DISCUSSION PAPER SERIES

IZA DP No. 5058

Extrinsic Rewards and Intrinsic Motives: Standard and Behavioral Approaches to Agency and Labor Markets

James B. Rebitzer
Lowell J. Taylor

July 2010

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Extrinsic Rewards and Intrinsic Motives: Standard and Behavioral Approaches to Agency and Labor Markets

James B. Rebitzer
Boston University,
NBER and IZA

Lowell J. Taylor
Carnegie Mellon University

Discussion Paper No. 5058
July 2010

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 5058
July 2010

ABSTRACT

Extrinsic Rewards and Intrinsic Motives: Standard and Behavioral Approaches to Agency and Labor Markets*

Employers structure pay and employment relationships to mitigate agency problems. A large literature in economics documents how the resolution of these problems shapes personnel policies and labor markets. For the most part, the study of agency in employment relationships relies on highly stylized assumptions regarding human motivation, e.g., that employees seek to earn as much money as possible with minimal effort. In this essay, we explore the consequences of introducing behavioral complexity and realism into models of agency within organizations. Specifically, we assess the insights gained by allowing employees to be guided by such motivations as the desire to compare favorably to others, the aspiration to contribute to intrinsically worthwhile goals, and the inclination to reciprocate generosity or exact retribution for perceived wrongs. More provocatively, from the standpoint of standard economics, we also consider the possibility that people are driven, in ways that may be opaque even to themselves, by the desire to earn social esteem or to shape and reinforce identity.

JEL Classification: D2, J0, M5

Keywords: agency, motivation, employment relationships, behavioral economics

Corresponding author:

James Rebitzer
Boston University School of Management
Department of Markets, Public Policy and Law
595 Commonwealth Avenue
Boston, MA 02215
USA
E-mail: rebitzer@bu.edu

* We are grateful for very helpful comments from Linda Babcock.

1. Introduction

Many of the most widely-discussed and contentious issues facing the U.S. economy concern the use of incentives to solve agency problems. Consider, for example, the problem of reforming the financial system following the recent collapse of financial markets. Explanations for the crash, as well as proposed strategies for effective reform, pivot around the adequacy of high powered financial incentives for ensuring that CEOs, rating agencies, financial advisors and brokers act in the interests of their constituents. Similarly, widely discussed proposals for improving health care quality and reducing costs involve "pay for performance" programs that reward the provision of "cost effective" health care. A growing literature in the economics of education is also exploring the efficacy of rewarding teachers for enhancing student performance.

In these debates, advocates argue that high powered incentives are necessary to get important work done efficiently. Thus, even the very large bonuses to top executives and elite financial engineers are "worth it" in the sense that expected gains from improved performance easily exceed the amount paid out. Critics counter that advocates for high powered incentive systems misunderstand human motivation. High powered incentives are unnecessary because appropriately motivated, selected and socialized agents will perform as well or better when stakes are lower. From this perspective incentives are inefficient because they generate unnecessary and potentially costly inequality within work groups or peer groups and because they needlessly divert agents' attention away from valuable aspects of their jobs that are hard to monitor and reward. In extreme cases, powerful incentives can cause agents to engage in malfeasance. Even more provocatively, some critics argue that the provision of extrinsic incentives undermine agents' intrinsic motives and, in this way, worsen the incentive problem they are designed to solve.

Although advocates and critics may not be aware of it, the public controversies about incentive pay are essentially disputes about the appropriate specification of a workhorse economic model: the principal agent model. In its basic form, this model supports the idea that extrinsic rewards can be an efficient means of motivating agents. The claims of

the critics are supported, however, when more realistic—and *ad hoc*—behavioral assumptions are introduced. Close examination of principal agent models reveals, furthermore, that debates about agency have implications far beyond issues of optimal incentive design. Indeed, the strategies firms adopt to resolve agency problems can have profound effects on labor markets broadly, affecting gender and racial inequality, labor market segmentation and unemployment.

In this chapter we review and analyze the principal agent model from a behavioral perspective. Although the literature is vast, our task is made simpler by the fact that conventional and behavioral principal agent models share a similar structure. In the simplest conventional models an agent is assumed to have utility that is increasing in earnings and decreasing in the provision of effort. Given this utility function, the principal can assess how the agent will react to a given reward structure, and can often link rewards to performance in a way that induces agents to supply efficient levels of effort—even if agents are entirely self-interested and even if measures of performance are noisy and imperfect. Behavioral models employ the same structure, but modify the agent's utility function to include additional psychological factors.

To complicate matters, in applications it is not sufficient to study an isolated agent responding to the policies of an isolated principal. Agents typically work as part of larger groups within organizations and society more broadly, and this can have important implications for the design of reward structures, especially when people have other-regarding preferences, care about inequality, or belong to groups with established norms of appropriate behavior. The policies adopted by firms may also have unexpected effects on labor market outcomes (e.g., can affect unemployment rates) and these outcomes may, in turn, alter the optimal policies of individual firms.

A second complication for conventional principal agent models is that pay structures often perform "double duty," e.g., they must resolve both a motivation problem and some other problem. For instance, principals (firms) might adopt compensation and employment practices that signal the principal's ability to make good on promises to agents. Conversely, employment practices might be designed so as to allow agents to signal some hidden characteristic about themselves, as in "rat race" models in which which individuals provide

"excessive" work hours as a means of signalling an otherwise unobserved personal inclination to work hard. Pay structure also performs "double duty" when workers must attend to multiple tasks. In these situations, rewards for high performance along one dimension draw effort and attention away from other valuable dimensions of performance. In turn, principals must be careful in the assignment of multiple tasks, and might also want to tilt toward lower-powered incentives.

Just as in conventional principal agent theory, "double duty" incentives play an important role in behavioral principal agent models. For instance, in behavioral models pay structures not only elicit effort, but also influence employee perceptions of the legitimacy of the reward structure. Indeed, from a behavioral perspective, a key task of management is to persuade employees of the legitimacy of tasks and rewards and so to help socialize them into the mutually reinforcing expectations of the group.

Personnel policies also do "double duty" if, as is commonly assumed in behavioral models, agents have intrinsic motivation. For example, when agents differ in their intrinsic alignment with an organization's mission, firms with especially evocative missions may design their pay structures so as to attract workers who identify with that mission. Signaling variants of "double duty" incentives are also prevalent in behavioral models of motivation. For example, firms might use compensation policies to signal workers about the likely motivations of co-workers, which can matter for workers who are inclined to conform to workplace norms. Signalling might be used also if the firm has hidden knowledge about a worker's suitability for a task. More provocatively, principals who are sensitive to psychological motivations might set up compensation policies to exploit the possibility that agents send signals to themselves, as a means of nurturing a sense of identity.

Our discussion proceeds as follows. In Section 2 we present a standard principal agent model. We begin with the simplest case—a single isolated agent working for an isolated principal. We then consider the complications that arise when we place this relationship into the context of a firm or a labor market. As we build our model in this section and throughout the paper, we refer to relevant empirical applications from experiments and from field data.

In Section 3 we introduce the problem of extrinsic rewards with "double duty" incentives. We discuss three applications: wages as a signal of firm fitness, rat races, and multi-tasking. In each case, the presence of double duty incentives greatly alters the market outcomes and employment relationships.

In Section 4, we introduce behavioral features to our agency model. To keep the discussion manageable, we focus on four issues: inequality aversion, the desire to reciprocate, behavioral norms, and identity/self-image. Many of the interesting applications in this section focus on professional settings and touch on professional norms.

Section 5 considers behavioral issues in the context of double duty incentives. The most interesting question we approach here is whether extrinsic rewards might "crowd out" valuable intrinsic motivation.

We conclude the paper by highlighting what we see as promising areas for future research.

2. Agency and Extrinsic Rewards

2.1. A Simple Agency Problem

There are a great many interactions in the labor market that can be fruitfully examined as a principal agent problem—an interaction in which the principal uses a reward structure to motivate an agent to pursue some desired action. As a baseline example, consider a principal who seeks to maximize profit, which depends on the "effort" of an agent and the compensation given to that agent, as specified by

$$g(e) - w. \tag{1}$$

In (1), $g(e)$ is the value produced for the principal as a consequence of an agent's effort e (which can be represented as a non-negative scalar) and w is compensation given to the agent. We assume that that $g'(e) > 0$ and $g''(e) < 0$ exist and are continuous. As for the agent, we assume simply that utility is $w - e$. Thus e is the money metric disutility of taking the action that benefits the principal. What makes this problem interesting is that the principal cannot directly observe effort.

Although information asymmetry is essential to our story, to set basic ideas, we ask initially what the solution would be if the principal could observe the agent's effort, and write a contract specifying effort and wage. The firm would then simply maximize (1), subject to the agent's *participation constraint*, which specifies that the utility resulting from the agreed-upon wage and effort must equal or exceed the utility available to the agent elsewhere, i.e., $w - e \geq v$. The principal finds it most profitable to operate with this latter constraint binding, so it immediately follows that the solution to this constrained maximization problem entails

$$g'(e^*) = 1. \tag{2}$$

This outcome is efficient: the marginal value of additional effort equals the marginal cost to the worker of supplying the effort. This agreed-upon effort level is the same as if the agent worked for herself. The resulting wage is $w^* = v + e^*$.

When, instead, the principal does *not* observe the agent's effort level, the principal must find an incentive mechanism to induce the desired effort level. One possibility is that pay

can be conditioned on the value of output $g(e)$. In some instances, though, neither output nor effort are readily observable. We consider such a case. We suppose, instead, that the principal has a noisy signal of effort,

$$x = e + \epsilon, \tag{3}$$

where ϵ is drawn from a differentiable, symmetric, single-peaked density f (with corresponding cumulative density F). There are variety of possible interpretations consistent with this set-up. For example, x might be some objective measure of performance, and ϵ is simply luck. Alternatively, we might interpret x as a principal's impression or opinion of how well a worker is performing, so x is unobservable by outside parties. The ϵ term in this latter case captures miscommunication and misinterpretation of effort.

We start with the case in which the realized value of x is common knowledge, adopting the assumption that the principal can be trusted to honor commitments in which compensation w is conditioned on x. In this case, there are many incentive schemes that will do. To set the stage for results to come, we work through a particularly simple scheme: we assume that the principal commits to a policy of paying wage w_0 if the observed performance x falls below some cut-off $\bar{x}$, and paying $w_1 > w_0$ if x is above that cut-off.[1]

To be clear, we have the following timing in mind: (i) The principal announces the policy (including $\bar{x}$), and posts w_0 and w_1. (ii) The agent decides whether to accept the job, and if she does, takes hidden action e. (iii) Nature plays x. (iv) Given x, the firm pays the agreed-upon wage.

We can easily solve for the optimal wage policy, (w_0^*, w_1^*). Conceptually, the first step is to account for the agent's *best response* to the wage policy. At effort level e, the probability of earning w_0 is $F(\bar{x} - e)$ and the probability of earning w_1 is $1 - F(\bar{x} - e)$. So the agent wants to maximize

$$w_0 F(\bar{x} - e) + w_1 \left[1 - F(\bar{x} - e)\right] - e, \tag{4}$$

[1]For the moment we take the cut-off as given, but as will become apparent, in this particular model it is important that the chosen cut-off be lower than the hoped-for level of performance, e^*.

which leads to the best response, $\hat{e}$, that solves

$$(w_1 - w_0) f(\bar{x} - \hat{e}) - 1 = 0. \tag{5}$$

From this last expression, we notice that the best response is a function of the difference between the higher wage and the lower wage, say $b \equiv w_1 - w_0$ (b is the "bonus" that accompanies the high-performance outcome).[2] Thus we can write $\hat{e}(b)$, noting, for future reference that

$$\hat{e}'(b) = f(\bar{x} - \hat{e})/[bf'(\bar{x} - \hat{e})] > 0 \tag{6}$$

(under the assumption that the second order condition holds). This makes sense; higher-powered incentives elicit greater effort.

Next, the principal must account for a *participation constraint*. The agent accepts the job only if the expected wage equals or exceeds the agent's opportunity cost:

$$w_0 F(\bar{x} - e) + w_1 \left[1 - F(\bar{x} - e)\right] \geq v + \hat{e}(b). \tag{7}$$

The principal's problem then turns out to be straightforward. Expected profit is output minus the expected wage, and given that the participation constraint binds, this is just

$$g(\hat{e}(b)) - [v + \hat{e}(b)]. \tag{8}$$

The first order condition to the principal's profit maximization problem is

$$[g'(\hat{e}(b^*)) - 1]\hat{e}'(b^*) = 0. \tag{9}$$

Above, we noted that $\hat{e}'(b) > 0$ for any best response, so the elicited effort level, $e^* \equiv \hat{e}(b^*)$, described by (9) solves

$$g'(e^*) = 1. \tag{10}$$

The solution thus entails the efficient level of effort, as in (2). The principal pins down b^* using (5), which can be read as giving b^* as an implicit function of e^*. Finally, given b^* and e^*, the firm sets the base wage w_0^* to be as low as possible, while still meeting the participation constraint (7).

[2]The second order condition is $-bf'(\bar{x} - \hat{e}) < 0$, which is satisfied if $\hat{e}$ exceeds the cut-off $\bar{x}$. Hence our observation in the previous footnote that $\bar{x}$ must be below the hoped-for level of effort.

We have obviously chosen to work out an extremely simple case as our prototypical principal-agent model.[3] As simple as the model is, it is nevertheless sufficient to make the point that a solution to the agency problem entails a strategy of conditioning pay on observed productivity. This reward structure can elicit efficient effort levels even when agents are entirely self-interested and when performance measures are noisy and imperfect.

2.2. Agency Matters

In this section we demonstrate the value of thinking carefully about agency in the context of three labor market applications: (i) CEO compensation, a case in which there is a single agent, (ii) personnel policies in a firm, which involves a single principal seeking to motivate multiple agents, and (iii) unemployment and labor market segmentation that can arise in labor markets in which multiple principals compete for agents, and in which the motivation problem is addressed by the threat of dismissal. In each of these three cases, solutions to the principal agent problem are seen to have important social consequences. In each case also, empirical evidence indicates that anomalies exist that point to the importance of behavioral aspects that are not included in the standard principal agent set-up.

2.2.1. CEO Compensation

In advanced economies with modern financial systems, top executives of publicly traded corporations and large financial firms play a central role in the allocation of productive resources. Thus the reward structure under which these executives operate is of considerable economic interest. The rapid increase in the pay of CEOs since the early 1980s is also one of the most striking labor market developments of the past 25 years. These pay increases have contributed in an important way to growing income inequality (Levy and Temin, 2007) and they have also been the target of intense public controversy. The rise in CEO compensation is inextricably linked to agency issues because most of the changes in pay are the result of increasing grants of stocks and stock options. For example, Hall and Liebman (1998) report the median elasticity of CEO compensation with respect to firm stock market performance more than tripled between 1980 and 1994, largely because of the rapid rise in stock based compensation. Bebchuck and Grinstein (2005) document a continuing rapid growth in equity-based compensation for CEOs and top five executives through 2003.

[3]Natural alternative conceptions would allow for risk aversion, as in the Holmström's (1979) classic paper.

One of the great appeals of the principal agent model is that it tells us what efficient CEO reward structures ought to look like. A central prediction of the model is that the efficient reward structure for CEOs and other top executives should have higher levels of expected pay and higher incentive intensity than for other employees. As a simple formal example, suppose that the value to a firm of a particular agent's effort is $g(e) = \theta h(e)$, where $h(e)$ is a concave function increasing in e, and θ is a positive constant that differs across individuals within an organization, depending on the importance of that individual's job to the organization's profitability. CEOs and top executives will typically have the highest values of θ. At the efficient level of effort, $\theta h'(e^*) = 1$ and $\theta h''(e^*) < 0$ (assuming that the second order condition holds), so $\frac{de^*}{d\theta} = \frac{-h'(e^*)}{\theta h''(e^*)} > 0$. Thus effort expectations are highest for CEOs and top executives and, because the agent's "best response" effort is increasing in the size of the bonus, b will also be highest for them. The size of the bonus, b, is likely to be very large, particularly in an environment in which it is difficult to assess the CEO's absolute performance.

One way of expressing the agency model presented above is that compensation should be set so that any agent becomes (at the margin and in expected value) the *residual claimant* with respect to her contributions to the firm; her own personal fortune rises or falls as a consequence of the value of the actions she undertakes on behalf of her firm, $g(e)$. Now in our set-up above, the firm conditions compensation on an imperfect measure of e, under an assumption about the infeasibility of measuring $g(e)$ itself for a typical employee. But in the case of the CEO, her actions might be so consequential to the firm that her contribution essentially represents *firm profit* itself. If so, perhaps the ideal contract would make her, roughly, the residual claimant to the entire corporation. To make that happen, one would want to tie CEO compensation tightly to firm profitability (i.e., stock values) and then give the CEO unlimited latitude with regard to actions she takes on behalf of the firm.

At first blush, incentives for CEOs appear to match well the predictions of the bare bones principal agent model. Top executives earn multi-million dollar salaries and the bulk of their compensation comes in the form of pay linked to stock-based performance measures, as one would expect if stock markets are efficient evaluators of firm value. Indeed, empirical analysis by Gayle and Miller (2009) indicates that the pattern of rising CEO pay and the

rising incentive intensity of this pay over a sixty year period can be explained largely by parameters emphasized in the principal agent model: increasing losses to the firm from CEOs pursuing their own goals rather than value maximization, and the deteriorating value of stock performance as a signal of CEO effort. The former is the result largely of the increasing size of firms.

While it is clear that CEOs ought to function under higher-powered incentives than other employees, it is not clear if compensation boards are setting incentives properly. In their seminal article, "Performance Pay and Top-Management Incentives," Jensen and Murphy (1990) estimate that during the period 1969 through 1983, CEO wealth increased by only \$3.25 in response to a \$1000 increase in firm value. This number would seem to offer a *prima facia* case for CEOs having inadequate incentives to increase shareholder value. Hall and Liebman (1998) present empirical evidence that in fact there is a substantially tighter link between CEO compensation and firm value, particularly when they examine more recent periods (1980-1994).[4]

Still, in large corporations CEOs are far from being residual claimants. As Hall and Liebman (1998) suggest, this might pose little problem for the proper alignment of some CEO actions but create large problems for others. For example, a CEO who receives \$1 in compensation per \$100 value created for a firm might be sufficiently motivated to make smart, carefully-reasoned decisions about resource allocation to proposed projects. But this same CEO gets an effective 99% discount on a multi-million jet purchased by the firm for his own use.

This latter point is easily illustrated with a slight modification to our baseline principal agent model. Suppose, now, that firm profit is

$$g(e) - w_S - w_N, \tag{11}$$

where now w_S is the CEO's salary and w_N is the non-salary cost that results from the CEO's actions that increase the CEO's welfare at the expense of the firm (e.g., expenditures on a

[4]In particular, they estimate that during the period 1980–1994 a typical CEO whose actions caused the firm to move from the 30th percentile of annual returns to the 70th percentile enjoyed an increase in annual compensation of 1 to 4 million dollars (1994 dollars), mostly through the increased value of the CEO's stocks and stock options. For stellar performance the increase in CEO wealth was estimated to be much higher.

jet for CEO use). We now let the CEO's utility be $w_S + u(w_N) - e$, where $u(w_N)$ is the money metric value to the CEO of non-salary expenditures—a function that is obviously increasing in w_N. We also expect $u'(0) > 1$ and $u''(w_N) < 0$.[5] If the firm could observe and direct e and w_N, it would choose e^* and w_N^* so that

$$g'(e^*) = 1 \quad \text{and} \quad u'(w_N^*) = 1. \tag{12}$$

Suppose instead the firm sets the variable component of the CEO's compensation equal to the share s of the firm's profit, i.e., $s[g(e) - w_N]$. The best response here will entail the CEO choosing

$$g'(e^{**}) = 1/s \quad \text{and} \quad u'(w_N^{**}) = s. \tag{13}$$

Comparison of (12) and (13) demonstrates the problem: If $s < 1$, we have too little CEO effort on behalf of the firm and too much squandering of resources on the CEO.

How should the corporation's compensation board respond? One argument is that s must be driven ever closer to 1, even if this entails a substantial direct surplus transfer to the CEO. An alternative might entail the judicious combination of monitoring and more-narrowly directed incentives—a process that would likely play on the hope that hard-to-observe excessive levels of w_N by the CEO would be limited by shame or a sense of obligation to shareholders. This latter strategy can only be studied in a set-up that takes such behavioral aspects into account.

In any event, it is infeasible for firms to set up pay structures in which CEOs literally become residual claimants. The issue at hand is readily visible in our baseline principal agent model. As we note above, if it is optimal to induce a CEO to exert a very high level of effort on behalf of the firm, it is necessary also to have a very high bonus. With a binding participation constraint, this means that the contract will specify *negative* base pay. If this is infeasible, i.e., if the CEO cannot be compelled to pay the firm when performance is poor, it will be necessary to modify the contract to take account of the CEO's *limited liability*.

A simple solution to compensation with limited liability might look like this: each year the CEO is offered high pay and is rewarded further by having her contract renewed for the

[5]Concavity is natural here. The assumption $u'(0) > 1$ simply allows for the fact that *some* expenditures on CEO wellbeing are efficient.

following year if she has a high observed performance level, but she is dismissed if observed performance is insufficiently high. We characterize this solution in detail in Section 2.2.3 below. Two key results that emerge in that analysis are germane here. First, the solution entails a surplus transfer from the firm to the CEO; the necessity of having a high-powered incentive leads to an especially high salary for the CEO when there is limited liability. Second, the more precise the principal is in assessing performance, the lower will be this surplus transfer.[6]

Our baseline principal agent model therefore predicts that profit-maximizing principals will, whenever possible, seek to reward performance and not luck.[7] There is persuasive evidence that, to the contrary, at least some CEOs are rewarded for observable luck. In particular, Bertrand and Mullainathan (2001) show that because of the way many firms tie CEO compensation to stock market performance, "CEO pay in fact responds as much to a lucky dollar as to a general dollar." For example, increases in the world price of oil causes stock price increases in the oil industry. In the baseline principal agent model, such "luck" should have no impact on CEO pay, but in reality some CEOs are observed to reap handsome rewards simply because of such luck.

The use of high-intensity incentives through the use of stock options—a common way of tying compensation and firm outcomes—can create additional problems. If options are very far under water, their value as incentives degrades to near zero—obviously an undesirable state of affairs.[8] Conversely, when stock prices are just below the stock price, the payoff to even small increases in the stock price are huge, and this can create irresistible temptations to game the compensation system. Heron and Lie (2009), for example, estimate that 13.6% of all option grants to top executives during the period 1996-2005 were backdated or otherwise manipulated.

