap, or low previous performance in a multiround contest) should maximize risk, and those in a strong position should minimize it.[10] In this vein, Chevalier and Ellison (1997) document that mutual fund managers with poor January-September performance increase the risk of their investment strategy in the fourth quarter, while managers with strong year-to-date performance reduce it. At a cost, investors have the potential to assess the risk associated with a manager's portfolio choice, since portfolios are announced periodically. Since poor first-half performers select distinctively high risk levels, presumably not many investors accept the cost to do such monitoring. Were rating firms like Morningstar to make risk assessments of mutual funds widely and cheaply available, gambles to "catch up" in the second half of the year would become less common.

3 Conclusion

We analyzed the levels of risk good and bad agents take on when they know their quality but outsiders do not. If the agents' risk choices are observed by outsiders, then, invoking reasonable criteria about market beliefs, a single pooling equilibrium is likely to emerge. Good types set the standard, so they select the equilibrium that is most favorable for them. At it, both good and bad types will choose a risk level below the one that maximizes their respective expected performance. We conclude that when risk choice is observable, agents with private information on their quality face strong incentives—regardless of their quality—to pick performance lotteries with low risk. When risk choice is conspicuous, conservatism helps good types separate themselves. Bad types will not like the equilibrium. Nevertheless, they will choose conservatism because that is better than maximizing their expected outcome but admitting their type.

Appendix

Proof of Claim 1 Suppose that $\widehat{V}$ gives such a pooling equilibrium. It is easy to see that if any out-of-equilibrium market beliefs support this equilibrium, then the extreme beliefs that set $\xi(V)=0$ for all V not equal to $\widehat{V}$ also support it, so let us suppose that the latter are the market's beliefs. In that case, the most attractive deviation from $\widehat{V}$ must be V^*. Therefore, necessary and sufficient conditions for incentive compatibility are that for both types we have

$$\mu\left(\widehat{V}\right)+\theta+\frac{1}{r}\left(\mu\left(V^*\right)+\Delta E\left[\widehat{\xi}(p,x)\middle|\theta,\widehat{V}\right]\right)\geq\mu\left(V^*\right)+\theta+\frac{1}{r}\left(\mu\left(V^*\right)+0\right)$$
$$\Leftrightarrow\Delta E\left[\widehat{\xi}(p,x)\middle|\theta,\widehat{V}\right]\geq r\left(\mu\left(V^*\right)-\mu\left(\widehat{V}\right)\right).$$

[10] Gollier (2004b) studies risk-choice externalities of a quite different sort, namely when an individual's marginal utility changes with the average consumption of others. Assume a positive relationship, the plausible case. Then the individual will be less tolerant of personal risks and more tolerant of market risks and collective gambles than he would be in the absence of such externalities.

Clearly the left side of the last inequality is larger for the good type than for the bad type, implying that if the bad type does not want to deviate then neither does the good. It follows that this inequality with θ set equal to 0 provides a necessary and sufficient condition for $\widehat{V}$ to be an equilibrium, as stated in the proposition.[11]

It is also clear that the left hand side is bounded while the right hand side goes to infinity as V approaches its upper or lower bounds, and that at $\widehat{V} = V^*$ the inequality is satisfied (with strict inequality). Q.E.D.

Proof of Claim 2 We need to show that, when comparing different pooling equilibria, the expected reputation for good agents is decreasing in the level of variance $\sigma^2 = V$. That is,

$$\frac{dER}{d\sigma}\left(\mu, \sigma^2, \Delta\right) < 0.$$

We will actually prove that

$$\frac{dER}{d\sigma}\left(\mu, \sigma^2, 0\right) > 0.$$

The former result for good agents follows immediately from the latter.