[6]We prove this latter result formally in Section 2.2.3 for the important case in which there is an ongoing (multiple period) relationship between the principal (firm) and agent (CEO).

[7]Notice that this result holds even though we assume that agents are risk neutral. The result is reinforced, of course, if the agent is risk averse.

[8]Hall and Knox (2004) estimate that at the height of the bull market in 1999, roughly a third of executive options were under water. Companies often respond to this non-linearity in stock option returns by increasing option grants following stock price declines.

It is especially surprising that the use of stock options is a part of compensation even for many corporate employees below the top executive level. Hall and Murphy (2003) report that in S&P 500 corporations, roughly 90% of the outstanding options are granted to employees below the top five executives. This pattern is very hard to reconcile with principal agent models because the efforts of individual employees below the top five executives, it would seem, can have little direct influence on the price at which their company's stock trades. In this case, a stock-based compensation of any sort is likely to have little effect on effort levels. The use of stock-based compensation for lower-level employees is even harder to understand if agents are risk averse. Here again, the simple principal agent model appears to be inadequate for explaining compensation practices, perhaps because of the omission of important behavioral aspects.

2.2.2. Personnel Policies

There have been a great many applications of the principal-agent model for the purpose of understanding compensation policies within firms more broadly. As a simple example, consider a profit-maximizing employer whose n workers produce output according to a production function, $Y = G(e^1, e^2, \ldots, e^n)$ per period, where e^i is worker i's effort. As above, output is increasing in effort: $\partial G/\partial e^i > 0$.

We continue to assume also that in a given period a worker chooses e and receives utility $w - e$, and that the firm cannot condition compensation on Y (or, in any event chooses not to). Importantly, for this application, we also assume that workers do not base their effort decisions on the effort or compensation of other workers.[9] Then we can treat the firm's agency problem with a given worker in terms of the function $g(e^i)$, which is the value of product that results from production $G(e^i, e^{(-i)})$, where $e^{(-i)}$ denotes effort levels of workers other than i. We assume that $g'(e^i)$ and $g''(e^i)$ are continuous, and that $g''(e^i) < 0$.

In setting up our baseline agency model in Section 2.1 we ignored an issue that is generally germane in the workplace: The indicator of performance, x in our model, is typically observed only by the manager and by workers within a firm, and thus cannot readily be used as the basis for forming contracts that can be enforced, say, by an outside court.

[9]This last assumption follows naturally enough, given the utility function we have specified. In our discussion of behavioral models in Section 4, we allow for the possibility that workers do care directly about the effort or compensation of other workers.

Workers who understand that they will have no recourse if a manager violates the implicit agreement—"pay a bonus for high observed effort"—will logically decline to accept the proposed agreement. How might a firm proceed in this case? One possibility is to set up a competition among it's n workers. Suppose, for example, that the firm cannot directly condition pay on x, but *can* commit to an evaluation process at the end of the period in which (i) workers are ordered on the basis of observed performance, and then (ii) the fraction P who are lowest-performing are paid w_0, while the remaining high-performing workers are paid $w_1 > w_0$. The key idea here is that while individual performance is not well observed, everyone can observe the agreed-upon reward structure and see if the firm is meeting that obligation.

We can easily find a Bayesian equilibrium the which all workers supply the same level of effort in response to the competition. Suppose that worker i believes that all other workers are going to play $\tilde{e}^{(-i)}$. Now what is her best response? The worker first uses her knowledge of P to accurately assess the cut-off value of observed performance, say $\tilde{x}$, which separates low- and high-performance assessments. That is, she takes note of the value $\tilde{x}$ that solves $F(\tilde{x} - \tilde{e}^{(-i)}) = P$. Given that value $\tilde{x}$, her optimal choice is to set effort level e^i so as to maximize

$$w_0 F(\tilde{x} - e^i) + w_1 \left[1 - F(\tilde{x} - e^i)\right] - e, \tag{14}$$

which leads to a best response given by

$$(w_1 - w_0) f(\tilde{x} - \hat{e}^i) - 1 = 0. \tag{15}$$

But this is exactly the worker best response we solved in our baseline example (compare (5) and (15)). Given this insight, it is easy to verify that the firm has a workable plan here: The firm starts by setting the "tournament prizes," (w_0, w_1), to be (w_0^*, w_1^*), as derived in our baseline example in Section 2.1. Then it chooses the fraction P^* so that $P^* w_0^* + (1 - P^*) w_1^*$ just satisfies the participation constraint (5). If worker i believes other workers are choosing effort level $\tilde{e}^{(-i)} = e^*$, she responds by also choosing e^*. All workers behave the same in equilibrium.[10]

[10]It is important here that the manager actually follows through on the promise to award the higher wage to workers who have the highest realized values of x. This might be sensible, especially if realized values of

The logic outlined in the preceding paragraph is the starting point of Malcomson's (1984) well-known paper on hierarchy and internal labor markets. He suggests that the "tournament prize" idea can be fruitful for thinking about the internal organization of the workplace. He works with a two-period model. In the first period of one's career, within a firm, each worker receives the same wage.[11] Then in the second period, the fraction $(1 - P)$ of workers who have been most successful as junior employees are promoted to high-paying jobs, while the fraction P who have been less successful are retained in low-paying jobs (at a wage that is high enough to keep them from moving to other firms). The tournament provides an extrinsic reward designed to elicit optimal effort from young workers.[12]

As Malcomson (1984) notes, the simple tournament model we have just outlined is consistent with some commonly observed features of organizations, e.g., that wage structures in organizations are often "hierarchical," with workers falling into distinct pay grades, that often workers in high-paid positions are promoted from within, that wages typically rise with seniority (perhaps by more than productivity), and that the variance of wages increases with seniority. Indeed, one of the major contributions of agency theory to labor economics is it's ability to help explain the origin of firm wage policies and hence clarify the contribution that personnel practices make to shaping the wage structure.[13]

As was true in its application to CEO compensation, the first-order predictions of the agency model receive considerable empirical support, but there are anomalies that suggest the model may not offer an altogether satisfactory guide to understanding the internal structure of organizations.

To begin, it is important to recognize that extrinsic incentives do matter within organizations, often in exactly the way predicted by simple models of agency: Lazear (2000) found an increase in effort when a glass installer went from fixed pay to pay-for-performance.

x are reasonably well known by people within the firm. After all, why wouldn't the manager want to reward to workers who have the highest performance outcomes? Having said that, if there is "favoritism" based on other criteria, the proposed incentive plan falls apart.

[11]This wage solves a two-period participation constraint. The first period wage is low, possibly negative.

[12]In Malcomson's (1984) set-up, the agency problem is left unresolved for older workers. The point is that young workers can be motivated by the promise of future prizes (promotions, raises, etc.). Such incentives are less likely to be effective for workers nearing retirement. This is quite typical in agency-based models of "internal labor markets," and it doesn't substantially alter the basic insights generated in these models.

[13]There are now a number of insightful overviews of the topic, including Lazear (1998), Prendergast (1999), Malcomson (1999), Gibbons (1998), and Oyer and Schaefer's chapter in this *Handbook.*

Kahn and Sherer (1990) document the effectiveness of an evaluation-and-bonus program at a manufacturing firm. Jacob (2005) shows that high-stakes testing in the Chicago Public Schools does alter teacher behavior—intensifying effort in improving student's test-specific skills, while substituting away from low-stakes subjects like science and social studies. Important work by Theodore Groves and John McMillan and their co-authors shows that strengthened incentives led to substantial productivity increases in Chinese industry and agriculture.[14] And, of course, there are many other examples in the literature.

Principal agent models also require that firms are choosing pay policies in an optimal way. It is hard to find direct evidence that pay policies are chosen in this way. Indeed, much of the literature showing that "incentives work" does so by exploiting the measured consequences of poorly designed incentives. That is, they clearly demonstrate that organizations—at least in some cases—do *not* choose incentives optimally. This is clear, for example, in Oyer's (1998) work, which calls attention to the fact that salespeople seem to intensify effort at the end of the fiscal year if by doing so they can surpass performance thresholds and earn a bonus. At the organizational level, Courty and Marschke (2004) similarly demonstrate that a large government organization strategically reported performance outcomes to increase earned rewards, and did so at the expense of productive activities. In work with Martin Gaynor (Gaynor, Rebitzer and Taylor, 2004), we document the effects of an HMO's incentive contract designed to limit expenditures by physicians, but our identification strategy relies on the observation that a key feature of the incentive contract was implemented haphazardly. An even more extreme example is Jacob and Levitt's (2003) demonstration that public school teachers responded to a shift toward higher-powered incentives by cheating, e.g., by altering questions on standardized tests taken by students.

Some of the ancillary predictions of principal agent models also lack empirical support. In the two period model we present above, compensation in period 1 ought to move inversely to expected compensation in period 2—a result that follows directly from a two-period participation constraint. In an earlier paper studying law firms (Rebitzer and Taylor, 1995a) we tested this hypothesis. We find, contrary to the predictions of our principal agent model, that large law firms with extremely high second period compensation (in the form

[14]See, e.g., McMillan, *et al.* (1989), Groves, *et al.* (1994), and Groves, *et al.* (1995).

of the high income of partnership) also pay their starting associates high salaries relative to smaller firms. This would seem to indicate that successful law firms use some form of "rent sharing"—a strategy that emerges when we add such behavioral features as "inequality aversion" (in Section 4.1).

A particularly jarring feature of the minimalist agency model of personnel practices is the "irrelevance of *ex post* inequality." The compensation structure emerging from our model might indeed be termed "pay for luck" rather than "pay for performance." The principal and agent(s) know that the equilibrium effort level is e^*. Even so, it is important that pay be based on the measure of observed performance so as to provide the crucial extrinsic incentives. This feature—that rewards or punishments are based on an observed outcome, not on the actual behavior (even though those behaviors can be deduced)—is very common in game-theoretic approaches, including much of the work presented below. Anybody who has spoken with managers (or chaired an academic department) knows that people don't respond well when they are paid less than co-workers for what appears to be arbitrary, capricious or random reasons. This observation has been widely examined in the behavioral literature on agency, and we will discuss it's implications in Section 4.

2.2.3. Involuntary Unemployment and Market Segmentation

In 1984, Shapiro and Stiglitz set out an influential "efficiency wage" model that illustrates an important feature of agency models: the actions an individual firm takes to resolve an agency problem can give rise to important social costs when adopted throughout the market. In the case of efficiency wage models, the social costs are those arising from involuntary unemployment and labor market segmentation.[15] The set-up we present here is a recasting of the Shapiro and Stiglitz model taken from Ritter and Taylor (forthcoming).

We consider a market in which there are a large number of identical profit-maximizing employers, each of which faces the agency problem we outlined above. Each firm in the model is assumed to behave as outlined in Section 2.1: the idea is to pay well for "good outcomes" while penalizing workers for "bad outcomes" to the maximum extent possible. *Limited liability* is invoked through the assumption that the only penalty that the firm can implement is to dismiss a poorly performing worker. The motivation problem is resolved by

[15]Similar points were also made in the important work of Bowles (1985).

employers making jobs sufficiently valuable that workers will provide effort so as to prevent dismissal.

To capture the idea that jobs have value, it is necessary to set the model up in a multi-period framework. The agency model outlined in Section 2.1 is thus assumed to pertain for each period indefinitely and workers are assumed to live indefinitely with a discount rate ρ.[16]

The Basic Set-Up

Employees are paid w for one unit of labor per period. In each period a worker chooses e, and this produces utility $w - e$. In this model, the alternative to employment is unemployment, which results in utility $v = 0$ in the period. The present value of being unemployed is V^u (which is not 0 because there is some prospect of being hired in the future). Hiring and termination are costless to the firm.

The model is a game between the firm and a worker with the following order of play in each round: (1) The firm offers a wage w. (2) The worker chooses effort level e. (3) Nature plays x using the distribution $f(\ )$. (4) The firm pays w. (5) The firm decides whether to retain the worker or end the game. We focus on the perfect Bayesian equilibrium in which the worker is retained if and only if x exceeds an endogenous threshold $\bar{x}$. We assume that $\bar{x}$ is common knowledge. (Workers can infer $\bar{x}$ by observing the frequency of terminations.)

The solution method mimics the steps we took in the simpler model above. In particular, we first find the worker's best response. Then we see how the firm will chose it's personnel policies ($\bar{x}$ and w) in light of the worker's best response.

The Worker's Best Response

Let $\hat{e}(w, \bar{x})$ be the worker's best response. To find that best response notice that for a person who chooses e in the current period, and then reverts to $\hat{e}$ in all future periods, lifetime utility is given by

$$V(e) = w - e + \frac{\big[F(\bar{x} - e)V^u + \big(1 - F(\bar{x} - e)\big)V(\hat{e})\big]}{1 + \rho}. \tag{16}$$

[16]Valuable long-term employment relationships are central to these models, and thus so are shocks to employment. The model here can be enriched to allow for these exogenous shocks to employment relationships, but for simplicity we omit this feature.

The employee maximizes $V(e)$ by choosing $\hat{e} > 0$. For an interior solution, the first order condition is

$$\frac{[V(\hat{e}) - V^u]}{1+\rho} f(\bar{x} - \hat{e}) - 1 = 0. \tag{17}$$

As in our baseline agency model, the second order condition holds when $f'(\bar{x} - e) > 0$. As we have noted, this incentive elicits effort because the job is valuable: $[V(\hat{e}) - V^u] > 0$. As is typical of models that invoke limited liability, the participation constraint does not bind.

Evaluating equation (16) at $e = \hat{e}$ and solving for $V(\hat{e})$, then substituting into the first order condition (17) produces

$$w = \hat{e} + \frac{\rho V^u}{1+\rho} + \frac{\rho + F(\bar{x} - \hat{e})}{f(\bar{x} - \hat{e})}. \tag{18}$$

This last expression implicitly defines the worker's best response, $\hat{e}(w, \bar{x})$.

Firm Profit Maximization

Now we can turn to the firm's objective. It seeks to maximize profits, taking into account the worker's best response, i.e., maximizes

$$g(\hat{e}(w, \bar{x})) - w.$$

The solution can readily be found using the same steps we followed in solving the agency problem in Section 2.1. In this instance the optimal employment policy again induces the socially optimal performance level regardless of $f(\ )$:[17]

$$g'(e^*) = 1.$$

The noise in the environment does, however, affect the distribution of surplus. In particular, Ritter and Taylor (forthcoming) establish the following results: First, when the firm optimally chooses w and $\bar{x}$, the resulting probability of retention, $F(\bar{x}^* - \hat{e}^*)$, is unaffected by the level of σ^2 (the variance of the density $f(\ )$).[18] Second, the optimal wage does depend

[17]This shows that efficiency wage motivation can lead to socially optimal effort levels, but this need not always be the case. For example, Allgulin and Ellingsen (2002) show that there can be distortions away from the socially optimal effort level when the principal has discretion over investments in monitoring.

[18]In fact, the probability of retention is shown to be $F^* = \frac{\phi(z^*)^2}{\phi'(z^*)} - \rho$.

on σ^2, as follows:

$$w^* = e^* + \frac{\rho V^u}{1+\rho} + \frac{\phi(z^*)}{\phi'(z^*)}\,\sigma, \tag{19}$$

where z^* is a "standardized" random variable, $z^* = (\bar{x}^* - e^*)/\sigma$, and $\phi(\,)$ is the "standardized p.d.f.," i.e., the p.d.f. of $\frac{\epsilon}{\sigma}$. The more intractable the agency problem—the greater the value of σ—the higher is the wage required to achieve efficient effort and so the greater the surplus accruing to the worker.

Equilibrium Unemployment

The equilibrium of the model we have just outlined generates unemployment.

Let V^* be the present value of lifetime utility of an employed individual who works at the optimal effort level. Recall that V^u is the expected lifetime utility for an individual who is unemployed. This utility level depends, clearly, on the probability of job acquisition. Let that rate be a.[19] Given that current-period utility of an unemployment person is zero, the expected lifetime utility of an unemployed individual is

$$V^u = 0 + \frac{\left[aV^* + (1-a)V^u\right]}{1+\rho}, \tag{20}$$

so in turn we can use (17) and (20) to solve for V^u and substitute into equation (19), giving

$$w = e^* + \frac{1}{\phi(z^*)}\left(a + \frac{\phi(z^*)^2}{\phi'(z^*)}\right)\sigma. \tag{21}$$

Equation (21) gives the locus of potential equilibrium values of w and a.

Now job loss among the employed occurs with probability $F^* \equiv F(\bar{x}^*, e^*)$, while job finding among the unemployed occurs with probability a. So if u is the steady state unemployment rate, we must have $(1-u)F^* = ua$. Solving for a and substituting into (21), and substituting also for F^* using footnote 18, gives

$$w = e^* + \frac{1}{\phi(z^*)}\left(\rho + \frac{F^*}{u}\right)\sigma. \tag{22}$$

This expression shows potential equilibrium wage and unemployment levels for the labor market.

[19]Note that the rate a is a known constant to any individual, but of course is endogenous to the economy as a whole. We solve for the equilibrium rate a shortly.

Figure 1 shows the market equilibrium when long-run labor demand is perfectly elastic, at w^E. (More general formulations are easily handled.) Equilibrium unemployment, u^E, solves

$$w^E = e^* + \frac{1}{\phi(z^*)}\left(\rho + \frac{F^*}{u^E}\right)\sigma. \tag{23}$$

Clearly $u^E > 0$. Also, inspection of (23) shows that an increase in σ increases unemployment. This outcome is intuitive. The weaker the link between the dismissal threat and employee behaviors, the stronger are the incentives required to elicit the desired effort level. In equilibrium, heightened incentives require higher unemployment.

This model of equilibrium unemployment—the Shapiro-Stiglitz model—has emerged as a workhorse for the analysis of macro-labor issues. It has proved to be useful also for evaluating policies like unemployment benefits, the public interest in regulating firms' layoff decisions (i.e., just-cause dismissal requirements, as discussed in Levine, 1991), and the potential of minimum wage policy to actually increase employment (Rebitzer and Taylor, 1995c). Having said this, economists are divided on the extent to which efficiency wages are an important source of equilibrium unemployment. Other forces, like labor market frictions, matter as well in determining equilibrium unemployment rates.[20] Efficiency wages are clearly not the whole story.

Labor Market Segmentation

Although we have focused on unemployment, the lost output from unemployment may not capture the full social costs of efficiency wage personnel policies. After all, if there are some jobs in the labor market where agency issues are of little importance, workers should generally be able to find jobs there. From an efficiency perspective, finding work in these "secondary jobs" is similar to unemployment in that individuals in the secondary labor market would prefer higher-productivity "primary jobs," but the equilibrium supply of qualified workers for these jobs exceeds the demand.

[20]Hornstein, Krusell, and Violante (2007) offer a review and discussion of models of unemployment resulting from search frictions. Search models produce both unemployment and wage dispersion, but search frictions sufficient to account for equilibrium unemployment imply far less wage inequality than is actually observed.

Labor market segmentation emerges if we enrich our efficiency wage model by allowing the difficulty of agency problems to vary across firms. Recall from (21), that in the Ritter-Taylor version of the Shapiro-Stiglitz model firms choose to pay

$$w = A + B\sigma, \tag{24}$$

where A and B are positive constants that are independent of σ. Firms that have low values of σ, i.e., who have production processes with accurate measures of worker effort, can pay wages that are relatively low. On the other hand, firms will choose to set wages high when effort is hard monitor or, equivalently, when they face high values of σ.

This latter observation was emphasized in Bulow and Summers' (1986) paper on "dual labor markets." In their conception, firms with severe agency problems pay high wages and are said to belong in the primary sector. The strategy of paying high wages is effective because workers are motivated by the prospect of retaining valuable jobs. Thus we also expect to observe low levels of voluntary exit from such firms and efforts on the part of firms to retain workers even in a down-turn. In contrast to the primary sector, firms that have modest agency problems can pay wages that are close to the market-clearing level. These secondary sector firms will be less concerned about worker turnover. In an extension of this argument (Rebitzer and Taylor, 1991) we show that firms which employ efficiency wages as a motivating device will also be led to hoard labor, i.e., employ labor above the value-of-marginal-product curve. By taking actions to ensure future employment—perhaps by hiring contingent workers to absorb demand shocks—firms can reduce the wage needed to provide optimal motivation to workers.

The most widely examined empirical prediction of efficiency wage models of labor market segmentation is that there will be cross-firm and cross-industry wage variation resulting from firm characteristics, rather than worker characteristics. There is considerable evidence for industry and firm wage effects (including well-known work by Krueger and Summers, 1988, on industry effects, and Brown and Medoff, 1989, on firm size effects) but it is often unclear how much this is due to factors emphasized in efficiency wage models (such as monitoring difficulties) or other market imperfections such as those emerging from search frictions (e.g., Burdett and Mortensen, 1998).

One potentially helpful approach entails the study of specific firms and industries, with an eye toward the central predictions of the model. Thus, Cappelli and Chauvin (1991) examined worker performance across plants within the same firm, examining the extent to which workers seem to choose performance on the basis of the value of their job relative to other opportunities in their local labor market. They find evidence that is generally supportive of the efficiency wage set-up. Similarly, in work with Daniel Nagin and Seth Sanders (Nagin, Rebitzer, Sanders, and Taylor, 2002), we evaluated a field experiment in which a firm manipulated monitoring rates across several work sites. Consistent with the effort-regulation model set up above, there was substantially more malfeasance in locations with low monitoring levels.

At the broadest level, the efficiency wage literature points to important social costs that emerge as a result of the strategies individual firms use to resolve agency problems. If firms indeed rely on the the fear of job loss to motivate employees, labor markets can be expected to waste human capital on a large scale through involuntary unemployment and labor market segmentation. If, however, other motivators can be mobilized to resolve agency problems, the situation may not be so grim. The costs of agency problems might be further reduced if schools can socialize children to be especially responsive to these alternative motivators. Indeed, some have speculated that such socialization may be the source of much of the social and private returns to investments in human capital. We take up some of these alternative motivators in Section 4 below. Before turning to these, however, we must first introduce another conceptual building block that is important for our story—incentives that are intended to work along more than one dimension.

3. Extrinsic Rewards and Dual-Purpose Incentives

In real-world applications, compensation policies are often asked to do "double duty." A well known and intuitive example of this is Lazear's (2000) study of compensation practices at Safelite, a windshield installation company. As might be expected from the basic principal agent model, the introduction of an explicit piece rate system induced many workers to perform at a higher intensity level. In addition, the piece rate system had a *selection* effect: workers who disliked having to choose between lower compensation and a faster pace of

work left the firm while, at the same time, the firm was able to attract workers drawn to the income-effort tradeoffs inherent in the piece rate system. In this case, incentive pay was serving a dual role: motivating and attracting employees.

At Safelite, the selection reinforced the effect of incentives on work effort. In many cases, however, there is a tension between the multiple effects of incentive pay, and thus employers must make compromises along one dimension in order to accomplish an objective along a second dimension. In the following three sub-sections we give examples of this phenomena and illustrate how the introduction of double duty incentives helps address well known anomalies. The discussion in this section also sets up of the discussion of behavioral models that follow. The special problems posed by dual purpose extrinsic incentives can be either ameliorated or sharpened by behavioral factors of the sort discussed in Section 4. In addition, dual purpose incentives play a key role in the models of intrinsic motivation presented in Section 5.

3.1. High Wages as a Signal of Firm Fitness

We begin by discussing a theoretical issue that is well known in the literature on efficiency wages. As we have seen, firms that pay efficiency wages must set wages above the market clearing level to elicit the desired level of effort. Effort can, however, be elicited more cheaply by a deferred compensation policy that causes employees to, in effect, *post a performance bond.* By judiciously back-loading pay, firms can create powerful work incentives while choosing a wage path whose discounted present value is equal to the market clearing wage. With this option available, why would employers ever select a more costly efficiency wage strategy? The practical relevance of this theoretical puzzle is sharpened by empirical work suggesting that even when very large amounts of deferred compensation were available, as is the case in the promotion from associate to partner in large corporate law firms, firms set wages as if they were pursuing an efficiency wage strategy (Rebitzer and Taylor, 1995a).

Ritter and Taylor (1994) tackled this issue by observing that for both efficiency wages and the performance-bond incentive, the power to shape behavior depends on the likelihood that the firm will honor its future commitments to employees. All else equal, firms will more effectively solve agency problems if employees expect them to be highly reliable in honoring future wage commitments.

Ritter and Taylor build upon this insight by positing a market in which there are two types of firms: highly reliable firms (i.e., firms that are unlikely to go bankrupt or otherwise renege on commitments) and less reliable firms (firms that are more likely to become bankrupt or renege). Reliability is known by the firms but not by anyone else, though the distribution of types is common knowledge. Firms would like to resolve their agency problem as cheaply as possible, and are inclined to do so by asking workers to post bonds in the form of deferred compensation. The posted bond is forfeited if the worker is judged to be working at a sub-standard effort level but is returned, with interest, if the worker's observed performance meets the expected standard.

Under some conditions, all firms pursue the same deferred compensation strategy. In this *pooling equilibrium,* workers will require a rate of return on their bonds that reflects the aggregate level of riskiness, based on the market-wide probability a firm will fail and be unable to return the bond.[21] This, of course, is a good deal for low-reliability firms—who benefit by paying a below-market interest rate on the bonds that workers have posted—but a bad deal for high-reliability firms. A more interesting possibility is that efficiency wage strategies might emerge for some firms as a *separating equilibrium.* In this equilibrium, a reliable firm that deviates from the bonding strategy—by paying a high wage up front—offers a credible *signal* that it is a highly desirable counterparty for long-term relationships. If the offered wage is sufficiently high, low-reliability firms will find it unprofitable to mimic this strategy, and the equilibrium thus satisfies the "Intuitive Criterion" of Cho and Kreps (1987). Equilibrium efficiency wages arise endogenously, in short, without a recourse to limited liability arguments.[22]

Our primary point here concerns the use of incentives policies to do "double duty." In the separating equilibrium, highly reliable firms use wage policies to solve an agency problem *and* to signal the fitness of the firm. In order to pursue both objectives, these firm must compromise on their use of deferred compensation and this compromise necessarily

[21]In this model, young workers have concerns about the realization of high earnings at the firm later in their careers. Wages paid to young workers thus depend on the degree to which firms that are judged to be unstable.

[22]In the law firm context, the term of art for paying very high wages to summer interns and associates is "paying full freight." Law firms that are able and willing to "pay full freight" signal that the value of their partnership is high, and this in turn allows them to attract the best talent.

introduces distortion. Thus, in the separating equilibrium there is a surplus transfer to workers employed by highly reliable firms and employment in the reliable-firm sector is inefficiently low.

3.2. The Rat Race

Rat race models build on a simple observation: early in their careers many successful professionals appear to be overworking. It is commonplace to find lawyers, consultants, and assistant professors complaining that the hours they work are simply "too much" and that they interfere with forming and raising a family. These strains are increased by the dramatic influx of women into professional occupations because overwork is most intense during prime years for family formation and childbearing. From the point of view of simple models of labor markets, this sort of overwork is anomalous. Firms are in competition for talent, and it would seem that the most successful competitors would be those who best accommodate employee preferences about work conditions—including work hours. In his famous paper on the "rat race," Akerlof (1976) offers a potential resolution to this anomaly based on unobservable worker heterogeneity.

Akerlof's set-up focuses on a production line. At the end of the day, line workers are jointly rewarded on the basis of total output. There are two types of workers—those inclined to work hard and those inclined to work less hard. To employers these workers appear identical, so they all earn the same wage. If both types of workers accept positions on the production line, this is a great deal for low-effort workers (who would earn lower pay than high-effort workers in a perfect-information world) but a bad deal for high-effort workers. This is precisely the situation that might lead firms to adopt rules that will provide high-effort workers the opportunity to credibly *signal* that they are in fact high-effort workers. Thus, Akerlof's proposed solution is that the firm set the production line at a speed that is uncomfortably fast for high-effort workers but more uncomfortable yet for low-effort workers—so uncomfortable indeed that the low-effort workers will opt out of working for the firm. The rat race thereby serves the useful function of screening out the low-effort workers.

In Akerlof's model, compensation policies are, quite clearly, doing "double duty." Compensation arrangements and work conditions are structured (i) to compensate workers at a

level necessary to induce them to accept employment at the firm, and (ii) to create incentives that attract the "right kind" of worker to the firm. The distortion here is that workers are being asking to provide effort that exceeds the first-best level. In a marketplace with many employers, the market can devolve into an equilibrium in which *all* firms that hire high-effort workers ask those workers to work at uncomfortable effort levels. This "adverse selection" equilibrium occurs because any one firm failing to adopt a rat race would be swamped by low-effort workers. The equilibrium is inefficient, in the sense that all firms would experience increased profitability if they coordinated on a lower-effort work norm.[23]

Akerlof's demonstration of an overwork equilibrium was presented in a self-consciously unrealistic example, but subsequent theoretical and empirical work suggests that it is an important phenomenon in professional labor markets. For instance, in a paper with Renee Landers (Landers, Rebitzer, and Taylor, 1996), we embed Akerlof's idea into a simple tournament-partnership model designed to shed light on work practices in large U.S. law firms. In our two-period model, young lawyers accept salaried positions as "associates" for one period, and if deemed suitable are promoted to be "partners" (equity shareholders) in the subsequent period. Partners share equally in firm surplus. This equal sharing rule gives incumbent partners powerful incentives to promote only highly motivated lawyers into the partnership.

We assume that there are two types of lawyers who are equally productive but have differing preferences over the hours they prefer to work: these are "short-hour" and "long-hour" workers. Now in our setting, firms have the incentive to attract workers who will be inclined to work long hours. The reason is that when workers become partners—at which point they share firm surplus—the long-hours individuals will engage in less free riding. As in Akerlof's model, an adverse selection equilibrium emerges. In our case, associates' willingness to put in extended hours over many years serves as a credible signal that they are long-hour individuals.

[23]Intuitively, the over-work equilibrium might persist even when there is a small number of low-effort workers. No one firm can deviate from the equilibrium without suffering harm from adverse selection. But if all firms backed away from overwork requirements, any one firm would get stuck with only a negligible number of low-effort workers. All firms would be better off.

Empirical evidence for the relevance of the over-work equilibrium comes from an empirical examination of work hours and work preferences in associates at two large East Coast law firms. In a survey conducted in these firms, we find that most associates express a preference for working shorter hours (with a correspondingly lower salary) but, importantly, their willingness to work shorter hours hinges on the work-hour norms adopted by other associates. In addition, we find that in making promotion decisions, partners use an associate's willingness to work long hours as an indicator of the motivation associates have to excel. These findings would not be expected in a conventional labor market, but are precisely what one would expect if work hours are being used as a signal of one's otherwise-unobservable type.

In our paper (Landers, Rebitzer and Taylor, 1996) we abstract from "career concerns" outside one's own firm, but it is clear that overwork early in one's career might be valuable not only as a signal within a firm but as a means of career advancement elsewhere. Completion of six years as an associate at a law firm known for abusing associates with grueling hours can be a valuable means of demonstrating an important but hard-to-observe trait to other employers in the marketplace.

The key idea that current on-the-job behaviors can affect one's future career, through their impact on reputation, is studied in insightful papers by Holmström (1999) and Gibbons and Waldman (1999). Gicheva (2009) shows how long work hours early in one's career can affect an individual's value in the market later in the career. Gicheva shows, further, that her model helps explain wage growth in a sample of workers who took the Graduate Management Admissions Test (GMAT). Specifically, she shows that among workers who worked above-norm hours when they were young (48 or more hours per week), subsequent wage growth was positively correlated with early career hours worked. The same was *not* true for those workers who worked fewer than 48 hours; for those workers wage growth was uncorrelated with hours worked.

Signalling can be particularly dysfunctional in situations in which workers can devote effort to more than one task—an issue we take up next—because it can distort effort allocation. An example is given in the work of Acemoglu, Kremer, and Mian (2008). In their model, career concerns can motivate excessive and misguided signaling by primary school

teachers—misguided because some effort is devoted to improving the signal (student performance on a proficiency test), without actually improving students' true human capital. An important conclusion in that paper is that the problem of excessive signalling can fundamentally shape the desirability of using markets *versus* the government for the provision of some services. Specifically, in a competitive market, there will be socially costly distortions as the result of excessive signalling. An advantage of having teachers employed in the public sector is that the government might be able to commit to policies that eliminate excessive signalling.

3.3. Multi-Tasking

Perhaps the most obvious case of dual-purpose incentives occurs when a principal seeks to regulate an agent's behaviors along more than one dimension. We have already encountered examples along these lines in our discussions above. For instance, in our examination of CEO compensation, we noted problems that arise when a compensation board seeks to create incentives for a CEO to exert effort toward increasing shareholder value *and* limiting wasteful expenditure in the executive suite. Our discussion of tournament incentives suggests a second obvious example: What happens in a tournament when each worker must be motivated to provide effort along his own assigned task *and* be motivated also to be cooperative with other workers?[24] A third example is Acemoglu, Kremer, and Mian's work, mentioned in the previous paragraph, in which teachers allocate effort that improves student human capital *and* effort that merely improves a student's test score ("teaching to the test").

Holmström and Milgrom (1991) establish a number of insightful and surprising results in precisely such contexts. The central point of their paper is both simple and profound: when an agent performs multiple tasks, incentives must perform the double duty of inducing appropriately high levels of effort generally *and* inducing a desirable allocation of an agent's attention across the various tasks inherent in the job.

We can get a feel for their analysis by making a simple extension to our baseline principal agent model. Let us suppose now that the agent can allocate effort along two dimensions, $e_1 \geq 0$ and $e_2 \geq 0$. We suppose also that the agent's utility is now $w + d(e_1 + e_2)$, and,

[24]A number of papers have taken up this issue, including Lazear (1989).

following Holmström and Milgrom, we make two key assumption about the function that gives money metric disutility of effort, $d(\ )$: (i) it is convex and (ii) it achieves a maximum at a positive level of effort. This last assumption means that in the absence of direct incentives the worker will be happiest when putting forth some effort. The principal's objective continues to be the maximization of value added by the worker, now given by $g(e_1, e_2) - w$.

A principal who has reasonably accurate measures of e_1 and e_2, say x_1 and x_2, will be able to construct an incentive scheme in which bonuses reward each dimension of effort appropriately. Matters are more interesting when the principal has good information along one dimension of effort but not the other. To take an extreme example, suppose the firm has a subjective measure x_1, but no measure at all of x_2. The firms best strategy then will depend crucially on the nature of the production function, $g(e_1, e_2)$. To see how this works, we set $d(e_1 + e_2)$ to be $-\frac{1}{2}(e_1 + e_2 - e_B)^2$, with $e_B > 0$ representing the agent's "bliss" level of effort, and work out the optimal incentives for two different production functions: first, a case in which the two types of effort are perfect substitutes, $g(e_1, e_2) = a_1 e_1 + a_2 e_2$; second, a case in which they are complements, $g(e_1, e_2) = e_1 e_2$.

The solution to the first case is easy to characterize. If the firm places any incentive whatsoever on the first type of effort (i.e., a bonus based on x_1), the worker's best response will be $\hat{e}_2 = 0$, and with this in mind the firm can follow the steps outlined for our baseline one-dimensional principal agent model. It is easy to confirm that the result will be to elicit effort $e_1^* = a_1 + e_B$. Intuitively, the firm will prefer this strategy if the value of the first type of effort is high relative to the second type of effort, which is certainly true if $a_1 \geq a_2$. On the other hand, if a_2 is sufficiently large, the principal might decide to use no explicit incentive and instead ask (nicely!) that the worker direct all his effort along the second dimension. Given a binding participation constraint, $w - \frac{1}{2}(e_1 + e_2 - e_B)^2 = v$, it is easily shown that value added by the worker is

$$g(e_1, e_2) - w = \begin{cases} a_1 e_B + \frac{1}{2} a_1^2 - v & \text{when incentives are placed on } e_1, \text{and} \\ a_2 e_B - v & \text{when agent effort is directed to } e_2. \end{cases} \tag{25}$$

As anticipated, if a_1 is sufficiently high relative to a_2, the firm will simply place incentives on the observable portion of performance. This is efficient when $a_1 \geq a_2$, and is second-best optimal even when a_1 is moderately lower than a_2. However, when a_1 is sufficiently low relative to a_2, the firm will instead try a "cooperative" strategy. No incentives are used; the worker is simply asked to direct all effort to the second dimension.

The solution to the second case, in which the two types of effort are complements, is also intuitive. When $g(e_1, e_2) = e_1 e_2$, the firm clearly must avoid a best response of $\hat{e}_2 = 0$, and so will *never* use an explicit incentive along the first effort dimension. In this case the firm instead directs the worker effort to be $e_1 = e_2 = \frac{e_B}{2}$ and hopes that the worker complies.

Simple as this analysis is, several interesting points emerge:

First, we see that there will be cases in which an employer will choose not to place an incentive on an easily-observed dimension of performance, even when that effort is valuable to the firm. This happens when there is a similarly-valuable second dimension of effort that is sufficiently difficult to observe and incentivize. Such an outcome is particularly likely when multiple tasks are complementary. In such cases the firm is best off using very low-powered incentives, i.e., simply paying a base wage.

Second, as shown by our second example, the principal's optimal incentive plan can result in a second-best outcome that is far from efficient. For instance, when $g(e_1, e_2) = e_1 e_2$, it is easily confirmed that the efficient level of output is e_B^2.[25] The firm's low-powered incentive scheme results, instead, in output $\frac{1}{4} e_B^2$. If possible the firm would very much like to find a way to make this worker a residual claimant, and indeed would be willing to suffer substantial cost along some other dimension to make this happen. In short, a strong motive exists here to outsource the task at to an independent contractor if this can be made workable.[26]

---

[25] Maximize $e_1 e_2 - w$ subject to $w - \frac{1}{2}(e_1 + e_2 - e_B)^2 \geq v$.

[26] From a legal perspective, employees are distinguished from independent contractors by the extent of control and supervision the principal exerts over the actions of the agent. A large literature focuses on the forces that drive firm boundaries, focusing on such issues as the direction of employee activities (e.g., Coase, 1937, and Simon, 1951), and firm ownership of assets (e.g., Williamson, 1985, and Grossman and Hart, 1986). The relationship between these issues, and the agency problem—particularly in the multi-tasking setting—is developed clearly in Holmström and Milgrom (1994).

Many papers examine the relationship between firm boundaries and employment relationships in specific industry setting. See, as examples, Arlen and MacLeod's (2005) analysis of physicians in managed care

Third, when a firm cannot use independent contracting, or for some compelling reason chooses not to, a central goal of the firm often entails structuring activities to improve management's ability to closely monitor and supervise key contributions by employees. Put another way, the issue of multi-tasking can matter a great deal for the organization of firm production.

Finally, and most important for our purposes, the multi-tasking approach developed by Holmström and Milgrom, and illustrated with the simple example above, clearly opens the door for "behavioral factors" to play a central role. For example, in our model, the "bliss" level of effort (e_B) is taken to be an exogenous constant. This unfortunate abstraction overlooks the many sociological and psychological factors that determine how intrinsically motivated individuals contribute to the success of their firm. These are the sorts of consideration that lead one into the territory of behavioral economics.

4. Behavioral Approaches to Agency and Motivation

In this section of the paper, we expand the psychological and sociological foundations of agency models. Our focus will be on behaviors that seem to us especially relevant to the understanding of agency—"social" or "other regarding" preferences.

Social preferences arise because people are naturally inclined to compare their own payoffs, sacrifices and behaviors to others, and often care about the impact of their actions on others. Economists have long understood that these "other regarding" preferences are important.[27] Recent progress in the behavioral economics literature has greatly deepened this understanding through the development of new theoretical models and novel empirical investigations using both experimental and observational data. As we discuss below, social preferences matter for agency problems because they offer an explanation for the norms and reference points that agents use to assess their pay, work effort and happiness.[28]

organizations, and the research one of us conducted on worker contracts in the petrochemical industry (Rebitzer, 1995).

[27]For example, several decades back Gary Becker initiated important strands of inquiry in economics by positing preferences that incorporate such factors as "altruism" (Becker, 1981) or "distaste" for interaction with people of a different race (Becker, 1957).

[28]Assessments based on reference points play an important role in behavioral economics generally, included features germane to labor economics. Kahneman and Tversky's "prospect theory" of decision making under uncertainty argues that individuals are loss averse and that they calculate gains and losses relative to (potentially manipulable) reference points (Rabin and Thaler, 2001). Sometimes reference points are

An important insight from the behavioral approach is that the role played by these norms and reference points varies depending on whether one is considering the agency problem in isolation or in a competitive setting.

Our discussion of social preferences in agency considers four distinct but related manifestations of other regarding preferences: pay status, effort norms, professional norms and identity. In each of these sections we begin by sketching a model that makes modest modifications to the standard agency models discussed above. We then consider how well the central predictions of the enhanced model are supported by available empirical analysis.[29]

4.1. Pay Status: Financial Incentives and Inequality Aversion Within Firms

People dislike inequality—especially when they have drawn the short straw. Indeed, there is substantial indirect evidence that wellbeing is shaped in large measure by comparisons with others.[30] This idea can matter within organizations because people are likely to compare themselves with others around them in the workplace. In turn this can be an important determinant in shaping firm compensation policies.

The idea that interpersonal comparisons matter to agents can easily be captured by including "asymmetric inequality aversion" into utility functions. Utility is increasing in

determined by the status quo or by inertia (e.g., Thaler and Sunstein, 2008, and Genesove and Mayer, 2001).

We do not discuss labor supply here, but note that reference points can be important in those models as well. Camerer, Babcock, Loewenstein, and Thaler (1997), for instance, argue that the labor supply of taxi drivers seems to entail drivers evaluating their daily income relative to a daily target. (See also Farber, 2005 and 2008, for additional evidence, some of it to the contrary, and DellaVigna, 2009, for a clarifying discussion.) Fehr and Goette (2007) provide evidence for a field experiment suggesting that loss aversion and reference points may be important in determining work intensity.

[29]Camerer and Loewenstein (2004) provide a nice justification for this approach: "Theories in behavioral economics . . . strive for generality—e.g., by adding only one or two parameters to standard models. Particular parameter values then often reduce the behavioral model to the standard one, and the behavioral model can be pitted against the standard model by estimating parameter values. Once parameter values are pinned down, the behavioral model can be applied just as widely as the standard one."

More generally, Camerer and Lowenstein's paper provides an insightful introduction to a broad and rich set of ideas in behavioral economics, including observations about the origins of modern behavioral economics, and suggestions about future directions for the field. DellaVigna (2009) gives a good recent review of behavioral economics, focusing on evidence drawn from the field.

[30]This is the basis, for example, of the well-known "Easterlin paradox"—the paradoxical results that (i) individuals who are low in a nation's income distribution report themselves to be unhappy, (ii) the average level of unhappiness does not much vary across nations with different levels of aggregate income, and (iii) countries do not get much happier as they get richer. While there is evidence to suggest that absolute income matters for happiness (e.g., Stevenson and Wolfers, 2008), it seems clear that one's standing on the economic totem pole matters as well. Frank's (1985) well known book provides an interesting and wide-ranging discussion on the human inclination for social comparison.

one's own income, of course, but decreasing in the income of other relatively-wealthier comparison individuals. The "asymmetry" refers to a presumption that agents suffer more from inequality that is to their material disadvantage than they gain from inequality that is to their material advantage (see, e.g., Fehr and Schmidt, 1999).[31]

As an example of how asymmetric inequality aversion can affect our agency models, recall the tournament model, as set out in Section 2.2.2. In that model, workers had utility given by $w-e$, and made a net contribution of $g(e)-w$. The firm conditioned pay on an imperfect measure of e: it paid a base wage w_0 to all workers, and in addition gave a bonus b to the fraction $(1-P)$ of workers who had the highest observed performance. (The "bonus" in this case would typically be a promotion to a higher-paying job.) Given workers' best-response effort choices, we saw that this simple tournament resulting in all workers supplying the efficient effort level, i.e., they choose $e^* = \hat{e}(b^*)$ that solves $g'(e^*) = 1$.

Suppose we take that same model but now introduce asymmetric inequality aversion by letting

$$\text{utility} = \begin{cases} w_0 + b - \delta_W b - e & \text{for "winners," and} \\ w_0 - \delta_L b - e & \text{for "losers."} \end{cases} \tag{26}$$

Here, $\delta_W \geq 0$ reflects the possibility that winners feel empathy for losers, proportional to the inequality generated (but of course people do like to win, so $\delta_W < 1$). $\delta_L > \delta_W$ reflects the fact that workers who do *not* win the bonus suffer an even large utility loss due to inequality aversion. Repeating steps outlined in Section 2.2.2, we can show that the principal's solution now has a first-order condition

$$[g'(\hat{e}(b^{**})) - 1]\hat{e}'(b^{**}) = (\delta_L - \delta_W)P > 0, \tag{27}$$

and, since $e'(b) > 0$, we have $g'(e^{**}) > 1$, which in turn means that the firm here settles for a second-best effort level, $e^{**} < e^*$.