Recall that we defined the expected reputation to be

$$ER\left(\mu, \sigma^2, \theta\right) = \int_{-\infty}^{+\infty} \Delta\, \widehat{\xi}\left(p, x \middle| \mu, \sigma^2\right) f\left(x \middle| \mu + \theta, \sigma^2\right) dx.$$

This expression is independent of μ, since adding the same constant to both agents' performance cannot affect their expected reputation. Formally, one just notes that the integrand is a function of x-μ, so substituting $y = x - m$ gives the result. It follows that

$$\frac{\partial ER}{\partial \mu} = 0,$$

and

$$\frac{dER}{d\sigma} = \frac{\partial ER}{\partial \mu}\frac{d\mu}{d\sigma} + \frac{\partial ER}{\partial \sigma} = \frac{\partial ER}{\partial \sigma},$$

so we need only prove the result for the partial derivatives. In order to do so, note that the Bayesian posterior probability distribution is given by

$$\widehat{\xi}\left(p, x \middle| \mu, \sigma^2\right) = \left[1 + \frac{1-p}{p} e^{-\Delta\left((x-\mu)/\sigma^2 - \Delta/2\sigma^2\right)}\right]^{-1},$$

[11] Notice that a Cho-Kreps "intuitive"-like criterion would not rule out these extreme beliefs. For suppose there are any unspecified beliefs that would make the good type deviate to some V. Then certainly the beliefs that attach probability 1 that any agent that plays V is good would make the bad type play V (because these beliefs would make the good type play V, and would do more for the bad type than the good type). Thus there is no "intuitive" reason to rule out such extreme beliefs.

so we can write

$$ER(\mu,\sigma^2,0) = \frac{\Delta}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[1 + \frac{1-p}{p} e^{-\Delta\left((x-\mu)/\sigma^2 - \Delta/2\sigma^2\right)}\right]^{-1} e^{-(x-\mu)^2/2\sigma^2} dx.$$

Setting $y = \frac{x-\mu}{\sigma}$ yields

$$ER(\mu,\sigma^2,0) = \frac{\Delta}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[1 + \frac{1-p}{p} e^{-\Delta\left(y/\sigma - \Delta/2\sigma^2\right)}\right]^{-1} e^{-y^2/2} dy,$$

and so

$$\frac{\partial ER}{\partial \sigma}(\mu,\sigma^2,0) = \frac{\Delta^2}{\sigma^2\sqrt{2\pi}} \left(\frac{1-p}{p}\right) \int_{-\infty}^{\infty} \left[1 + \frac{1-p}{p} e^{-\Delta\left(y/\sigma - \Delta/2\sigma^2\right)}\right]^{-2} \left(\frac{\Delta}{\sigma} - y\right) e^{-\Delta\left(y/\sigma - \Delta/2\sigma^2\right)^{-y^2/2}} dy.$$

We now make a further substitution, setting $z = y + \frac{\Delta}{\sigma}$ to give

$$\begin{gathered} \tfrac{\partial ER}{\partial \sigma}(\mu,\sigma^2,0) = -A \textstyle\int_{-\infty}^{\infty} \eta(z)^{-2}\left(z - \frac{2\Delta}{\sigma}\right) e^{-z^2/2} dz, \\ where\, A = \tfrac{\Delta^2}{\sigma^2\sqrt{2\pi}}\left(\tfrac{1-p}{p}\right) e^{\Delta^2/\sigma^2} > 0 \\ and\, \eta(z) = 1 + \tfrac{1-p}{p} e^{-\Delta\left(z/\sigma - 3\Delta/2\sigma^2\right)}. \end{gathered}$$

We break this up as the difference of integrals over $[2\Delta/\sigma, \infty]$and over $[-\infty, 2\Delta/\sigma]$, as follows:

$$\frac{\partial ER}{\partial \sigma}(\mu,\sigma^2,0) = -A\left(\int_{2\Delta/\sigma}^{\infty} \eta(z)^{-2}\left(z - \frac{2\Delta}{\sigma}\right) e^{-z^2/2} dz + \int_{-\infty}^{2\Delta/\sigma} \eta(z)^{-2}\left(z - \frac{2\Delta}{\sigma}\right) e^{-z^2/2} dz\right).$$

Using the transformation $z' = 4\Delta/\sigma - z$ turns the latter into an integral over $[2\Delta/\sigma, \infty]$, so we can combine the two to get

$$\begin{aligned} \tfrac{\partial ER}{\partial \sigma}(\mu,\sigma^2,0) &= -A \int_{2\Delta/\sigma}^{\infty} \left(z - \tfrac{2\Delta}{\sigma}\right) e^{-z^2/2} \left[\eta(z)^{-2} - \eta(4\Delta/\sigma - z)^{-2} e^{4\Delta z/\sigma - 8\Delta^2/\sigma^2}\right] dz \\ &= -A \int_{2\Delta/\sigma}^{\infty} \left(z - \tfrac{2\Delta}{\sigma}\right) e^{-z^2/2} \eta(z)^{-2} \eta(4\Delta/\sigma - z)^{-2} \\ &\qquad \left[\eta(4\Delta/\sigma - z)^2 - \eta(z)^2 e^{4\Delta z/\sigma - 8\Delta^2/\sigma^2}\right] dz. \end{aligned}$$

To show that the above expression is positive, it is therefore sufficient to show that the integrand is negative, or equivalently, that

$$\eta(4\Delta/\sigma - z)^2 - \eta(z)^2 e^{4\Delta z/\sigma - 8\Delta^2/\sigma^2} < 0.$$

This expression can be factored as

$$-(t^2-1)(1+2ut+t^2),\ where$$
$$t=e^{\Delta\left(z/\sigma-2\Delta/\sigma^2\right)}>1,\ and$$
$$u=\tfrac{1-p}{p}e^{-\Delta^2/2\sigma^2}>0.$$

Hence the expression is negative. Thus we have shown that

$$\frac{dER}{d\sigma}\left(\mu,\sigma^2,0\right)>0.$$

Since we must have

$$pER\left(\mu,\sigma^2,\Delta\right)+(1-p)ER\left(\mu,\sigma^2,0\right)=p\Delta,$$

it must be that

$$\frac{dER}{d\sigma}\left(\mu,\sigma^2,\Delta\right)<0.\quad Q.E.D$$

Proof of Claim 3

(1) As long as μ(V) is differentiable, the cost in terms of expected revenue of moving ε away from V^* is initially of second order in ε, while the change in expected reputation is of the first order.

(2) Moving toward V^* from the right gives the good type both a higher expected performance and a better expected reputation, and hence increases its total expected payoff. For the more general result, note that moving to a lower V has the same effect on the expected performance of both types of agents, but increases the expected reputation of the good type while it hurts the expected reputation of the bad type.

(3) Consider the payoff $\Pi_B(V)$to the bad type from both types' pooling at a value V, ignoring for the moment the question of whether or not V actually gives an equilibrium (that is, whether or not pooling at V is incentive-compatible). We have just shown in the proof of part (1) that $\Pi'_B>0$ at V^*. However, if we assume that the marginal cost, in expected revenue forgone, increases as we raise V higher and higher, then we will eventually have $\Pi'_B<0$ and reach a point $\widehat{V}$ with $\Pi_B\left(\widehat{V}\right)=\Pi_B\left(V^*\right)$ (point D in Figure 1). Pooling at this value is an equilibrium. To show that this is so, we need only check that the incentive-compatibility condition of Claim 1 holds. By definition of $\widehat{V}$ we have

$$\mu\left(\widehat{V}\right)+\frac{1}{r}\left(\mu\left(V^*\right)+\Delta.E\left[\widehat{\xi}(p,x)\middle|0,\widehat{V}\right]\right)$$
$$=\mu\left(V^*\right)+\frac{1}{r}\left(\mu\left(V^*\right)+\Delta.E\left[\widehat{\xi}(p,x)\middle|0,V^*\right]\right).$$

So

$$\frac{\Delta}{r}E\left[\widehat{\xi}(p,x)\middle|0,\widehat{v}\right] = \mu\left(V^*\right) - \mu\left(\widehat{V}\right) + \frac{\Delta}{r}E\left[\widehat{\xi}(p,x)\middle|0,V^*\right] > \mu\left(V^*\right) - \mu\left(\widehat{V}\right).$$

In other words, since the bad type is indifferent between pooling at $\widehat{V}$ and pooling at V^*, it will strictly prefer pooling at $\widehat{V}$ to deviating to V^*, as the latter is strictly worse than pooling at V^* (since deviating to V^* will lead to an expected reputation of 0).

Since the incentive-compatibility condition holds with strict inequality at $\widehat{V}$, an equilibrium also exists at $\widehat{V}+\varepsilon$. This equilibrium will be worse for the bad type than pooling at $\widehat{V}$,[12] hence worse than pooling at V^*. We know by part (2) that $\widehat{V}$ is also worse for the good type than V^*, so it is strictly dominated by V^* as claimed. Q.E.D.