Inequality aversion causes the incentive pay parameter to do the "double duty" of eliciting work effort and determining the extent of expected pay inequality in the firm. As a result,

[31]The assertion that utility ins influenced by inequality aversion represents a "stripped down" way of characterizing the behavioral phenomena under study. Our initial treatment makes no distinction about agent attributions concerning the nature of inequality (e.g., what the inequality might say about the principal's intentions or other agents' intentions). We consider more sophisticated approaches below.

the firm must compromise along an important dimension by lowering incentive pay, b, and reducing the effort level elicited from workers. Equation (27) also shows that in this setting the firm will want to be careful how it sets its promotion rate. Here the firm would like (all else equal) to set P near 0, which would allow effort to approach first-best. Intuitively, the cost to the firm of inequality is lowest when there are relatively few people who are affected by the inequality, i.e., when the promotion rate, $1 - P$, is close to 1.

The logic of this model of income comparisons underlies Frank's (1984) seminal article on inequality aversion in labor markets. In his treatment, an employee gains in utility from being high in the firm's pay hierarchy and loses utility from having a low position. Just as in our model, these concerns cause firms to operate with lower-powered incentives. Frank presents evidence drawn from many different types of organizations; he sees, for example, a dampening effect in commissions paid to car salesman and realtors as well as pay compression among college professors.

Encinosa, Gaynor and Rebitzer (2007) present a similar analysis of these issues in the context of incentive pay within medical partnerships. In these professional organizations, physicians determine incentive intensity by choosing how broadly they wish to share the income they generate with others in the practice. For example, groups often choose to share income equally across physicians—in a practice with n physicians, each physician keeps $1/n$ of profits—which minimizes inequality. The practice of equal sharing rules has the potential disadvantage, though, of offering the lowest possible level of incentive intensity to partners. The model of inequality aversion set out by Encinosa, *et al.* shows that the tension between these forces makes sharing rules less attractive in large partnerships than in smaller partnerships—a result supported by the available data. The authors also present evidence consistent the notion that physicians compare effort as well as income. We take up effort comparisons in the next section.

As a second example of the potential impact of inequality aversion in a principal agent setting, consider the multi-tasking model we examined in Section 3.3. Recall that in that model, we assumed an agent's utility is represented by $w - \frac{1}{2}(e_1 + e_2 - e_B)^2$, where e_B is a positive constant, and we assumed further that the principal had a good signal for e_1 but not e_2. So when the principal's payoff is $e_1 e_2$, the best the principal can do is pay the agent

a fixed wage w that meets the participation constraint (i.e., so that $w = v$) and then direct the agent to allocate effort so that $e_1 = e_2 = \frac{e_B}{2}$. Now suppose that the agent is inequality averse, just as in (26), but that in this case her comparison is the *principal's* income, π. So for the agent, utility is

$$w - \frac{1}{2}(e_1 + e_2 - e_B)^2 - \delta(\pi - w), \quad \text{if} \quad w < \pi, \tag{28}$$

where $\delta > 0$ reflects the extent to which the agent is inequality averse. The important point here is that the agent can always costlessly enforce perfect equality by simply adjusting effort allocation (keeping total effort at e_B); she can reduce the principal's profits, while causing no harm to herself. If the principal sets the wage to v, the agent will adjust effort so that π also equals v. So the principal typically finds it profitable to increase w above the participation constraint, i.e., the principal uses "rent sharing." Reducing wages to the level of the employee's outside option would be self-defeating in this context because it risks that the agent will become disgruntled and take steps to "even the score."

The idea that inequity aversion supports rent sharing has been extensively explored in laboratory and field experiments involving variations on the Ultimatum Game. In this bargaining game a *proposer* offers to divide a fixed amount of money between himself and a responder. The *responder* can accept or reject this offer. If the offer is accepted, the money is divided according to the offer. If the offer is rejected by the responder, however, neither the responder nor the proposer gets any money. The conventional game theoretic solution to this bargaining problem for selfish players is for the proposer to make an offer in which he keeps all (or nearly all) of the money while the responder accepts any offer.

It turns out, however, that the participants in these games don't behave as entirely selfish players. Proposers routinely make offers close to an equal division of the pie and responders routinely reject low offers. These anomalies can be resolved, of course, by introducing equity concerns into players' utility functions. Inequality averse proposers get less utility than entirely selfish players do by proposing a division of the pie that greatly favors themselves. Conversely, inequality averse responders can credibly threaten to destroy the surplus if highly unequal divisions are proposed. In practice, reasonably egalitarian offers are made and accepted.

Most employment relationships exist in the context of labor markets. Thus it is not sufficient to demonstrate that individuals prefer more equitable pay practices. Economists must also establish that these preferences matter for the equilibria that emerge in markets. Fehr and Schmidt (1999) examine this central issue by considering whether the egalitarian rent sharing observed in the Ultimatum Game survives in an environment in which there are many proposers and a single responder who must accept or reject the best offer received. They find that competition between proposers leads to lower levels of egalitarianism. To see the logic, consider a situation in which n proposers each offered 50 percent to the responder, leaving each proposer with a $1/n$ chance of having his offer accepted. An individual proposer could clearly do better by simply offering a 51 percent share to the responder, thereby insuring that his offer was selected (the probability increases from $1/n$ to 1). But all proposers are driven by this same logic, and in the end the responder gets all the surplus. This, of course, is the outcome we would observe if proposers had no inequality aversion.

The irrelevance of inequality aversion stems from the fact that with many competitors, no single player can prevent an inequitable outcome. If no individual action can reduce the inequality of the equilibrium outcome, then inequality aversion cannot be an important determinant of behavior. Fehr and Schmidt conclude that matters are different when individual players have a way to impose a cost on the counterparty to a highly unequal offer (as does the agent in the example we consider with equation (28)). Specifically,

> ... competition renders fairness considerations irrelevant if and only if none of the competing players can punish the monopolist by destroying some of the surplus and enforcing a more equitable outcome. This suggests that fairness plays a smaller role in most markets for goods than in labor markets. This follows from the fact that, in addition to the rejection of low wage offers, workers have some discretion over their work effort. By varying their effort, they can exert a direct impact on the relative material payoff of the employer (Fehr and Schmidt, 1999).

In short, agency problems of the sort depicted in our model above (that allows retaliation motivated by equity concerns), can survive market competition.

The Fehr-Schmidt conjecture has been examined experimentally by Fischbacher, Fong and Fehr (2009).[32] They find that increasing proposer competition in the Ultimatum Game,

[32]That paper also provides reference to a large relevant experimental literature.

by adding extra proposers, causes a large increase in mean accepted offers. Similarly, increasing responder competition causes a reduction in mean accepted offers. Both of these results suggest that competition undercuts the influence of equity norms on bargaining outcomes, although in each case the increase in inequality of outcomes is less than one would predict on the basis of competition alone.

The idea that workers exact retribution upon employers when treated unfairly is supported by Bewley's (1999) extensive qualitative interviews. He found that managers and other labor market participants believe that there is a connection between employee morale and performance. Bewley focuses on the morale effects of cutting wages in recessions: Employers are averse to cutting wages because of the fear of a backlash from employees.

Evidence that the fear of backlash is reasonable emerges from a number of recent studies by Mas and co-authors. In a remarkable study, Mas (2006) finds that when New Jersey police officers loose in final offer arbitration, so that the wage they receive is lower than the requested wage, arrest rates and average sentence lengths decline, while crime reports rise. This evidence is consistent with the idea that workers are less inclined to provide effort when the wage falls below a salient reference wage. Krueger and Mas (2004) report evidence that a long and contentious strike and the hiring of replacement workers in a Bridgestone/Firestone plant contributed to the production of defective tires. Mas (2008) finds Caterpillar plants that underwent contract disputes experienced reduced workmanship and reduced product quality. In this latter study he estimates that the contract dispute was associated with at least $400 million in lost service flows due to inferior quality.

The papers cited in the previous paragraph study workers in unionized environments. The presence of a union likely facilitates the sort of collective retaliation that punishes employers who take morale lowering actions. The central idea, that perceptions of unfairness can damage effort, is likely to be important outside the union sector. Evidence along these lines appears, for example, in our field experiment with Nagin and Sanders (Nagin, *et al.,* 2002), which manipulated monitoring levels at call centers collecting donations for charitable causes. In that study, conducted in a non-union environment, we had access to a direct measure of malfeasance on the part of individual employees as well as direct measures of individual employee perceptions about the employer. The measure of employee

malfeasance is the rate at which employees artificially inflated their level of sales in order to earn extra commissions at the expense of the firm. The employer could catch some, but not all, of this activity through costly monitoring of a random sample of calls. When employees worked in centers with very low apparent rates of monitoring, opportunistic behavior increased. However, this increase was observed only for a subset of employees, and, importantly, increased opportunism was most prevalent among among workers who had expressed feelings that the employer treated them unfairly, did not care about them, and provided a bad place to work.[33] Employees who perceived themselves to be unfairly treated struck back at the employer (and added to their own income) when the opportunity to do so arose.

Inequality aversion on the part of employees has a number of interesting ancillary predictions about the way employment relationships are organized. Firms that employ people in both low-wage and high-wage occupations must go to some length to be sure that employees in the low-wage occupation do not include the high-wage occupation in their reference group. Failure to make this separation can lead to pressures to either pay employees in the low-wage occupation too much or employees in the high-wage occupation too little. Indeed, it would not be at all surprising to see firms choosing outsourcing to other firms to avoid just these sorts of invidious comparisons.

Similarly, if employees respond to perceived inequities by retaliating along important, but hard-to-monitor dimensions of work effort and quality, firms that engage in highly unequal pay practices ought to seek out ways to reduce the perceived level of inequality. Secrecy regarding pay is a common human resource practice and it obviously makes invidious pay comparisons more difficult. Some companies, such as Walmart and Lincoln Electric, famously go to great lengths to discourage ostentatious executive perks and the depressing effects on morale they might engender.

For publicly traded corporations the compensation of top executives is a matter of public record. In practice, however, these companies adopt complex and opaque compensation practices that make it difficult to understand exactly how much and in what ways top

[33]These attitude questions were collected in an anonymous survey of employees conducted before the field study began.

executives are paid. Bebchuck and Fried (2003) argue that anomalous features of executive compensation—such as the reliance on "at the money" stock options rather than stock grants—are best understood as efforts to camouflage pay and so avoid "outrage" from shareholders, employees, regulators and other interested parties.

Levy and Temin (2007) examine these outrage costs from an institutional and historical perspective. They argue that the Federal government enforced a set of informal yet egalitarian social norms on executive pay from the post World War II era through the early 1970's. These norms were part of a larger set of institutional arrangements that included powerful unions, high minimum wages, and high marginal tax rates for high earners. The actions taken by Reagan administration in the 1980s (notably the firing of the air-traffic controllers) signaled that the Federal government was leaving such decisions as CEO pay strictly to market forces. The degree to which income norms can be shaped by national institutions and economic policy is an important question that would benefit from additional empirical and historical research.

4.2. Effort Norms

Fehr and Schmidt's key idea is that equity concerns constrain firm behavior even in competitive labor markets, because of behavioral features in agency problems inherent in employment relationships. If, as appears to be the case, employees respond to unequal or unfair treatment by taking actions that punish the employer, it is a short further step to presume that morale enhancing activities ought to motivate employees to take actions in the interest of their employer. This is the premise of Akerlof's (1982) gift exchange model of efficiency wages.

Akerlof's model plays a central role in behavioral labor economics because it relies on very different sociological and psychological mechanisms than the standard agency approach presented in Section 2 above. Instead of engaging in calculations about the costs and benefits of working harder, employees in Akerlof's model are motivated by norms governing behavior in the exchange of gifts. If the employer pays employees a wage higher than some reference wage, the employee perceives himself to have received a gift from the employer. This gift creates an obligation to give something valuable in return. In the employment context, the

obvious way to reciprocate is to provide the firm with more than the minimally acceptable level of work effort and attention.

Akerlof's approach to the problem of agency rests critically on the concept of effort norms, i.e., on the idea that individuals are motivated to provide effort in ways that enable them to conform to their self image or social identity. Decent people, so the reasoning might go, return kindness with kindness and so, wishing to preserve the self image of decency, the employee responds to a high wage by returning the favor in the form of high effort to the employer.[34]

Experimental investigations suggest that reciprocity of the sort identified in Akerlof can survive in competitive environments. For example, Fehr, *et al.* (1998) report results from a laboratory experiment in which sellers have the opportunity to select quality levels above the minimum level enforceable by buyers. In treatments where sellers have the opportunity to do so, they reciprocate high prices with high quality levels. Anticipating this behavior, buyers profit by offering high prices far in excess of the seller's reservation prices. In treatments where sellers do not have the opportunity to reciprocate, buyers offer lower prices.

If employee effort responds to the perceived "fairness" of wage offers, then policy makers must pay special attention to policies that might shift perceptions of the "fairness" of a wage offer. Policy may be especially likely to affect fairness if individuals care about employer intentions as well as outcomes. A wage of X in the absence of a minimum wage might be perceived to be quite fair because employers could have offered a good deal less but choose not to. If, on the other hand, the minimum wage was set to be X, then the employer might have to offer a wage above X to demonstrate good intentions.

In an important paper, Falk, Fehr and Zehnder (2006) investigate whether minimum wage laws influence the perceived fairness of wage offers. They set up an experimental labor market in which individual employees (students paid to participate in the experiment) have to decide whether or not to accept a job offered by a firm. Contrary to the conventional self-interest model, but consistent with a fairness-concerns model, individuals had reservation

[34]This idea is modeled in an insightful way in an important paper by Rabin (1993). Charness and Rabin (2002) provide a clear statement of the ideas, and give reference to further literature. They also give compelling evidence from laboratory experiments on reciprocity. For a discussion of additional experimental evidence, see the chapter by Kuhn and Charness in this *Handbook*.

wages significantly above zero. There was also considerable heterogeneity in reservation wages, giving individual firms an upwardly sloping supply curve for labor. Introducing a minimum wage in this labor market had the effect of increasing individual reservation wages considerably—a result consistent with the hypothesis that the perceived intentions of the firm matter in determining fairness. Surprisingly, there appeared to be hysteresis in the effect of minimum wage laws on reservation wages: subjects exposed to the laws after participating in labor markets with no minimum wage laws increased their reservation wages, but subjects who first participated in labor markets with minimum wages did not revise reservation wages downwards when the laws were removed. These results, if they hold outside the laboratory, have implications that extend far beyond the issue of minimum wage laws. Labor market regulations that influence employer scope of action must take into account how these regulations are likely to affect employee perceptions of employer intentions. More provocatively, the hysteresis result also raises the possibility that regulators might not be able to "undo" the effects of policy simply by reversing previous decisions.

In strong form, well-functioning norms can have considerable social value. They can serve to reduce the problems created by agency in many contexts, including employment relationships within firms.

Given their considerable economic value, it is important to understand the social processes that generate and sustain socially valuable effort norms. In a path-breaking economic analysis, Frank (1988) emphasizes the role of emotions in resolving the "commitment problem," i.e., the problem of eliciting a commitment to constructive cooperation. He argues that rational calculation is often not sufficient to sustain cooperation because by the time the misbehavior occurs, the benefit of punishing the bad actor has often already passed. Emotions, in contrast, can be the foundation of much more powerful sanctions because the commitment to follow through on the action is rooted in the primitive reward structure of the brain. Thus, "cross me and you'll never work in this town again" is a weak deterrent when uttered by a rational calculator who may decide after the fact that it is not worth the effort to punish the double-crosser. It is a strong deterrent, however, when uttered by someone who gets visceral satisfaction in carrying out his threat regardless of the cost to himself.

From a psychological perspective, emotions emerge from a genetically determined neurological reward system. The triggers of this reward system, however, are shaped by an intense and costly socialization process that trains individuals to have a "conscience," i.e., to feel strong emotions when they lie, cheat or otherwise disappoint others' expectations. Evolutionary considerations led Frank to expect that these efforts at socialization will not be entirely successful. Society will be composed of a mixture of types: opportunists who take advantage of chances to free ride and reciprocators who will devote resources to monitoring the behavior of their counterparties and cooperate so long as they perceive others doing the same.[35]

The idea that populations contain a mixture of opportunists and reciprocators is supported by experiments involving public good contribution games. Fehr and Gaetcher (2000) study such games and contrast two treatments. In the "no punishment" treatment, anonymous individuals are randomly allocated to groups of four and are given the opportunity to make contributions towards a public good. Payoffs are such that the dominant strategy is to make no contributions towards the public good. In the "punishment" treatment a second stage is added which gives each individual the opportunity to punish others by subtracting from their payout. Punishing poor contributors is costly, however, and no one interested in maximizing their monetary payoff will choose to punish after the damage is done.[36] For this reason one would expect the dominant strategy to be one of "no contribution" in both punishment and no-punishment treatments. This prediction is confirmed in the "no punishment" game: average contributions converge to almost complete free-riding. In contrast, in games with the option to punish in the second stage, individuals do make substantial

[35]In a population entirely composed of cooperators there will be little reason to devote resources to monitoring the actions of others. This is an environment in which opportunists will thrive. Conversely, in an environment with many opportunists, cooperators will enjoy an advantage so long as they devote resources to monitoring their counterparts. Evolutionarily stable equilibria will therefore involve a mix of opportunists and cooperators with the latter spending resources seeking to weed out or punish the former. See, e.g., Gintis, *et al.* (2003).

[36]The fact that some individuals will punish opportunists even when it is not in their direct material interest suggests that punishment is supported by psychological reward mechanisms rather than rational calculation. Consistent with this view, recent brain imaging studies taken during an economic experiment involving trust and retaliation suggest that punishment of individuals who violate trust activates a brain region, the Caudate, that is involved in actions motivated by anticipated rewards (de Quervain, *et al.,* 2004). High Caudate activation likely reflects the anticipated satisfaction from punishing defectors.

contributions to the public good and these contributions do not fall over time. Consistent with our effort norms model, subjects are more heavily punished the more his or her contribution falls below the average contribution of other group members. Individuals also exhibit heterogeneous tendencies to free-ride and punish. Depending on the definition, the authors estimate that between 20 and 53 percent of subjects in their study were free riders. Heterogeneity in the tendency to behave opportunistically has important implications for labor markets and personnel practices that we explore further in Section 5.

Emotions can support pro-social behavior in ways other than sustaining irrationally high levels of retaliation against defectors. Ekman (2001), for example, argues that emotional states can be read from the facial expressions of individuals. It follows from this that lying and other opportunistic activities that can elicit strong emotions are harder to sustain during face to face interactions. Valley, *et al.* (1998) investigate this hypothesis in a bargaining experiment which requires negotiators to elicit private information about the true value of an underlying asset when the incentives in the experiment do not support revealing this information truthfully. The study finds that face to face negotiations are more likely to reach mutually beneficial solutions than negotiations conducted over the phone or in writing.

The feelings of guilt and shame that support truth telling and honesty are similar to the emotions that support effort norms, and these emotions are generally thought to be strengthened by physical proximity and face to face interactions (Sally, 2002). A nice laboratory experiment by Falk and Ichino (2006) provides evidence along these lines. In particular, in that study the authors observed "peer effects" in which subjects who would otherwise have provided low effort were motivated to increase effort when physically paired with high-productivity workers.[37]

In a remarkable study of cashiers at a national supermarket chain, Mas and Moretti (2008) find that substituting a worker with below average productivity for a worker with above average productivity is associated with a 1 percent increase in the effort of other workers on the same shift. Low productivity workers are especially responsive to the composition of their co-workers and this peer effect occurs only for low productivity workers who are

[37]On net, Falk and Ichino estimate a positive impact on output due to these peer effects similar to estimate of peer effects on absenteeism behavior found in Ichino and Maggi's (2000) study of workers in different branches of a large Italian bank.

in the line of vision of the high productivity workers. The effect of high output peers on the productivity of others declines with distance and with the frequency of interaction as measured by the degree to which shifts overlap.

If effort norms indeed require close proximity and frequent interactions within a work group, it is natural to ask whether these motives can operate in large organizations. Very little empirical work has focused on this important issue.[38]

Effort norms clearly matter within organizations and work groups, and may have important implications for broader labor markets as well. In the gift exchange model, as first set out by Akerlof (1982), the "gift" that results in the optimal reciprocal responses from agents is a wage that exceeds the market-clearing wage. The consequence is equilibrium unemployment (see also Akerlof and Yellen, 1990). If some firms and industries find it important to use gift exchange as a motivating tool, and others do not, then the gift exchange model can be used to explain "dual labor markets," i.e., to help understand cross-firm and cross-industry wage variation.

As Akerlof and Yellen (1985) and Akerlof, Dickens and Perry (2000) argue, the gift exchange logic—that worker performance depends on a firm's current wage relative to a reference wage and to the unemployment rate—can be a building block for macroeconomic models. In Akerlof, *et al.* (2000), for instance, a reference wage model is combined with an assumption that some principals adopt "near rational" wage setting rules whereby they ignore the effect of inflation on reference wages when inflation rates are sufficiently low. The consequence is a long-run Phillips curve with the property that a modest rate of inflation

[38]A nice exception is Knez and Simester (2001). This case study of Continental Airlines finds evidence of the apparent success a firm-wide incentive scheme that paid out a modest sum of money to almost all employees if the airline's aggregate on-time departure statistics cleared a certain threshold. The authors argue that effort norms, enforced by the relatively small and homogeneous ramp and ground crews at each airport, could and did augment the low-powered financial incentives inherent in the bonus plan.

An important question concerns the extent to which norms can persist in cross-functional work groups where the social distance between members of the group might be high. Such work groups—composed of employees with widely different levels of income, status and education—play an important role in the health care system (e.g., in teams that include surgeons and high school educated technicians working together to improve processes). The failure of effort norms and peer pressure to operate in these settings likely contributes to inefficiencies in our health care system (as described, e.g., in Cebul, *et al.,* 2008, and other papers referenced therein).

(approximately 3% in their calibration) is associated with a lower unemployment rate than is either 0 inflation or high inflation.[39]

4.3. Professional Norms

In professions such as law and medicine, the principal agent problem takes on a special importance. Professionals are in theory the agents of their clients, but professionals enjoy advantages of education, credentials, status and specialized knowledge that make their clients especially vulnerable to exploitation. In order to protect clients from abuse, professions go to great lengths to inculcate norms of professional conduct. This makes professions an especially important venue for analyzing the effect of norms.

In our analysis of physician incentives in a managed care organization with Martin Gaynor (Gaynor, Rebitzer and Taylor, 2004), we develop a simple model of professional norms that we adapt here. The model follows the approach used throughout this essay: we modify the agent's utility function, in this instance to include physicians' regard for their patients. We posit, in particular, that the socialization of physicians causes them to experience disutility when they adopt a practice style that delivers medical services that are less than the level that the patient would select for themselves (if they were as well informed as the physician). Think of this level of services, m_B, as the level (measured here in dollars) that results when the physician incorporates a patient's own preferences into his utility function.

We write the utility of a physician treating $i = 1, \ldots, n$ patients as a function of income earned, w, the deviation of medical services, m_i from the ideal level:

$$w + \sum_{i=1}^{n} \mu_i d(m_i - m_B^i), \tag{29}$$

where $d(\ )$ is a convex function that achieves a maximum when m_i is equal to a subjective "best" levels of care, m_B^i, and the μ_i parameters indicate the weight the physician places on each patient's well-being. Thus physician utility is increasing in both income and services provided when they adopt a practice style with $m_i < m_B^i$. In a fee-for-service environment where insurers don't try to "manage" the care physicians provide, one would expect physicians to deliver care at or close to m_B^i.

[39]See also Bewley's (2000) comments on the paper, in which he argues that internal wage comparisons within the firm's wage structure can be a key force in shaping macroeconomic outcomes.

Managed care organizations, such as Health Maintenance Organizations (HMOs), often try to influence physician practice styles through the use of financial incentives. The managed care organization we studied, for example, adopted a simple incentive strategy designed to restrict utilization without substantially harming patients: the principal (HMO) offered agents (physicians) a bonus b if total annual medical expenditures fell below a target $\bar{m}$. The probability that a physician's expenditures on behalf of patients fell below this threshold depended of course on the decisions made on behalf of each patient and also random factors. The probability of earning the bonus, given expenditures m_i and target $\bar{m}$, is given by the c.d.f. $F(\bar{m} - \sum m_i)$.

Giving the physician responsibility for the allocation of resources across a panel of patients in this way makes sense when physicians have practice norms of the sort characterize in (29). We can illustrate the idea easily with the case in which disutility is $d(m_i - m_B^i) = -\frac{1}{2}(m_i - m_B^i)^2$. In this case the physician's best response to a policy, b and $\bar{m}$ is found by maximizing

$$bF(\bar{m} - \sum m_i) - \frac{\mu_i}{2}(m_i - m_B^i)^2. \tag{30}$$

This leads to the best response function for the treatment of each patient ($i = 1, \ldots, n$):

$$\hat{m}_i(b) = m_i^B - \frac{b}{\mu_i} f(\bar{m} - \sum m_i). \tag{31}$$

The extrinsic reward induces the physician to conserve resources on behalf of the HMO, and if μ_i is the same across all patients, say μ (i.e., there is no favoritism), the physician does so in a sensible way.[40] Also, it is easy to confirm that if the second order condition holds, an increase in the bonus induces the physician to reduce expenditures on patients,

$$\hat{m}_i'(b) = \frac{f(\bar{m} - \sum m_i)}{bf'(\bar{m} - \sum m_i) - \mu_i} < 0. \tag{32}$$

From (31) and (32) we can see that the intrinsic value the physician places on patients (represented by μ_i) governs the level of expenditures chosen for patients, as well as the power of the extrinsic incentive to alter chosen expenditures.

In our empirical analysis of internal records in an HMO (Gaynor, Rebitzer, and Taylor, 2004), we found results consistent with the prediction in (32); increased incentive intensity

[40]Indeed, the outcome is potentially efficient.

led physicians to reduce expenditures on patients. We also found that physicians cut costs most for outpatient and elective procedures, but not at all for inpatient procedures. Consistent with the model, this suggests that physicians cut costs most where the consequences for patient welfare were lowest.

Two obvious implications of this treatment of professional norms and incentives merit mention. First, patients are not necessarily harmed by incentives that impose constraints on physician actions. Indeed, it is clearly in the interests of patients for their physicians to allocate resources in a reasonable way, because ultimately patients pay for misallocation through higher insurance premiums. Second, physician practice norms of the sort specified above serve to protect patients from potential abuses introduced by cost-containment incentives, especially if the internalization of patient utility is allocated evenly across patients.

Given the pivotal role of professional norms in protecting clients, it is important that attention be paid to the ways in which these norms are established and how they might be undermined. A key example of the latter phenomenon is professional "conflict of interest." In health care, drug companies famously used gifts and aggressive marketing to influence the prescribing activities of physicians (Avorn, 2004). Conflicts of interest arise in other contexts as well. For example, Jackson (2008) observes that in many financial services markets (including the market for health insurance), brokers who represent one side of the transaction are paid by the other side. These arrangements clearly threaten the ability of brokers to represent the interests of their clients.

Economists have devoted relatively little attention to understanding why practices that create such obvious conflicts of interest persist in markets where principals greatly depend on the independent judgement of professionals. An important exception is a provocative set of articles, Dana and Lowenstein (2003) and Moore and Loewenstein (2004), which argue that even small gifts can trigger a norm of reciprocity that introduces largely unconscious biases into professional judgements.[41] The fact that these biases are unconscious prevents them from inducing the negative feelings that otherwise cause professionals to conform to norms of acceptable behavior. Laboratory experiments suggest that clients who rely on professional

[41]For evidence that judgements of professionals (and others) can be shaped by unconscious but self-serving biases, see Babcock and Loewenstein (1997) and references therein.

judgements do not adequately adjust their interpretation of professional advice even when they are informed that their agents might be biased. Although the evidence for this view of conflict of interest is far from definitive, the implications for the successful resolution of principal agent models in professional settings are both profound and unsettling.[42]

4.4. Identity

Our discussion of professional norms focused on the idea that physicians might experience disutility—perhaps profound feelings of discomfort or anxiety—if they deviate from proscribed behaviors with respect to their clients or patients. This approach to economic sociology is discussed at length in the work of Akerlof and Kranton (2000, 2005) on "identity." Here is the key idea:

> The term *identity* is used to describe a person's social category—a person is a man or a woman, a black or a white, a manager or a worker. The term identity is also used to describe a person's self-image. It captures how people feel about themselves, as well as how those feeling depend upon their actions. In a model of utility, then, a person's identity describes gains and losses in utility from behavior that conforms or departs from the norms for particular social categories in particular situations.
>
> This concept of utility is a break with traditional economics, where utility functions are not situation-dependent, but fixed. In our conception, utility functions can change, because norms of appropriate and inappropriate behavior differ across space and time. Indeed, norms are taught—by parents, teachers, professors, priests, to name just a few. Psychologists say that people can *internalize* norms; the norms become their own and guide their behavior (Akerlof and Kranton, 2005).

The idea that "category" and "situation" can be fundamental elements in preferences enormously expands the range of principal agent models. For instance, to the extent that identity can be manipulated within an organization, identity-based incentives might substitute for extrinsic rewards.[43] Just as families and religious communities undertake important and costly investments to ensure that their children internalize a set of values and practices consistent with passing on the family or group's social identity, so organizations might make

[42]Another largely neglected theme in the economics of professional norms is whether these norms are strengthened or weakened by market competition. Cooper and Rebitzer (2006) argue that competition between HMOs for patients and providers actually magnifies the importance of physician practice norms, and limits the willingness of managed care organizations to control costs via incentive contracts.

[43]For example, practices at West Point are designed to "inculcate non-economic motives in the cadets so that they have the same goals as the U.S. Army" (Akerlof and Kranton, 2005), and firm or workgroup loyalty can be found more generally in many organizations.

investments in practices that persuade employees to adopt goals of the organization, and so mitigate agency problems. These investments are likely to be greatest where financial rewards are most costly to the organization, e.g., when performance measures are especially noisy and where high effort (or high effort at peak times) is critical to the organization's success. Investment in identity incentives will also be greatest where inculcating identity is cheap, and it is reasonable to suppose that imparting identity is cheapest when agents are young and/or when highly motivated individuals self-select into the organization—an issue to which we return in Section 5.

The great virtue of identity models is that they are highly flexible and therefore able to account for behavior that is anomalous from the perspective of simple agency models. This virtue is a curse, however, when it comes to generating falsifiable hypotheses for testing identity models themselves. One way around this problem might be to focus on a particular relevant social category and seek to understand key norms that can be studied systematically and characterized in a parsimonious way.

A template for this latter approach can be found in a series of careful and nuanced investigations of psychological factors that generate *gender* differences in economic behaviors. For example, work by Babcock and her co-authors, demonstrates a profound gender-based difference in the inclination to initiate negotiation; "women don't ask."[44] A simple and clear demonstration emerges in an experimental study in which subjects are asked to complete a simple task, and are then put in a position in which there is ambiguity with regard to the payment. In a typical experiment, subjects were told in advance that the payment would be between three and ten dollars. Then, at the conclusion of the session, the experimenter says, "Here's three dollars. Is three dollars okay?" Eight times as many men as women asked for more money in this experiment. Even in a variant of the experiment in which the experimenter provides cues to signal the social acceptability of negotiation (e.g., with a prompt, "the exact payment is negotiable"), far more men than women take up the opportunity (Small, Gelfand, Babcock, and Bettman, 2007).

[44]Babcock and Laschever (2003) provide an engaging and wide-ranging discussion. The authors present real-world evidence about women's general disinclination to ask, and they include observations about the implications for gender inequality.

In laboratory and field experiments, this disinclination by women to "ask" affects outcomes in negotiated settlements, leading women to do less well than men. Importantly, though, when a woman advocates on behalf of *someone else,* she is typically more successful than when she negotiates for herself, and indeed is generally more effective than men in this capacity (Bowles, Babcock, and McGinn, 2006). Part of the reluctance to "ask," it appears, comes from a desire to avoid self promotion.

Along these same lines, Gneezy, Niederle, and Rustichini (2003) find that women respond differently than men to tournament style incentives when these contests involve both men and women. When paid by piece rate or when competing in single sex tournaments, women's performance is similar to those of men. Niederle and Vesterlund (2007) provide experimental evidence that in comparison to men, women generally shy away from incentives schemes that involve tournament competition.

Gender identity, in short, matters in economically important ways. It is tempting to assert that female identity includes a component that guides women to shy away from competition with men and to reject self promotion. However, it is important to understand that this might not be the whole story, or even the most important part of the story, when using identity to explain gender differences in behavior. There is considerable evidence in psychology that a "kinder, gentler image" is expected of women (to use the expression in Rudman and Glick's 1999 article on the topic). Women who violate that norm by engaging in self promotion face the potential of backlash, which can entail psychological and material costs (as when a woman is bypassed for promotion because she is seen as "inappropriately assertive"). Thus, even a woman who feels no particular disinclination for self promotion might find it in her self interest to adopt the expected "kinder, gentler" norm (Bowles, Babcock, and Lai, 2006).[45]

[45]The point here is that individuals are not passive carriers of their social identities and, as Akerlof and Kranton note, there are many instances in which identities are supported by sometimes severe social sanctions meted out to those whose behavior deviates from proscribed behaviors.

Standard agency models, discussed in Sections 2 and 3, do have implications for gender in labor markets.[46] The recent work by Babcock, Niederlie, Vesterlund, and their co-authors, discussed above, adds a new and promising perspective for understanding the role of gender in organizations and labor markets. Babcock and Lashever (2003) provide extensive evidence that women's reluctance to ask often includes an unwillingness to negotiate their own salaries. It follows logically that in labor markets in which there is rent sharing, this psychological phenomenon contributes to male-female wage and income gaps. On the other hand, the title to Neiderle and Vesterlund's 2007 paper—"Do Women Shy Away from Competition? Do Men Compete Too Much?"—suggests an important point: Cooperation, and the willingness to work hard on the behalf of others, are valuable traits, which should receive positive value in the labor markets. An important research agenda going forward is the incorporation of new findings on gender from psychology and behavioral economics into models of organizations and labor market equilibrium for the purpose of investigating these very issues.[47]

There are certainly other important identity categories that deserve attention from behavioral economists who study organizations and labor markets. *Ethnicity* and *sexual orientation* are additional identity categories that are important in many contexts, including, quite possibly, the labor market.[48] Berman's (2000) economic analysis of ultra-orthodox

[46]For example, as Bulow and Summers (1984) note, if women generally have lower labor market attachment than men (perhaps because they are more likely to withdraw from the market for bearing and raising children or elder care), efficiency wages will be less effective in motivating women than in motivating men. This leads to an equilibrium in which a higher proportion of women than men will end up in the "secondary sector." As a second example, long-hours work norms that emerge in rat race models, such as those of Landers, Rebitzer, and Taylor (1996) might be particularly disadvantageous to women. See Landers, Rebitzer, and Taylor (1997) for a discussion of this latter issue.

[47]In this essay we focus on the effect of gender identity on agency problems, but gender identity is also likely to be very important for understanding female labor supply. In an intriguing and ingenious study, Fernandez, Fogli, and Olivetti (2004), find that married women are more likely to work outside of the home if they are married to a man whose mother worked outside the home. A causal link is established by the use IV estimation, with cross-State variation in male World War II mobilization rates as the instrument.

[48]In the broad social sphere, Bisin and Verdier's (2000) analysis of *ethnic* identity and intermarriage makes a strong prediction that if families value homogamous matches (matches between men and women in the same ethnic group), minority families will make greater investments in identity-preserving activities than majority families, because there is a greater chance that their children will enter heterogamous matches. In a field study of one workplace, Bandiera, Barankay, and Rasul (2009) document workplace favoritism based on nationality (presumably because of social connections between those who share language and national origin) that is costly to the firm. As for *sexual orientation,* it seems possible that when gay individuals take the (possibly very costly) break from powerful expectations to adopt a heterosexual identity and norm,

Jewish groups indicates a strong behavioral impact of *religious identity,* which induces many Israeli ultra-orthodox men to engage in fulltime yeshiva study into their early 40s, thereby impoverishing themselves and their families.

4.5. Miscommunication and Race

Another identity category of indisputable importance is *race.* There is little theoretical work in economics that explores the role of race in organizational form and compensation practices. A very important exception is Lang's (1986) "language theory" of statistical discrimination, which focuses squarely on agency and performance within organizations. The starting point of Lang's analysis is the observation that misunderstanding and misinterpretation are common workplace problems. Lang draws on a wide body of literature in psychology and linguistics to argue that these problems are exacerbated when managers and workers are from different "cultural" or "linguistic" groups.[49]

Following Lang's lead, Ritter and Taylor (forthcoming) consider a labor market in which there is potential for race-based workplace misunderstandings. Their focus is the possibility that this force contributes to black-white gaps in unemployment. The model of unemployment is the agency-based efficiency wage model outlined in Section 2.2.3.

Suppose that most supervisors in the U.S. are white (perhaps because capital is disproportionately in the hands of whites in the U.S.), and that these managers are more successful evaluating the performance of white employees than black employees. Now recall that in the efficiency wage model set out above, σ (the standard deviation of the noise) reflects the precision with which managers evaluate workers. The logic of Lang's arguments leads to the conclusion that σ is relatively higher when white managers evaluate black workers. If so, the unemployment rate will be higher for blacks than for whites.

---

this reduces costs for deviation from traditional norms along other dimensions, such as occupational choice. Along these lines, Black, *et al.* (2000) show that during the Korean War era (1950–1954), military service rates were 12 times higher for lesbian women than other women, and Black, Sanders, and Taylor (2007) show that lesbian college graduates sort into traditionally male majors at substantially higher rates than other women. We know of no work in economics that explores implications for organizations and labor markets.

[49]See also the excellent discussion by Cornell and Welch (1996). The idea that minority individuals might be more difficult to assess than non-minority workers is of course also at the root of the classic work on statistical discrimination. (See, e.g., Arrow, 1998, for references to earlier literature, and a thoughtful discussion.) Austen-Smith and Fryer (2005) provide an additional important perspective.

To see the logic of the Ritter-Taylor result, recall, from (23), that in an economy with homogenous workers, an efficiency wage strategy of worker motivation leads to the following relationship between the equilibrium wage (w^E) and unemployment rate (u^E):

$$w^E = e^* + \frac{1}{\phi(z^*)}\left(\rho + \frac{F^*}{u^E}\right)\sigma. \tag{33}$$

Now suppose that black and white workers are equally productive and thus in equilibrium must be paid the same wage. Suppose also that, as discussed in the last paragraph, there is more noise in the evaluation of black workers than white workers: $\sigma^B > \sigma^W$. Then the following must hold in equilibrium:

$$e^* + \frac{1}{\phi(z^*)}\left(\rho + \frac{F^*}{u^B}\right)\sigma^B = e^* + \frac{1}{\phi(z^*)}\left(\rho + \frac{F^*}{u^W}\right)\sigma^W, \tag{34}$$

where u^B is the unemployment rate for black workers and u^W is the unemployment rate for white individuals. Clearly, $u^B > u^W$. We therefore have a potential explanation for racial differences in unemployment rates.

This model is thus consistent with evidence, such as Neal's (2006), that among men, black-white gaps in the *wage* are small when one conditions on a measurement of human capital taken when the men were youths (the AFQT), but black-white gaps in *unemployment* are large. Ritter and Taylor (forthcoming) show that black-white unemployment gaps persist when one conditions on the AFQT, and that unemployment rates are highest for black men who attended high schools in which other students were mostly black. Under the assumption that these men are most likely to find on-the-job interactions with their boss difficult, this evidence is consistent with the model of racial differences in unemployment we have just outlined.[50]

The "miscommunication model of unemployment" outlined above calls attention to a general point that pertains broadly in models of behavioral agency: If the efficient resolution to agency problems is economically important, than labor markets will tend to reward

[50]Grogger (2009) provides another piece of evidence consistent with the idea that impediments to black-white interactions spill over into the labor market. Even when he controls for skill and family background, blacks with speech patterns that sound distinctively black (according to anonymous listeners) are found to be relatively less successful in the labor market.

Even so, the miscommunication story we have outlined is likely a modest part of the profound racial divide in the U.S., as indicated by the black-white gap in unemployment and labor force participation, as well as many other economic and social dimensions.

individuals who possess scarce preferences and traits that enhance the effectiveness of firms' strategies to evaluate, monitor, and provide incentives. Bowles, Gintis and Osborne (2001) offer a creative assessment of the labor market returns to such incentive-enhancing traits—traits that might include include a low rate of time preference, an intense sense of shame at being without work, perseverance, identification with work goals, and the psychological predisposition to see personal initiative and self determination as important relative to external luck or fate (i.e., to have "internal control" rather than "external control" on Rotter's "locus of control").

Bowles, *et al.* (2001) offer an overview of empirical evidence that wages are correlated with measures of some such traits. For example, people with a high degree of internal control earn higher wages.[51] Of course, correlation does not establish causality, and the Rotter measure may simply stand in for on-the-job productivity. Still, the authors persuasively argue that incentive-enhancing traits can matter for labor market outcomes, and may be important for understanding the large amount of unexplained variation typically observed in estimated wage regressions.

A number of recent theoretical papers in behavioral economics explore the implications of heterogeneity in agent traits along some key dimension (identification with the task, degree to which the agent is pro-social, etc.). In Section 5 below we discuss several of these papers. In each case the distribution of traits as taken to be exogenous—a reasonable approach given the goals in each paper. But, of course, in the broader scheme, many of these key traits are shaped by individuals' home and school environments. This latter point is developed in the seminal work of Bowles and Gintis (1977), who argue that if the education system is to be successful in preparing students for the labor market, the objective function ought to include the development of both cognitive skills and incentive-enhancing behaviors.

5. Dual-Purpose Incentives: Can Pay Destroy Intrinsic Motivation?

In Section 3 we discussed dual-purpose incentives as they arise in conventional models of extrinsic rewards, noting that they arise in many contexts, including the use of compensation

[51]Similarly, Ritter and Taylor (forthcoming) use the Rotter measure as a control in one of their unemployment regressions, finding that men with high internal control (measured when they are young) have lower subsequent unemployment. (The same is not true of women, though.)

practices to signal some otherwise-unobserved characteristic of the firm, to avoid adverse selection of workers along some otherwise-unobserved characteristic, and to deal with multi-tasking. We return in this section to the study of incentives that must do "double duty," but now the second role concerns intrinsic motives.

5.1. Pay and Selection on Dedication

To set the stage, we begin with an extremely simple model in which increasing pay induces adverse selection along an important dimension—dedication to the job. Our case considers a group of workers who are qualified for a particular occupation that for many is a "calling" or "vocation." There are many examples of this sort of dedication: religious ministry, policy advocacy, nursing, early childhood education, public-interest law, etc. The "calling" in this case is a potentially important form of intrinsic motivation. In this section we take a vocational inclination to be an unobserved feature of preferences. In subsequent sections, however, we allow for the possibility that the intensity of an agent's "calling" might be reinforced or eroded by the behavior of the principal or peers.

Specifically, we consider the model set up by Heyes (2005) and analyzed further by one of us (Taylor, 2007). The analysis, which uses the market for nurses as the focus of discussion, begins with simple behavioral assumptions. There are L qualified nurses, each of whom falls into two categories: (i) A proportion of these nurses, $1-\pi$, view nursing as simply a job. These individuals receive utility equal to their wage, w, and they produce value q_L on the job. (ii) The remaining proportion, π, is comprised of nurses who view work as a "vocation." They provide higher-quality nursing along an unobservable dimension, $q_H > q_L$. They also find their work fulfilling, and thus earn money metric utility m beyond the earned wage w.

Each individual has a reservation wage r which is drawn from a log concave p.d.f., $f(r)$, that has a corresponding c.d.f., $F(r)$. The function $f(\ )$ is assumed to be the same for both types of worker. Thus, at wage w, the quantity of nursing labor supplied is

$$\tilde{L}(w) = [\pi F(w+m) + (1-\pi)F(w)]L, \tag{35}$$

and the average quality is of nursing care is

$$\tilde{q}(w) = \theta q_H + (1-\theta)q_L, \tag{36}$$

where θ is the proportion of employed nurses for whom nursing is a vocation, i.e.,

$$\theta = \frac{\pi F(w+m)}{\pi F(w+m) + (1-\pi)F(w)}. \tag{37}$$

Heyes' key insight is that this latter proportion is declining in the wage.[52] Thus, the higher the wage, the lower will be the quality of services provided.