(4) Since the set of pooling equilibria is closed (by Claim 1) and those equilibria with $V>V^*$ give the good type a lower payoff than V^*, there must be a pooling value V' (not necessarily unique) that maximizes the payoff to good agents over all pooling equilibria.

Proof of Claim 4 Suppose we have a separating equilibrium, which we can denote (V_b,V_g), where V_b and V_g are the equilibrium choices of the bad and good types respectively. If $V_b \neq V^*$, then deviating to V^* improves the bad agent's expected performance and cannot damage its "equilibrium reputation" $\Delta E[\theta|V_b,x]$, since separation means that the agent's type is fully revealed in equilibrium (so a bad agent has an equilibrium reputation of zero). Thus we must have $V_b=V^*$. Then the condition for the bad type requires that

$$\mu\left(V^*\right) + \frac{1}{r}\left(\mu\left(V^*\right)+0\right) \geq \mu\left(V_g\right) + \frac{1}{r}\left(\mu\left(V^*\right)+\Delta\right),$$

(since the right side gives the payoff to pretending to be good). For the good type we have

$$\mu\left(V_g\right) + d + \frac{1}{r}\left(\mu\left(V^*\right)+\Delta\right) \geq \mu\left(V^*\right)\Delta + \frac{1}{r}\left(\mu\left(V^*\right)+0\right).$$

Combining these produces

$$\mu(V^*) = \mu\left(V_g\right) + \frac{\Delta}{r}.$$

(See Figure 2.) Thus, there are exactly two equilibria. Q.E.D.

In either of these equilibria, it is easy to see that any market beliefs that would rationalize deviation for the bad type also rationalize it for the good type. Consider the difference, for either type, between playing some deviation V and playing the bad type's equilibrium strategy V^*. This quantity is at least as large for the good type as

[12] At least this is generically so: some particular functional form for $\mu(V)$ might give a turning point $\Pi'_B = 0$ at $\widehat{V}$.

for the bad type, since the change in current payoff is the same while the reputation effect is (weakly) better for the good type, who gets a higher performance and hence better posterior reputation on average. Therefore if it is positive for the bad type (i.e., the bad type wants to deviate) then it is positive for the good type, and the good type prefers playing V to playing V^*. However, the incentive-compatibility conditions tell us that the good type is indifferent between playing V^* and playing its own strategy V_g, so it will prefer V to V_g, i.e., will choose to deviate.

It follows from this that an equilibrium refinement in the spirit of Banks and Sobel's "divinity" concept (Banks and Sobel 1987) would require out-of-equilibrium beliefs to satisfy $\xi(p, V) \geq p \text{ for } V \notin \left\{ V^*, V_g \right\}$. Since the good type in a separating equilibrium is indifferent between playing its own strategy V_g and playing V^*, it will prefer to deviate to within epsilon of V^*, earning the same performance (to first order) as at V^* and a positive reputation (by the "divinity" condition above). We can therefore rule out these two signaling equilibria via this equilibrium refinement.

References

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

Chevalier, Judith and Glenn Ellison. (1997). "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy* 105, 1167–1201.

Cho, In-Koo and David M. Kreps. (1987). "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics* 102, 179–221.

Crawford, Vincent P. and Joel Sobel. (1982). "Strategic Information Transmission," *Econometrica* 50, 1431–1451.

Degeorge, François, Boaz Moselle and Richard Zeckhauser. (2004). "The Ecology of Risk Taking," *Journal of Risk and Uncertainty* 28(3), 195–215.

Gollier, Christian. (2004a). *The Economics of Risk and Time*. Cambridge: MIT Press.

Gollier, Christian. (2004b). "Misery Loves Company: Equilibrium Portfolios with Heterogeneous Consumption Externalities," *International Economic Review* 45, 1169–1192.

Schelling, Thomas C. (1960). *The Strategy of Conflict*. Cambridge: Harvard University Press.

Spence, A. Michael. (1974). *Market Signaling: Informational Transfer in Hiring and Related Processes*. Cambridge: Harvard University Press.

Tsetlin, Ilia, Anil Gaba and Robert L. Winkler. (2004). "Strategic Choice of Variability in Multiround Contests and Contests with Handicaps," *Journal of Risk and Uncertainty* 29(2), 143–158.