Consider an employer acting in isolation, e.g., a monopsonistic National Health Service (NHS), that hires nurses. (In different markets one could think of the Roman Catholic church setting wages for priests or nuns, or Habitat for Humanity setting wages for professional builders who take part-time positions constructing affordable housing.) Heyes shows that an employer who understands the adverse selection properties of high wages will set the wage to be lower than would otherwise be chosen. Thus, an NHS that maximizes surplus generated by nurses will operate with an apparent "shortage" of nursing, in the sense that the expected net value of product will be positive for the marginal nurse.

It is possible indeed that the principal will be driven to a corner solution, with pay set to zero. Thus, Habitat for Humanity has the well-known policy of using unpaid volunteers for many key tasks. Organization that seek to remedy injustice in the legal system often rely on *pro bono* attorney services. Historically, many religious workers take a "vow of poverty," accepting compensation at near-subsistence levels. The idea, of course, is that the lower the pay, the higher will be the dedication level of individuals willing to adopt the vocation.

Taylor (2007) extends Heyes' analysis to show that a monopsonist that seeks to maximize surplus generated by workers will always set the wage *lower* than the socially efficient level. The reason is that the monopsonist fails to take account of the surplus generated to those individuals who view their work as a vocation. Because wages are too low, too few vocationally-oriented workers end up in an occupation in which they create the greatest social value.

On the other hand, a parallel analysis in Taylor (2007) shows that if the labor market is perfectly competitive, the equilibrium wage will be inefficiently *high.* To see how this happens, notice that under the assumption that all workers must be paid the same wage, a

[52] To see that point, take the derivative of θ with respect to w. The derivative has the same sign as $\frac{F(w+m)}{f(w+m)} - \frac{F(w)}{f(w)}$, which is negative for a log concave function.

social planner would want to maximize

$$\pi\left[(q_H+m)F(w_H+m)+\int_{w_H+m}^{\bar{r}} rf(r)dr\right]+(1-\pi)\left[q_LF(w_L)+\int_{w_L}^{\bar{r}} rf(r)dr\right], \tag{38}$$

where r is taken to be the value of the worker in some other capacity (with $\bar{r}$ being the highest value in the distribution). Maximization of (38) leads to the wage being set to be the weighted sum,

$$w^*=\hat{\theta}q_H+(1-\hat{\theta})q_L, \quad \text{with} \quad \hat{\theta}=\left[\frac{\pi f(w+m)}{\pi f(w+m)+(1-\pi)f(w)}\right]. \tag{39}$$

Next notice that in a competitive market the wage instead will equal average productivity,

$$w^c=\tilde{\theta}q_H+(1-\tilde{\theta})q_L, \quad \text{with} \quad \tilde{\theta}=\left[\frac{\pi F(w+m)}{\pi F(w+m)+(1-\pi)F(w)}\right]. \tag{40}$$

Using the property of log concavity given in footnote 52, we can compare wages in (39) and (40), finding that $w^c > w^*$. The problem with the competitive market is that each firm makes hiring decisions on the basis of *average* market productivity. A social planner would instead make decisions on the basis of the productivity of the *marginal* worker, i.e., would take account of the fact that as the wage increases in the market, the productivity of the marginal worker declines.

This simple model serves as a first illustration of an important point that reappears throughout this section of our essay: Pay policies can affect intrinsic motivation, often in surprising ways. Here, high pay *reduces* intrinsic motivation in a workforce in a particularly transparent way. Low-pay environments attract workers for whom the job is a vocation—workers who have an intrinsic inclination to provide high-quality service. The higher the pay, the greater will be the proportion of workers for whom the job is simply a job, i.e., workers who will provide lower-quality service.

A particularly interesting feature of this simple behavioral model is that markets can lead to wages being either too high or too low relative to an efficient benchmark, depending on the market's structure. To see the logic of this point, consider this question: "If you were a falsely convicted death-row inmate, would you rather be in a State in which you must rely on an organization that reviews cases using *pro bono* attorney services, or in a State that purchases legal services on the competitive market?" In the State that relies on

pro bono services, attorneys who work on death-penalty cases will be highly dedicated to justice, and will provide excellent legal aid, but that aid will be in short supply. In contrast, in a State that purchases legal services for death-row inmates, access to attorneys may be more extensive, but those attorneys will have a lower expected level of dedication. Our model shows that in the State that relies on *pro bono* attorneys, the wage is too low and the quality level too high relative to the efficient benchmark. But in the State that uses the competitive market, the wage is too high and the quality too low. Theory alone does not identify the socially preferred second-best outcome.

Recently, a number of papers have examined models in which agents differ in their level of intrinsic motivation. Delfgaauw and Dur (2007), for example, have a wage posting model in which a monopsonist faces the same tension discussed above: the higher the posted wage, the higher the probability of filling a vacancy, but the lower the expected motivation level of workers who apply. In their model, workers have private information about their utility—information which they may wish to signal to or conceal from an employer. Besley and Ghatak (2005) and Delfgaauw and Dur (2008) study public sector employment under the assumption that some agents have a "public service motivation" that takes the form of intrinsic value derived from making a contribution to one's organization.[53]

The model we have examined in this section omits, obviously, several relevant issues that merit further consideration. For instance, the set-up abstracts from the core problem of agency; workers are simply assumed to supply effort on the basis of their internal intrinsic values. Second, workers are assumed to be *steadfast*; motivation is not affected by the actions of those around them. Thus, a worker who is inclined to provide high quality service is not de-motivated when she is surrounded by others who provide low quality service. In short, the analysis abstracts from "social preferences" of the sort discussed in Section 4. Third, the model does not take account of the possibility that a worker's motivation can be affected by attributions the agent might place on the intentions of the principal. It is

[53]See also Prendergast (2007), who sets out a model in which there is variation in the degree to which agents care about the outcome of some action they might take, as when bureaucrats vary in the extent to which they have altruism and empathy for individuals they are serving. One example he develops concerns social workers hired to determine eligibility for public assistance programs. While a "client-serving ethic" is important for this occupation generally, that same trait may be an impediment for the bureaucratic task at hand.

easy to see that such attributions might be germane, though, in the case of the "service motivations" we have been discussing.[54] Finally, the set-up does not allow for the possibility that a worker's intrinsic motivation (m in the model above) can be reinforced or undermined by pay policies. We turn next to models that take up such issues.

5.2. Social Preferences, Conformism, and the Principal's Use of Extrinsic Rewards

Recent work by Sliwka (2007) considers the question of agency in a model that draws on a "social preference framework," i.e., allows for agents' motivation to be shaped in part on the behavior of those around them. As in the model set out in the last section, there is heterogeneity in worker motivation, and an agent's motivational inclinations are initially hidden to the principal and to other agents. There are two types of "steadfast" agents. One type is "strictly selfish;" these are agents who care only about their own payoff. Agents of the second type are "fair" in the sense that they care about the wellbeing of others; specifically, for these agents, utility is increasing in the principal's payoff. The key innovation is to assume that there is yet a third group, "conformists," whose inclination toward fairness depends the values of those around them. To keep matters simple, Sliwka assumes that when a conformist learns what agent type is in the majority, the conformist behaves like the majority-type agent.

In this set-up, a principal who understands that most of his steadfast workers are fair, might be able to use compensation policies as a credible signal to "conformists." In turn, if conformists believe that others around them are "fair" they behave like fair agents.

To see how this works, we set up a simple example similar to that developed by Sliwka. In our example, the principal first posts a policy that specifies wage as a function of effort (which is assumed to be observable *ex post*), $w(e)$. An agent's best response to the announced policy depends, of course, on his preferences over effort and money, and those preferences in turn vary by type. In particular,

$$\text{utility} = \begin{cases} w(e) - \frac{e^2}{2} & \text{for a steadfast selfish agent, and} \\ w(e) - \frac{e^2}{2} + \mu\pi & \text{for a steadfast fair agent,} \end{cases} \tag{41}$$

[54]For example, when an organization hires motivated agents to pursue some jointly-shared social goal, agents must believe that they indeed are advancing that goal. Presumably, religiously-oriented individuals will be demoralized if they discover that are working for a corrupt church. See, e.g., Besley and Ghatak (2005) for a discussion along these lines.

where μ reflects a fair agent's level of identification with the principal's objective (with $0 < \mu < 1$).

We assume that the principal sets compensation to have a fixed component and a "bonus" that is a linear function of effort, $w(e) = w_0 + \beta e$. Given that types are unobserved, the principal's posted wage-bonus policy applies to all agents. It is easy to see that for an announced compensation policy, best responses are

$$\hat{e}(\beta) = \begin{cases} \beta & \text{for a steadfast selfish agent, and} \\ (1-\mu)\beta + \mu & \text{for a steadfast fair agent.} \end{cases} \tag{42}$$

With this in mind, consider a profit-maximizing principal who earns surplus

$$\pi = e - w(e) \tag{43}$$

for a given agent. Given the best responses in (42), it is clear that effort is increasing in the incentive intensity; $\hat{e}'(\beta) > 0$ for both types of steadfast agent. Sufficiently high effort-contingent bonuses would seem to be in order.[55] Remarkably, it might nonetheless be in the principal's best interest instead to set β equal to 0 and increase the baseline wage, i.e., to rely solely on low-powered incentives.

The key is the emergence of a separating equilibrium in which conformists become convinced that most steadfast agents are *fair*. It is assumed that the firm has private information about the proportion of steadfast agents who belong to each type, which for simplicity is taken to have a *low* value or a *high* value. It is a matter of simple algebra to confirm that there are parameter values for which the following holds: If the firm has a *low* number of fair agents, it pays a high bonus, $\beta > 0$, and a low wage. If the firm has a *high* number of fair agents, it pays no bonus, $\beta = 0$, and a relatively higher wage, i.e., it uses low-powered incentives. Here low-powered incentives serve as a credible signal, so conformists follow suit and behave like fair agents. This makes sense, because a principal who has a *high* fraction of steadfast fair agents will incur a smaller loss than a principal with a *low* fraction of fair agents when it sets the bonus to 0. A willingness by the principal to raise the fixed wage

[55] With perfect information, the bonus would be $\beta_S = \frac{1}{2}$ for a selfish agent and $\beta_F = \frac{1}{2} - \frac{\mu}{2(1-\mu)} < \beta_S$ for a fair agent. Ideally, the principal prefers a larger bonus for a selfish agent, but will want to have a positive bonus for fair agents as well as long as $\mu < \frac{1}{2}$. Notice also that fair agents have "intrinsic motives;" they provide effort even when the bonus is 0.

further strengthens the signal. Conformists, in response to this credible signal, behave like fair agents. Profit for this firm is higher than if it used higher-powered incentives.

The operating logic of the model is like the Taylor-Ritter model discussed in Section 3.1, in which the firm uses compensation policy to signal hidden information about itself (i.e., the firm's financial fitness). In that model, low-powered incentives signal relatively good financial fitness, which allows the firm increased profitability.[56] Here, the principal's hidden information is the mix of worker type. There is a cost to low-powered incentives, of course, as effort is set by all agents to be less than first-best. But the low-powered incentive persuades some workers—the "conformists"—to behave in an altruistic fashion, when they would otherwise have not.

Sliwka interprets his model as generating "trust as a signal of a social norm." In Silwka's setting, the principal observes effort and can, if he chooses, condition rewards on effort. By setting no explicit incentives, the principal expresses trust in his workers. This trust directs a social norm. Some workers are more generous in their efforts than they would have been if they perceived a different norm.

The most intriguing possibility in the Sliwka's model is that *monetary incentives crowd out intrinsic motivation.* If a firm moves from a "high trust" low-powered incentive scheme to a "low trust" high-powered incentive scheme, the firm shifts the norm and undermines the intrinsic portion of worker's motivation (the "social component" in the utility of a worker who would otherwise behave as a "fair agent"). Sliwka develops his theory further by looking at employee self-selection into firms. Here again, low-powered incentives can serve to attract workers with high intrinsic motives (fair agents), which serves to reinforce the positive work norms that influence those who conform to others.

The key behavioral underpinning of the Sliwka model is the observation that many people seem to want to conform to those around them. As we have noted, there is considerable evidence about the key component of this story—that many people are influenced by *norms.* For example, some individuals feel bad about a particular action only in situations in which

[56]Similar logic pertains in Spier's (1992) model, in which a principal knows more about the profitability and riskiness of a project than does an agent, and in Allen and Gale (1992), in which a supplier has superior information about his ability to distort a signal of production costs.

they think others would experience remorse for that same action.[57] In previous sections of this paper, we cited empirical work supportive of the social forces that create norms, e.g., studies by Ichino and Maggi (2000), in which worker absenteeism in an Italian bank was affected by the absenteeism of those around them, and Mas and Moretti (2009), in which effort by supermarket checkout workers was affected by other similar workers in their sightline. Yet another study, by Bandiera, Barankay, and Rasul (2009), shows that the productivity of farm workers is affected by the productivity of friends on the job. Jackson and Bruegmann (2009) document peer learning for teachers, which might be read as providing additional evidence on conformity to norms. Of course, considerably more empirical work will be required to know if conformism plays a sufficiently strong role to generate in real-world organizations the crowding out of intrinsic motivation predicted by the Sliwka model.

5.3. Extrinsic Incentives when Agents Value the Principal's Esteem

Ellingsen and Johannesson (2008) present a model that, like Sliwka's, relies on social preferences. Also, like Sliwka's, their model opens up the possibility that extrinsic incentives can undermine valuable intrinsic motivations.

The key innovation in the Ellingsen-Johannesson model is the postulate that human motivation is often rooted in *social esteem*—the desire to be well regarded by others. In this conception, an agent reasons as follows: "I wish for others to hold a high opinion of me. While I cannot know with certainty what others think of me, I do have beliefs about what others think, and these beliefs about others' opinions are an important source of pleasure or discomfort." The identity of the audience that the agent wishes to impress plays a key role in this model, and the agent might well have multiple audiences. For example, a college professor might care about opinions of her students, her dean, other professors in her department, and/or colleagues in the profession more generally. She desires the *respect* of the intended audience(s), i.e., gains utility if she believes that others think highly of her.

[57]There is a great deal of empirical work across disciplines on norms. One particularly evocative story is told by Fisman and Miguel (2007): When United Nations diplomats in New York were given immunity for parking violations, violations were much higher among diplomats from countries that have high levels of corruption than from countries that have low levels of corruption. This suggests a powerful role for cultural norms. At a theoretical level, work by Bernheim (1994) is important. Fischer and Huddart (2008) discusses the role of endogenous social norms on organizational design.

Ellingsen and Johannesson focus on the case in which the relevant audience to an agent is the principal.[58] To simplify use of personal pronouns, we let the principal be male and the agent be female. In the model, then, the agent's utility depends on the "respect" she earns from the principal, which is defined to be *her beliefs about his beliefs about her.*

A simple example shows how the desire to earn respect can affect an agent's effort decisions.[59] Suppose there are two types of agent, "talented" and "untalented," and type is not initially observable to the principal. An agent hired by a principal is paid an agreed-upon wage w, and then chooses any effort level she likes, $e \geq 0$. We suppose that her utility is the sum of three components: (1) compensation w, (2) the cost of effort, which is $-c_1 e$ for a talented worker and $-c_2 e$ for an untalented worker, with $c_2 > c_1 > 0$, and (3) "respect" of the principal, which has value $rp(e)$, where r is a positive constant and $p(e)$ is the agent's subjective probabilistic assessment of the principal's belief that the agent is talented. In sum,

$$\text{utility} = \begin{cases} w - c_1 e + rp(e) & \text{for a talented agent, and} \\ w - c_2 e + rp(e) & \text{for an untalented agent.} \end{cases} \tag{44}$$

Now, given (44), an *untalented* agent is clearly better off supplying effort 0 than supply effort greater than $\frac{r}{c_2}$, even if the higher effort level would "earn maximum respect" (i.e., would induce the principal to believe with probability 1 that the agent is talented). So we have a separating equilibrium satisfying the Intuitive Criterion if a *talented* agent supplies effort $\frac{r}{c_2}$, which is just high enough so that the *untalented* agent will decline to mimic.[60]

In this example, a talented agent provides positive effort. She is *not* motivated by her own material wellbeing, as she would be in a standard principal agent model. Nor is she motivated by an innate desire to see the principal's wellbeing improve, as with the other-regarding preferences assumed in the Sliwka model (or other such models discussed in

[58]Kandel and Lazear's (1992) important work on peer pressure, in contrast, focuses on the case in which an agent values the regard of co-workers.

[59]The example set out here follows Ellingsen and Johannesson (2007).

[60]For such an equilibrium to exist, we need for the difference between c_2 and c_1 to be large enough to support separate actions by the two types. Suppose it is common knowledge that proportion π of agents are talented. If all agents were to supply $e = 0$, agents would earn respect $r\pi$. But in such a proposed equilibrium, it would be worth it to a maverick talented agent to play $e = \frac{r}{c_2}$, and thereby earn respect r, only if $r - c_1 \frac{r}{c_2} > \pi r$, which boils down to $(1 - \pi)c_2 > c_1$.

Section 4). She provides effort because by so doing she can be confident that the principal holds a high opinion of her. Put another way, she is motivated by *social esteem*—the desire to earn respect.

With this basic logic in place, we can set out the The Ellingsen-Johannesson model of principal-agent interaction. The model is built around three components. First, agents and principals hold *social preferences.* They care about their own material wellbeing as well as well as the material wellbeing of others. Second, there is unobserved heterogeneity in the extent to which agents and principals value others' wellbeing. In particular there are two types on people, who vary in the extent to which they are pro-social. Third, and most distinctively, both the agent and the principal are motivated by *social esteem,* so the agent cares about what the principal thinks about her, and the principal cares about what the agent thinks of him. Both want to be thought of as pro-social by the other. Moreover, the agent's concern about the principal's opinion is highest if she thinks highly of him, i.e., believes it is likely that he is highly pro-social. Similarly, the principal places greater weight on the agent's opinion if he believes that she is highly pro-social.

With these assumptions in place, Ellingsen and Johannesson examine the equilibrium of a game in which the principal takes an initial action (e.g., makes a wage offer, or makes a decision about how much discretion to allow the agent in her work), and then the agent takes an action which affects both her material wellbeing and the principal's wellbeing. As in the simpler example in the preceding paragraphs, there is a set of parameters on preferences and the distribution of types such that a separating equilibrium emerges that satisfies the Intuitive Criterion. In that equilibrium a pro-social principal can take a credible action that signals that he is pro-social, and, if she is sufficiently pro-social, the agent responds with an action that benefits the principal. The key driving behavioral force is that a pro-social agent wishes to be highly regarded by a pro-social principal. Having learned that the principal is pro-social, the agent takes a pro-social action herself as a means of securing the knowledge that the principal believes her to indeed be a pro-social individual.

Figure 2 provides a nice illustration of the behavior predicted in this model. The game presented is the two player sequential "trust game" of McCabe, Rigdon, and Smith (2003):

In Case 1, the first player (the principal) can choose "not trust" (NT), which leads to payoffs (20, 20) for the principal and agent respectively, or "trust" (T), which accords discretion to the agent. If the the principal plays T, the agent can reward the trust (R), giving payoffs (25, 25) or not (N), giving payoffs (15, 30). With conventional preferences, the subgame perfect equilibrium is clearly "not trust." However, in the McCabe-Rigdon-Smith experiments, many principals chose T, and most agents responded by rewarding such trust by playing R.

In Case 2, the principal has no choice but to play T. In contrast to Case 1, here most agents responded by playing N. Thus, the intentionality of the principal's trust appears to matter for the agent's response.

The Ellingsen-Johannesson model provides a rationale for these observed outcomes in the trust game. Start with Case 1. There are parameters in the model such that two key conditions are met. First, the principal will play T only if he is sufficiently pro-social, i.e., only if he cares sufficiently about the wellbeing of the agent. Second, a pro-social agent, having received a credible signal that the principal is pro-social, cares sufficiently about the respect of the principal that she in turn takes the action R, confirming that she is pro-social.

In contrast, in Case 2 the principal has no opportunity to signal that he is pro-social. In turn, the agent cares less about his respect, and so she plays N.

Ellingsen and Johannesson (2008) show, using similar logic, that their model predicts behavior consistent with Falk and Kosfeld's (2006) experimental evidence on the hidden "cost of control" in a principal agent game. In the Falk-Kosfeld game, agents are given an endowment of 120 and can transfer $x \leq 120$ to the principal, who in turn receives payoff $2x$, thus resulting in payoffs $(2x, 120 - x)$. The important twist is that in some conditions of the game, there is a first stage in which the principal can play "control" by imposing a minimum transfer (e.g., a transfer of 10) from the the agent to the principal, or can choose instead to "trust." Agents motivated solely by material gain would always play the minimum available x, and knowing this, the principal would always "control" to the maximum extent allowed. But, in fact, consistent with the Ellingsen-Johannesson set-up, many principals sent a signal of being pro-social themselves by choosing "trust" when they are allowed to do so, and in

such cases many agents responded with larger values of x that if the principal had played "control."

One nice way to see the distinctive contribution of the "esteem model" is to view it in the context of Akerlof's (1982) concept of gift exchange. Recall that in Akerlof's model (discussed above in Section 4.2), an agent's best response to a sufficiently generous "gift" by the principal is to reciprocate by providing high effort. The agent's motivation to do so is captured in a clear, but stripped down fashion—with a utility function in which the agent experiences disutility only when her effort level exceeds a psychologically determined threshold (the effort "norm"). Ellingsen and Johnnesson take an additional step, positing an explicitly specified behavioral mechanism ("esteem") that drives this motivation. This approach, the authors show, allows them to predict gift exchange behavior. But there are two advantages to the Ellingsen-Johannesson model:

First, the model gives a clear way of understanding the role of a principal's *intentions* in shaping agents' responses. This is important, given experimental evidence (such as Charness, 2004) that intentionality is important to understanding gift exchange.

Second, the model provides a rigorous way of approaching an important and under-appreciated aspect of principal agent problems as they apply in the workplace—the delegation of decision rights. Essentially, the delegation of consequential actions to agents plays the role of a "gift" here, and provides the agent with the opportunity to earn the respect of the principal. In contrast, highly intrusive job design diminishes intrinsic motivation.

These ideas are potentially valuable for understanding otherwise-inexplicable practices within organizations. For example, charitable organizations like Habitat for Humanity often rely on volunteers who receive little on no pay, and then delegate key decisions to these same individuals. By providing low pay, we have suggested (in Section 5.1 above), the organization is less likely to attract opportunists. The the "esteem model" shows the important advantages to relinquishing bureaucratic control. A second example is the widespread use of "psychological contracts" (Rousseau, 1995) in which contracting parties find it advantageous to leave many elements unspecified, relying instead on mutual goodwill.

Conversely, the esteem model indicates why high pay and clearly delineated direction might be required in other circumstances. This can happen when the desire for esteem

leads agents to have intrinsic motivation that works at cross purposes to the principal. For example, very strong financial incentives might be required to induce a physician to cut costs if the physician values the esteem of her patients more than the esteem of her boss (e.g., the managed care organization she works for).[61]

5.4. Extrinsic Rewards and Reputation

The Ellingsen-Johannesson model we have just discussed is one of a number of recent papers that focus on the interaction of pro-social motivation and reputation or esteem. This literature starts with the observation that people undertake altruistic and reciprocal actions (in the workplace and elsewhere), but recognizes that such behavior is often difficult to rationalize solely by other-regrading preferences. The degree to which people undertake pro-social behavior often depends on the social context and economic environment, and is driven in part by the desire to be highly regarded by others.

An important contribution to this literature is the recent work of Bénabou and Tirole (2006). Among the remarkable insights of this paper is a clear demonstration that extrinsic rewards can undermine intrinsic motivation when people care about reputation.

Consider the following anecdote: One of us has a particularly personable colleague who was asked by the dean to accept a somewhat onerous task—advising masters students—that would have high value to his colleagues. In exchange, the dean offered a $2000 bonus. The professor replied that it was certainly not worth taking on the task for $2000, but that he would be willing to do the job for free!

To demonstrate how the Bénabou and Tirole approach explains the behavior of this public-spirited professor, we set up a simplified case of their more general model. We suppose that a principal asks agents to undertake a pro-social activity by providing effort

[61]Quite possibly, carefully constructed models of esteem formation can make predictions about the interaction between organizational design and society's class structure. Esteem motives, after all, break down if the principal is not a member of the audience, i.e., the class of individuals whose respect the agent values. Thus, if female identity is formed in a way that leads women to especially value the esteem of men, male bosses would have a distinct advantage over female bosses when supervising women. Bosses in high positions in racial hierarchies, ethnic hierarchies, or other socially determine hierarchies would be similarly advantaged.

$e = 1$. The agent can decline, instead providing $e = 0$. The agent's effort choice is observable by all, including members of an audience who's opinion matters to the agent.[62]

The agent's effort decision is assumed to affect his utility via four channels, which for simplicity are taken to be additive: (1) The agent is other-regarding, and so earns direct utility, $v_e e$, from providing effort e; (2) he stands to earn a material reward in the form of a bonus of $b \geq 0$, which provides utility $v_m be$ (where v_m is the marginal utility of money); (3) he faces an effort cost of ce; and, most distinctively, (4) he stands to gain from the reputation-enhancing effect of his effort choice.

Central to the Bénabou-Tirole model are the following two assumptions, which drive development of "reputation." First, people differ in the extent to which they have other-regarding inclinations, and in the extent to which they value money. So each individual's set of preference parameters, v_e and v_m, is drawn from a known distribution.[63] An agent's preference type can be thought of as his or her "identity." Second, reputation is taken to be other's views of one's own identity. This reputation is *increasing* in the degree to which one is seen as having concern for others (having a high value of v_e) and decreasing in the degree to which he is seen as materialistic (having a high value of v_m).[64] Thus, reputation is taken to be $R(e, b) = \mu_e E[v_e|e, b] - \mu_w E[v_m|e, b]$, where μ_e and μ_w are weights that reflect the degree of image-consciousness (and are taken to be common knowledge constants here).

To summarize,

$$\text{utility} = \begin{cases} v_e + v_m b - c + R(1, b) & \text{if } e = 1 \text{ and} \\ R(0, b) & \text{if } e = 0. \end{cases} \tag{45}$$

So our agent provides effort if

$$v_e + v_m b + [R(1, b) - R(0, b)] > c. \tag{46}$$

[62]To simply matters, suppose that the agent is already in the principal's employ and is now being asked to undertake a task that was not originally part of the job (as in the example of the public-spirited professor). It would be a worthwhile task to apply the Bénabou-Tirole model in a labor market generally (which would require attention to participation constraints, and to the way in which an announced compensation policy might affect selection into a firm).

[63]To keep things simple here, we suppose that the parameters are independent.

[64]Concern for others ("kindness") and moderation in materialistic pursuit ("temperance") are but two of the seven virtues. By leaving the other five virtues unstudied, perhaps Bénabou and Tirole signal "patience" (leaving them for future work) and "humility" (deference to other behavioral economists who wish to study those virtues).

The three terms on the left-hand side of (46) are, respectively, the agent's intrinsic, extrinsic, and reputational motivations. Effort is provided when the sum of these motivations exceeds the cost of providing effort.

What makes matters interesting here is that the reputation the agent earns depends on the level of the bonus b chosen by the principal. To see this point, consider first the effort decision if the principal chooses $b = 0$, so that motivation is strictly intrinsic and reputational. Then the agent provides effort only if his "concern for others" exceeds the cut-off $\hat{v}_e$, where

$$\hat{v}_e \equiv c - [R(1,0) - R(0,0)]. \tag{47}$$

Such highly pro-social identity types lie to the right of the vertical "No Bonus" line in Figure 3. Notice that in this situation the agent's audience can draw an informative inference about the agent's v_e (i.e., can infer if the agent is in an other-regarding pool to the right of $\hat{v}_e$ or narcissistic pool to the left of $\hat{v}_e$) simply by observing the effort level, but can learn nothing about his materialism v_m.

Now suppose that the principal provides a bonus b. For the moment, ignore any impact on reputation. The bonus has no effect on an agent with $v_m = 0$, of course, but for all other agents (those with $v_m > 0$) the bonus is motivating. Agents now provide effort if they have an identity that lies to the right of the negatively sloped line marked "Positive Bonus ($b > 0$), Unadjusted for Reputation." Absent reputational effects, an extrinsic reward expands the pool of agents providing effort.

However, this is not the end of the story. Inspection of Figure 3 makes it is clear that the average level of v_e (concern for others) has declined in the pool of agents providing effort and also in the pool of agents not providing effort. It is similarly clear that the average level of v_m (greed) has risen for the pool of those providing effort and declined for the pool not providing effort. So the overall effect on reputation is ambiguous. The most interesting possibility is that the effect of the bonus is to drag down overall reputation for those providing effort. This is demotivating. In Figure 3 this is illustrated by the parallel shift to the right in the sloped line dividing those who provide effort and those who do not, with $\tilde{v}_e$ now giving the value of v_e that separates the two pools for individuals along the

horizontal axis (i.e., for individuals with $v_m = 0$). On net, effects of the bonus are two-fold: Some identity types—those in Area A of Figure 3—switch behavior to providing effort. Others—those in Area B—are induced to switch from effort provision to *not* providing effort. Overall, an extrinsic reward can increase or decrease effort, depending on the distribution of identity types.

The public-spirited professor in our anecdote is apparently an individual with an identity of the sort represented by Area B. Such a person has a relatively high concern for others and a relatively low level of greed. By offering a bonus for the task, the dean deprived the professor of the opportunity for the professor to demonstrate his public spiritedness.

In short, in the Bénabou-Tirole model, extrinsic rewards can spoil the reputational value pro-social action, thereby crowding out intrinsic motivation. The logic of the model is one of "signal extraction." People take pro-social actions in part to signal one's own identity to others. Extrinsic rewards, even very small extrinsic rewards, can serve to increase the noise-to-signal ratio of such actions.[65]

One of the most interesting ideas in the Bénabou and Tirole model appears when the authors reinterpret the "reputational" terms in (46) to instead be the reinforcement of one's own *self image.* The idea is described as follows:

> When making a decisions affecting others' welfare, an individual will often engage in a self-assessment: "How important is it for me to contribute to the public good? How much do I care about money? What are my real *values*?" Later on, however, this information may no longer be perfectly "accessible" in memory—in fact, there will often be strong incentives to recall it in a self-serving way. Actions, by contrast, are much easier to remember than their underlying motives, making it rational to define oneself partly through one's past choices: "I am the kind of person who behaves in this way" (Bénabou and Tirole, 2006).

Thus, the public-spirited professor might have found the offer of a $2000 bonus to be demotivating even if his colleagues were unaware of the bonus. Accepting the task without pay, in this conception, served to reinforce his identity; the professor can look at himself in

[65]Indeed, in more general versions of their model, Bénabou and Tirole show that extrinsic rewards can reverse the sign of the signal! Armed with this logic, the authors establish interesting and surprising insights around the use of non-monetary motivators as *praise* and *shame.* They show, for example, that the excessive use of praise can backfire if pro-social behavior "becomes suspected of being motivated by appearances." They also study the equilibrium development of social norms.

the mirror and argue convincingly, "I must be a pro-social person. Otherwise I wouldn't have taken on this task with no pay."[66]

The striking prediction that extrinsic rewards can crowd out desired behaviors does have empirical support. One widely cited example in economics is work by Gneezy and Rustichini (2000b) showing that the imposition of a monetary penalty for late child pick-up at a daycare center increased the likelihood of late pick-up. Bénabou and Tirole's reputation/self-respect model strikes us as applicable here. When there was no explicit monetary penalty for on-time pick-up, parents were presumably motivated by genuine concern for daycare center workers and by a desire to project to others (or to oneself) character traits of *responsibility* or *concern for others*. The imposition of a monetary penalty of course increased the inclination for on-time pick-up among those parents with a materialistic orientation and a low level of concern for others. This very fact led parents more generally to no longer view on-time pick-up as a reliable signal of kind and responsible identity, and so reduced the strength of those motivating forces. Similar arguments apply to Gneezy and Rustichini's (2000b) demonstration that extrinsic incentives reduced effort by school children collecting donations for a charitable organization.[67]

Evidence of a potentially important form of crowding out is also found also in Frey and Oberholzer-Gee's (1997) analysis of public reaction to the siting a nuclear waste facility in one's community. Their paper indicates that the provision of substantial compensation to residents of a host community reduces willingness to accept such a facility. The reputation/self-respect model might speak to these results, but it is quite possible that mechanisms described in Bénabou and Tirole (2003) are more germane. In that paper the authors set up a problem in which a principal seeks to motivate an agent to take a desired action in an environment in which the principal has better information than does the agent about some crucial aspect of the task—for example, the cost the agent will incur if she undertakes the task, the personal satisfaction she will experience if the task is completed

[66]Bénabou and Tirole (2006) note that the key idea—"that individuals take their actions as diagnostic of their preferences"—is found in psychology in Bem's (1972) *self-perception theory* and is related to the Festinger and Carlsmith's (1959) theory of *cognitive dissonance.*

[67]Frey and Jegen (2001) provide further reference to the literature, and discuss crowding effects from an economic perspective.

successfully, or the likelihood that the agent will indeed successful at the task. In this setting, the offer of a substantial monetary reward for some action can signal "bad news" to the agent about one of the elements of the action. Thus, if residents of a potential host community are asked to site a nuclear waste facility *and* are offered substantial compensation for doing so, that compensation might be seen as "bad news" about the eventual consequences of the facility.

The two papers we highlight in this section of our paper, Bénabou and Tirole (2003 and 2006), are but two of a number of recent contributions in behavioral economics that might form solid building blocks for a new generation of behavioral principal agency models.[68] The challenge going forward is to place the psychological subtleties introduced in these new economic models into workable (and testable) theories of firm organization and labor markets. It is important to have carefully constructed, psychologically correct models in behavioral economics, but important also to work forward to understand the implications of these models for the allocation of resources in markets and in society broadly.[69]

5.5. A Concluding Puzzle

The economic approach to agency places a primary emphasis on the use of material incentives (pay, promotion, etc.) as devices to resolve principal agent problems. The economic literature offers a rich and varied set of evidence in support of the critical efficacy and importance of material incentives. The theoretical literature reinforces these empirical findings. There are many situations in which firms eschew high-powered incentives, but for the most part this is the result of incentives having a powerful effect on behaviors. It is possible, as we have seen, to construct models where extrinsic rewards undermine intrinsic motives, but these models appear largely as elaborations and qualifications of the fundamental message:

[68]Among the many other potentially relevant examples are the models of social image in Bernheim and Severinov (2003) and Andreoni and Bernheim (2009). The Bernheim-Severinov model is designed to explain the common practice of equal division of bequests. The model posits that children care about the extent to which they are loved relative to other siblings, and then studies bequests as a mechanism by which parents can signal love. Equilibrium behavior tends to pool at equal bequest division. Similar logic might explain the frequent organizational practice of equality in treatment (pay, work conditions, etc.) of workers who might differ quite widely in productivity. Andreoni and Bernheim's refinement of these ideas might serve as a valuable microfoundation for studying the role of fairness (e.g., Fehr and Schmidt, 1999) in principal agent relationships.

[69]The beautiful work of Akerlof and his co-authors—in papers on the economic implications of reciprocal motives, cognitive dissonance, social distance and identity—provides a template in this regard.

well designed extrinsic rewards are crucial to the resolution of fundamental and ubiquitous agency problems.

Things are quite different in the field of psychology. Here there has accumulated a vast amount of evidence that extrinsic rewards actually undermine intrinsic motives.[70] What explains the difference?

One important part of the explanation is a cross-disciplinary difference in the definition of *intrinsic motives.* Psychologists typically view as extrinsic any sort of action undertaken for instrumental reasons. Thus many of the pro-social and other regarding preferences we discuss in Sections 4 and 5 would be regarded as part of an extrinsic reward system in the psychology literature.[71]

Economics is concerned with the efficient use of society's material resources. In societies characterized by specialization and a sophisticated division of labor, almost all economic activity involves some degree of instrumental motives. Thus by defining the notion of intrinsic rewards so narrowly, psychologist's have restricted their attention to a very small subset of economically relevant behaviors.

The focus, as peculiar is it might appear from an economist's perspective, makes perfectly good sense from the perspective of psychology. After all, psychology is concerned with understanding the reward structures that drive human behavior. Why then should psychology privilege economically relevant motives?

---

[70]Important theoretical constructs include Lepper, Greene, and Nisbett's (1973) *overjustification theory* and Deci and Ryan's (1985) *self-determination theory.* A large number of carefully constructed experiments provide evidence favoring these theories, including many that demonstrate crowding out of intrinsic motivation (e.g., Deci, Koestner, and Ryan, 1999).

[71]For instance, in Ryan and Deci's (2000) taxonomy, *intrinsic motivation* is reserved for "the doing of an activity for its inherent satisfactions rather than for some separable consequences." *Extrinsic motivation,* on the other hand, "pertains whenever an activity is done in order to attain some separable outcome." Such extrinsic motivation includes "external regulation" with a material reward or punishment, but also includes "introjection," which focuses on approval from others or from oneself, and also, remarkably, "integrated regulation," which occurs when an agent comes to assimilate the external driver as an internal driver. To quote Ryan and Deci (2000), "The more one internalizes the reasons for an action and assimilates them to the self, the more one's extrinsically motivated actions become self-determined. Integrated forms of motivation share many qualities with intrinsic motivation, being both autonomous and unconflicted. However, they are still extrinsic because behavior motivated by integrated regulation is done for its presumed instrumental value with respect to some outcome that is separate from the behavior, even though it is volitional and valued by the self."

There is another important difference in the ways that psychology and economics analyze extrinsic and intrinsic motivation: the handling of *autonomy.* Deci, Ryan, and other psychologists argue that feelings of autonomy and competence are fundamental to human happiness. To the extent that they cause people to become accustomed to responding to rewards rather than their own intrinsic drive for self-realization, extrinsic rewards undermine a fundamental determinant of psychological wellbeing.[72]

Positive economics, in contrast, conceives of autonomy simply as a means for achieving some productive end. For example, in standard principal agent models, high levels of autonomy are warranted when an agent has better information than does the principal about the consequences of actions, and can be rewarded on the basis of the value created by selecting the best action from a choice set. Even in Ellingsen and Johannesson's esteem model, autonomy awarded by the principal serves the instrumental purpose of allowing the agent to signal valuable information to the principal. Economists have only begun to explore the interesting and provocative possibility that autonomy has value in and of itself and that the use of targeted extrinsic rewards (in the psychological sense) undermines an individual's feeling of autonomy and competence.[73]

6. Conclusions

Our purpose in writing this chapter is to assess the contribution of behavioral economic ideas to the study of agency in employment relationships. In Section 2 we introduce the basic logic of standard agency models and in Section 3 we discuss the complications that arise when incentives must serve "double duty" as is the case where firms have to worry about adverse selection or multi-tasking. In Section 4 we introduce the core behavioral idea of "other regarding preferences" and consider effects on agency relationships of various

[72]Deci and Ryan's theory of self-determination theory, for example, emphasizes the innate psychological needs for a sense of *competence* and *autonomy.* The authors suggest that "interpersonal events and structures (e.g., rewards, communications, feedback) that conduce toward *feelings of competence* during action can enhance intrinsic motivation for that action because they allow satisfaction of the basic psychological need for competence. Accordingly, for example, optimal challenges, effectance promoting feedback, and freedom from demeaning evaluations are all predicted to facilitate intrinsic motivation" (Ryan and Deci, 2000).

[73]See, e.g., Benz and Frey's (2008) research on the value of independence and Dur and Glazer's (2008a) work on the desire by workers for impact.

manifestations of these preferences—equity considerations, effort norms, norms of professional practice and identity. In Section 5 we return to the theme of double duty incentives, and consider the possibility that incentives have the two-fold effect of motivating desired behaviors while also reinforcing (or undermining) intrinsic motives.

The narrow focus of our paper has caused us to give short shrift to many important contributions that behavioral economics has made to our discipline. We say relatively little about such important behavioral economic topics as prospect theory, hyperbolic discounting, mental accounting, status-quo biases and default rules, cognitive dissonance, or bounded rationality. Perhaps more noteworthy than the behavioral issues we have left out of this essay are the standard methodological approaches that we have kept in. Our intention has been to remain theoretically grounded and methodologically conservative. In each section of the paper we represent purposive behavior by analyzing equilibrium behaviors that emerge when individual agents maximize a utility function subject to participation constraints and the constraints imposed by incentive and monitoring systems. Also, consistent with standard economic analysis, we are careful to consider the ways in which equilibrium outcomes are shaped by market competition and by the selection of agents into employment relationships.

Even with this deliberately conservative approach, we find that the introduction of behavioral features into agency models leads to novel and important results: Inequity aversion among agents leads to lower powered incentives than would otherwise be the case, but this effect can be undone in certain competitive environments. Effort norms and "gift exchange" can support high effort levels even when monitoring and incentives are problematic, but reliance on effort norms requires that principals be exquisitely attuned to the ways in which their actions influence employee morale. Professional norms can have the effect of protecting consumers from exploitation by professionals and this effect can be reinforced by properly designed incentives. The protective value of these norms can, however, be undermined by self-serving biases that distort the judgement of professionals in unconscious ways. Identity matters for the resolution of agency problems within employment relationships and can help explain important empirical anomalies in labor markets. High powered extrinsic incentives

can have a corrosive effect on the motivation of employees, especially when the employees work in "mission driven" or "caring" organizations or when preferences or identity are endogenously shaped by the incentives to which employees are exposed.

The application of behavioral economics to agency in employment relationships is a relatively new area of research. It is worthwhile then to speculate on what might be especially promising areas for future research. We highlight four such areas:

First, given the pivotal importance of professional norms for well functioning markets in health care and financial services, we think it would be useful to investigate more thoroughly the behavioral foundations of conflicts of interest. Very little is known about the ways these conflicts shape the psychology of decision making, and having a clearer understanding of this issue may be quite important for designing efficient and effective regulatory policies.

Second, models of identity have a great deal of appeal, because families, schools, and firms appear to devote enormous resources to shaping and refining the identity of their participants. As currently specified, however, models of identity are so flexible that they may not generate falsifiable conclusions. A satisfactory understanding of the economics of identity will therefore require either a more structured modeling approach or, more likely, the accumulation of additional sociological and psychological data on the nature of identity so that the parameters of the models can be empirically constrained.

Third, much more needs to be learned about the relationship between public policy and income and effort norms. Are Levy and Temin, for example, correct in their assertion that changes in Federal government policy in the 1980's shifted the tolerance for income inequality throughout the labor market? Are Akerlof, Dickens and Perry correct that the effectiveness of monetary policy is determined by the workings of reciprocity and gift exchange in the workplace? At present we do not have definitive answers to these questions.

Finally, although the theory is new and the evidence not yet conclusive, we are intrigued by the notion that extrinsic rewards can undermine intrinsic motives. In health care, corporate governance, education, and other important settings, standard models typically prescribe some sort of of "pay for performance" for resolving agency issues. This prescription must be greatly modified if we can identify people and contexts where high powered financial incentives undermine employee motives to do the right thing.

Clearly, there is much more to discover about the behavioral economics of agency in employment relationships.

References

Acemoglu, Daron, Michael Kremer, and Atif Mian (2008), "Incentives in Markets, Firms, and Governments," *Journal of Law, Economics, and Organization*, 24(2): 273-306.

Akerlof, George A. (1976), "The Economics of Caste and of the Rat Race and Other Woeful Tales," *Quarterly Journal of Economics,* 90(4): 599-617.

Akerlof, George A. (1982), "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics,* 97(4): 543-569.

Akerlof, George A. (1984), "Gift Exchange and Efficiency-Wage Theory: Four Views," *American Economic Review,* 74(2): 79-83.

Akerlof, George A. and William T. Dickens (1982), "The Economic Consequences of Cognitive Dissonance," *American Economic Review,* 72(3): 307-319.

Akerlof, George A., William T. Dickens, and George L. Perry (2000), "Near-Rational Wage and Price Setting and the Long-Run Phillips Curve," *Brookings Papers on Economic Activity,* 2000(1): 1-44.

Akerlof, George A. and Rachel Kranton (2000), "Economics and Identity," *Quarterly Journal of Economics,* 115(3): 715-53.

Akerlof, George A. and Rachel Kranton (2005), "Identity and the Economics of Organizations," *Journal of Economic Perspectives,* 19(1): 9-32.

Akerlof, George A. and Janet L. Yellen (1985), "A Near-Rational Model of the Business Cycle, with Wage and Price Inertia," *Quarterly Journal of Economics,* 100(5): 823-838.

Akerlof, George A. and Janet L. Yellen (1990), "The Fair Wage-Effort Hypothesis and Unemployment," *Quarterly Journal of Economics,* 105(2): 255-283.

Allen, Franklin and Douglas Gale (1992), "Measurement Distortion and Missing Contingencies in Optimal Contracts," *Economic Theory,* 2(1): 1-26.

Allgulin, Magnus and Tore Ellingsen (2002), "Monitoring and Pay," *Journal of Labor Economics,* 20(2 part 1): 201-216.

Andreoni, James and B. Douglas Bernheim (2009), "Soial Image and the 50-50 Norm: A theoretical and Experimental Analysis of Audience Effect," *Econometrica,* 77(5): 1607-1636.

Andreoni, James, William Harbaugh and Lise Vesterlund (2003), "The Carrot or the Stick: Rewards, Punishments, and Cooperation," *American Economic Review,* 93(3): 893-902.

Arlen, Jennifer and MacLeod, W. Bentley (2005), "Torts, Expertise, and Authority: Liability of Physicians and Managed Care Organizations," *RAND Journal of Economics,* 36(3): 494-519.

Arrow, Kenneth J. (1998), "What Has Economics to Say about Racial Discrimination?" *Journal of Economic Perspectives,* 12(2): 91-100.

Austen-Smith, David and Roland G. Fryer, Jr. (2005), "An Economic Analysis of 'Acting White,"' *Quarterly Journal of Economics,* 120(2): 551-83.

Avorn, Jerry (2004), *Powerful Medicines: The Benefits, Risks and Costs of Prescription Drugs.* New York: Alfred A. Knopf.

Babcock, Linda and Sara Laschever (2003), *Women Don't Ask: Negotiation and the Gender Divide.* Princeton and Oxford: Princeton University Press.

Babcock, Linda and George Loewenstein (1997), "Explaining Bargaining Impasse: The Role of Self-Serving Biases," *Journal of Economic Perspectives,* 11(1): 109-26.

Babcock, Linda, George Loewenstein, Samuel Issacharoff and Colin Camerer (1995), "Biased Judgments of Fairness in Bargaining," *American Economic Review,* 85(5): 1337-43.

Babcock, Linda, Xianghong Wang, and George Loewenstein (1996), "Choosing the Wrong Pond: Social Comparisons in Negotiations That Reflect a Self-Serving Bias," *Quarterly Journal of Economics,* 111(1): 1-19.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2005), "Social Preferences and the Response to Incentives: Evidence from Personnel Data," *Quarterly Journal of Economics,* 120(3): 917-962

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2007), "Incentives for Managers and Inequality Among Workers: Evidence From a Firm-Level Experiment," *Quarterly Journal of Economics,* 122(2): 729-773

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2009), "Social Connections and Incentives in the Workplace: Evidence from Personnel Data," CEPR Discussion Paper.

Bebchuk, Lucian and Jesse M. Fried (2003), "Executive Compensation as an Agency Problem," *Journal of Economic Perspectives,* 17(3): 71-92.

Bebchuk, Lucian and Yaniv Grinstein (2005), "The Growth of Executive Pay," *Oxford Review of Economic Policy,* 21(2): 283-303.

Becker, Gary, S. (1957), *The Economics of Discrimination,* Chicago: University of Chicago Press.

Becker, Gary S. (1981), *A Treatise on the Family,* NBER Books.

Bem, Daryl J. (1972), "Self-Perception Theory," in Leonard Berkowitz, *ed., Advances in Experimental Social Psychology,* Vol. 6. New York: Academic Press.

Bernheim, B. Douglas (1994), "A Theory of Conformity," *Journal of Political Economy,* 102(5): 841-877.

Bernheim, B. Douglas, and Sergei Severinov (2003), "Bequests as Signals: An Explanation for the Equal Division Puzzle," *Journal of Political Economy,* 111(4): 733-764.

Bénabou, Roland, and Jean Tirole (2002), "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics,* 117(3): 871-915.

Bénabou, Roland, and Jean Tirole (2003), "Intrinsic and Extrinsic Motivation," *Review of Economic Studies,* 70(3): 489-520.

Bénabou, Roland, and Jean Tirole (2006), "Incentives and Prosocial Behavior," *American Economic Review,* 96(5): 1652-1678.

Benz, Matthias and Bruno S. Frey (2008), "Being Independent is a Great Thing: Subjective Evaluations of Self-Employment and Hierarchy," *Economica,* 75: 362-383.

Berman, Eli (2000), "Sect, Subsidy, and Sacrifice: An Economist's View of Ultra-Orthodox Jews," *Quarterly Journal of Economics,* 115(3): 905-953.

Bertrand, Marianne, and Sendhil Mullainathan (2001), "Are CEOS Rewarded for Luck? The Ones without Principals Are," *Quarterly Journal of Economics,* 116(3): 901-932.

Besley, Timothy and Maitreesh Ghatak (2005), "Competition and Incentives with Motivated Agents," *American Economic Review,* 95(3): 616-36.

Bewley, Truman F. (1999), *Why Wages Don't Fall During a Recession,* Cambridge, MA: Harvard University Press.

Bewley, Truman F. (2000), "Comments and Discussion," *Brookings Papers on Economic Activity,* 2000(1): 45-50.

Bisin, Alberto and Thierry Verdier (2000), "'Beyond the Melting Pot': Cultural Transmission, Marriage, and the Evolution of Ethnic and Religious Traits," *Quarterly Journal of Economics,* 115(3): 955-88.

Black, Dan A. Gary Gates, Seth Sanders, and Lowell J. Taylor (2000), "Demographics of the Gay and Lesbian Population in the United States: Evidence from Available Systematic Evidence," *Demography,* 37(2): 139-54.

Black, Dan A. Seth Sanders, and Lowell J. Taylor (2007), "The Economics of Lesbian and Gay Families," *Journal of Economic Perspectives,* 21(2): 53-70.

Bowles, Hannah Riley, Linda Babcock, and Lei Lai (2006), "Social Incentives for Gender Differences in the Propensity to Initiate Negotiations: Sometimes it Does Hurt to Ask," *Organizational Behavior and Human Decision Processes,* 103(1): 84-103.

Bowles, Hannah Riley, Linda Babcock, and Kthleen L. McGinn (2005), "Constraints and Triggers: Situational Mechanics of Gender in Negotiation," *Journal of Personality and Social Psychology,* 89(6): 951-65.

Bowles, Samuel (1985), "The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian, and Marxian Models," *American Economic Review,* 75(1): 16-36.

Bowles, Samuel and Herbert Gintis, (1977), *Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life.* New York: Basic Books.

Bowles, Samuel and Herbert Gintis (1986), *Democracy and Capitalism: Property, Community and the Contradictions of Modern Social Thought.* New York: Basic Books.

Bowles, Samuel, Herbert Gintis, and Melissa Osborne (2001), "The Determinants of Earnings: A Behavior Approach," *Journal of Economic Literature,* 39(4): 1137-1176.

Brown, Charles and James Medoff (1989), "The Employer Size-Wage Effect," *Journal of Political Economy,* 97(5): 1027-1059.

Bulow, Jeremy I., and Lawrence H. Summers (1986), "A Theory of Dual Labor Markets with Application to Industrial Policy, Discrimination, and Keynesian Unemployment," *Journal of Labor Economics,* 4(3): 376-414.

Burdett, Kenneth and Dale T. Mortensen (1998), "Wage Differentials, Employer Size, and Unemployment," *International Economic Review,* 39(2): 257-273.

Cain, Daylian M., George Loewenstein, and Don A. Moore (2005), "The Dirt on Coming Clean: Perverse Effects of Disclosing Conflicts of Interest," *Journal of Legal Studies,* 34(1): 1-25.

Camerer, Colin, Linda Babcock, George Loewenstein, and Richard Thaler (1997), "Labor Supply of New York City Cabdrivers: One Day At A Time," *Quarterly Journal of Economics,* 112(2): 407-41.

Camerer, Colin, and George Loewenstein (2004), "Behavioral Economics: Past, Present, Future," in Colin Camerer, George Loewenstein, and Matthew Rabin, eds., *Advances in Behavioral Economics,* Princeton, NJ: Princeton University Press.

Cappelli, Peter and Keith Chauvin (1991), "An Interplant Test of the Efficiency Wage Hypothesis," *Quarterly Journal of Economics,* 106(3): 769-787.

Cebul, Randall, Ray Herschman, James Rebitzer, Lowell Taylor and Mark Votruba (2008), "Organizational Fragmentation in the U.S. Health Care System," *Journal of Economic Perspectives,* 22(4).

Charness, Gary (2004), "Attribution and Reciprocity in an Experimental Labor Market," *Journal of Labor Economics,* 22(3), 665-688.

Charness, Gary and Martin Dufwenberg (2006), "Promises and Partnership," *Econometrica,* 74(6), 1579-1601.

Charness, Gary and Peter Kuhn (2010), "Lab Labor: What Can Labor Economists Learn from the Lab?" this volume.

Charness, Gary and Matthew Rabin (2002), "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics,* 117(3): 817-69.

Cho, In-Koo and David M. Kreps (1987), "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics,* 102(2), 179-221.

Coase, Ronald H. (1937), "The Nature of the Firm," *Economica,* 4(16): 386-405.

Cooper, David J. and Rebitzer, James B. (2006), "Managed Care and Physician Incentives: The Effects of Competition on the Cost and Quality of Care," *B.E. Journals in Economic Analysis and Policy: Contributions to Economic Analysis and Policy,* 5(1): 1-30.

Cornell, Bradford and Ivo Welch (1996), "Culture, Information, and Screening Discrimination," *Journal of Political Economy,* 104(3): 542-71.

Courty, Pascal and Gerald Marschke (2004), "An Empirical Investigation of Gaming Responses to Explicit Performance Incentives," *Journal of Labor Economics,* 22(1): 23-56.

Dana, Jason and George Loewenstein (2003), "A Social Science Perspective on Gifts to Physicians from Industry," *Journal of the American Medical Association,* 290(2): 252-55.

Deci, Edward L., Richard Koestner, and Richard M. Ryan (1999), "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation," *Psychological Bulletin,* 125(6): 627-668.

Deci, Edward L., and Richard M. Ryan (1985), *Intrinsic Motivation and Self-Determination in Human Behavior.* New York: Plenum Press.

Delfgaauw, Josse and Robert Dur (2007), "Signaling and Screening of Worker Motivation," *Journal of Economic Behavior and Organization,* 62: 605-24.

Delfgaauw, Josse and Robert Dur (2008), "Incentives and Workers' Motivation in the Public Sector," *Economic Journal,* 118: 171-91.

DellaVigna, Stefano (2009), "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature,* 47(2): 315-72.

deQuervain, Dominique J. F., *et al.* (2004), "The Neural Basis of Altruistic Punishment," *Science,* 305(5688): 1254-58.

Dur, Robert and Amihai Glazer (2008a), "The Desire for Impact," *Journal of Economic Psychology,* 29(3): 285-300.

Dur, Robert and Amihai Glazer (2008b), "Optimal Contracts When a Worker Envies His Boss," *Journal of Law, Economics, and Organization* 24(1): 120-137.

Ekman, Paul (2001), *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage,* 3rd ed. New York City: W. W. Norton and Company.

Ellingsen, Tore and Magnus Johannesson (2007), "Paying Respect," *Journal of Economic Perspective,* 21(4), 135-49.

Ellingsen, Tore and Magnus Johannesson (2008), "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review,* 98(3): 990-1008.

Encinosa, William E., III, Martin Gaynor, and James B. Rebitzer (2007), "The Sociology of Groups and the Economics of Incentives: Theory and Evidence on Compensation Systems," *Journal of Economic Behavior and Organization,* 62(2): 187-214.

Fairlie, Robert and William Sundstrom (1999), "The Emergence, Persistence, and Recent Widening of the Racial Unemployment Gap," *Industrial and Labor Relations Review,* 52(2): 252-70.

Falk, Armin, Ernst Fehr, and Christian Zehnder (2006), "Fairness Perceptions and Reservation Wages—The Behavioral Effects of Minimum Wage Laws," *Quarterly Journal of Economics,* 121(4): 1347-81.

Falk, Armin and Andrea Ichino (2006), "Clean Evidence on Peer Effects," *Journal of Labor Economics,* 24(1): 39-57.

Falk, Armin and Michael Kosfeld (2006), "The Hidden Costs of Control," *American Economic Review,* 96(5): 1611-1630.

Farber, Henry S. (2005), "Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers," *Journal of Political Economy,* 113(1): 46-82.

Farber, Henry S. (2008), "Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers," *American Economic Review,* 98(3): 1069-82.

Fehr, Ernst, Martin Brown, and Christian Zehnder (2009), "On Reputation: A Microfoundation of Contract Enforcement and Price Rigidity," *Economic Journal,* 199: 333-53.

Fehr, Ernst, and John A. List (2004), "The Hidden Costs and Returns of Incentives—Trust and Trustworthiness Among CEOs," *Journal of the European Economic Association,* 2(5): 743-771.

Fehr, Ernst and Simon Gaetchter (2000), "Cooperation and Punishment in Public Goods Experiments," *American Economic Review,* 90(4).

Fehr, Ernst and Klaus M. Schmidt (1999), "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics,* 114(3): 817-68.

Fehr, Ernst and Lorenz Goette (2007), "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment," *American Economic Review,* 97(1): 298-317.

Fehr, Ernst, Georg Kirchsteiger, Aron Riedl (1998), "Gift Exchange and Reciprocity in Competitive Experimental Markets," *European Economic Review,* 42(1): 1-34.

Fernandez, Raquel, Alessandra Fogli, and Claudia Olivetti (2004), "Mothers and Sons: Preference Formation and Female Labor Force Dynamics," *Quarterly Journal of Economics,* 119(4): 1249-99.

Festinger, Leon and James Carlsmith (1959), "Cognitive Consequences of force Compliance," *Journal of Personality and Social Psychology,* 58(2): 203-210.

Fischbacher, Uris, Christina M. Fong, and Ernst Fehr (2009), "Fairness, Errors and the Power of Competition," *Journal of Economic Behavior and Organization,* 72(1): 527-545.

Fischer and Huddart (2008), "Optimal Contracting with Endogenous Social," *American Economic Review,* 98(4): 1459-75.

Fisman, Raymond and Edward Miguel (2007), "Corruption, Norms, and Legal Enforcement: Evidence from Diplomatic Parking Tickets," *Journal of Political Economy,* 115(6): 1020-48.

Frank, Robert H. (1984), "Are Workers Paid their Marginal Products?" *American Economic Review,* 74(4): 549-71.

Frank, Robert H. (1985), *Choosing the Right Pond: Human Behavior and the Quest for Status.* New York: Oxford University Press.

Frank, Robert H. (1988), *Passions Within Reason: The Strategic Role of the Emotions.* New York City: W. W. Norton and Company.

Frey, Bruno S. (1997), *Not Just for the Money: An Economic Theory of Personal Motivation.* Cheltenham: Edward Elgar.

Frey, Bruno S. and Reto Jegen (2001), "Motivation Crowding Theory," *Journal of Economic Surveys,* 15(5): 589-611.

Frey, Bruno S. and Felix Oberholzer-Gee (1997), "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out," *American Economic Review,* 87(4), 746-755.

Gayle, George-Levi, and Robert A. Miller (2009), "Has Moral Hazard Become a More Important Factor in Managerial Compensation?" *American Economic Review,* 99(5): 1740-1769.

Gaynor, Martin, James Rebitzer, and Lowell J. Taylor (2004), "Physician Incentives in Health Maintenance Organizations," *Journal of Political Economy,* 112(4).

Genesove, David and Christopher Mayer (2001), "Loss Aversion and Seller Behavior: Evidence from the Housing Market," *Quarterly Journal of Economics,* 116(4): 1233-60.

Gibbons, Robert (1998), "Incentives in Organizations," *Journal of Economic Perspectives,* 12(4): 115-32.

Gibbons, Robert and Michael Waldman (1999), "A Theory of Wage and Promotion Dynamics Inside Firms," *Quarterly Journal of Economics,* 114(4): 1321-1358.

Gicheva, Dora (2009), "Working Long Hours and Career Wage Growth," Working Paper, Yale University.

Gintis, Herbert, Samuel Bowles, Robert Boyd and Ernst Fehr (2003), "Explaining Altruistic Behavior in Humans," *Evolution and Human Behavior,* 24(3): 153-72.

Gneezy, Uri, and John A. List (2006), "Putting Behavioral Economics to Work: Field Evidence of Gift Exchange," *Econometrica,* 74(5): 1365-84.

Gneezy, Uri, Niederle, Muriel, and Rustichini, Aldo (2003), "Performance in Competitive Environments: Gender Differences," *Quarterly Journal of Economics,* 118(3): 1049-74.

Gneezy, Uri, and Aldo Rustichini (2000a), "A Fine is a Price," *Journal of Legal Studies,* 29(1): 1-17.

Gneezy, Uri, and Aldo Rustichini (2000b), "Pay Enough or Don't Pay at All," *Quarterly Journal of Economics,* 115(3): 791-810.

Grossman, Sanford J. and Oliver D. Hart (1986), "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy,* 94(4), 691-719.

Groves, Theodore, Yongmiao Hong, John McMillan, and Barry Naughton (1994), "Autonomy and Incentives in Chinese State Enterprises," *Quarterly Journal of Economics,* 109(1): 183-209.

Groves, Theodore, Yongmiao Hong, John McMillan and Barry Naughton (1995), "China's Evolving Managerial Labor Market," *Journal of Political Economy,* 103(4): 873-92.

Hall, Brian, and Thomas Knox (2004), "Underwater Options and the Dynamics of Executive Pay-to-Performance Sensitivities," *Journal of Accounting Research,* 42(2): 365-412.

Hall, Brian and Jeffrey Liebman (1998), "Are CEOS Really Paid Like Bureaucrats?" *Quarterly Journal of Economics,* 113(3): 653-91.

Hall, Brian J. and Kevin J. Murphy (2003), "The Trouble with Stock Options ," *The Journal of Economic Perspectives,* 17(3): 49-70.

Heron, Randall and Erik Lie (2009), "What Fraction of Stock Option Grants to Top Executives Have Been Backdated or Manipulated?" *Management Science,* 55(4): 513-525.

Heyes, Anthony (2007), "The Economics of Vocation or Why is a Badly Paid Nurse a Good Nurse?" *Journal of Health Economics,* 24, 561-569.

Holmström, Bengt (1979), "Moral Hazard and Observability," *Bell Journal of Economics,* 10(1), 74-91.

Holmström, Bengt (1999), "Managerial Incentive Problems: A Dynamic Perspective," *Review of Economic Studies,* 66(1): 169-82.

Holmström, Bengt and Paul Milgrom (1991), "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization,* 7, Special Issue (Papers from the Conference on the New Science of Organization): 24-52.

Holmström, Bengt and Paul Milgrom (1994), "The Firm as an Incentive System," *American Economic Review,* 84(4), 972-991.

Hornstein, Andreas, Per Krusell, and Giovanni Violante (2007), "Frictional Wage Dispersion in Search Models: A Quantitative Assessment," NBER Working Paper.

Ichino, Andrea, and Giovanni Maggi (2000), "Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm," *Quarterly Journal of Economics,* 115(3): 1057-90.

Jackson, Howell E. (2008), "The Trilateral Dilemma in Financial Regulation," in Anna Maria Lusardi, ed., *Improving the Effectiveness of Financial Education and Savings Programs.* Chicago: University of Chicago Press.

Jackson, Kirabo C. and Elias Bruegmann (2009), "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," Manuscript.

Jacob, Brian A. (2005), "Accountability, Incentives and Behavior: the Impact of High-Stakes Testing in the Chicago Public Schools," *Journal of Public Economics Volume* 89(5-6), 761-796.

Jacob, Brian A. and Steven D. Levitt (2003), "Rotten Apples: An Investigation of The Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics,* 118(3), 843-877.

Jensen, Michael, and Kevin J. Murphy (1990), "Performance Pay and Top-Management Incentives," *Journal of Political Economy,* 98: 225-64.

Kahn, Lawrence M. and Peter D. Sherer (1999), "Contingent Pay and Managerial Performance," *Industrial and Labor Relations Review,* 43(3): 107S-120S.

Kandel, Eugene and Edward P. Lazear (1992), "Peer Pressure and Partnerships," *Journal of Political Economy,* 100(4): 801-17.

Knez, Marc and Simester, Duncan (2001), "Firm-Wide Incentives and Mutual Monitoring at Continental Airlines," *Journal of Labor Economics,* 19(4): 743-72.

Krueger, Alan B. and Alexandre Mas (2004), "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires," *Journal of Political Economy,* 112(2): 253-89.

Krueger, Alan B. and Lawrence H. Summers (1988), "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica,* 56(2): 259-293.

Landers, Renee, James Rebitzer, and Lowell J. Taylor (1996), "Rat Race Redux: Adverse Selection in the Determination of Work Hours in Law Firms," *American Economic Review,* 86(3).

Landers, Renee, James Rebitzer, and Lowell J. Taylor (1997), "Work Norms in Professional Labor Markets," in Francine Blau and Ronald Ehrenberg, eds., *Gender and Family Issues in the Workplace.* New York: Russell Sage Press.

Lang, Kevin (1986), "A Language Theory of Discrimination," *Quarterly Journal of Economics,* 101: 363-82.

Lazear, Edward P. (1989), "Pay Equality and Industrial Politics," *Journal of Political Economy,* 97(3): 561-80.

Lazear, Edward P. (1998), *Personnel Economics for Managers.* New York: John Wiley and Sons.

Lazear, Edward P. (1999), "Culture and Language," *Journal of Political Economics,* 107(6, part 2): S95-S126.

Lazear, Edward P. (2000), "Performance Pay and Productivity," *American Economic Review,* 90(5): 1346-1361.

Lazear, Edward P. and Sherwin Rosen (1981), "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy,* 89(5): 841-64.

Lepper, Mark R., David Greene, and Richard Nisbett (1973), "Undermining Children's Intrinsic Interest with Extrinsic Reward: A Test of the 'Overjustification' Hypothesis," *Journal of Personality and Social Psychology,* 28(1): 129-137.

Levine, David I. (1991), "Just-Cause Employment Policies in the Presence of Worker Adverse Selection," *Journal of Labor Economics,* 9(3): 294-305.

Levy, Frank and Peter Temin (2007) "Inequality and Institutions in 20th Century America," NBER Working Paper.

MacLeod, W. Bentley and James Malcomson (1989), "Implicit Contracts, Incentive Compatibility, and Involuntary Unemployment," *Econometrica,* 57: 447-80.

Malcomson, James M. (1984), "Work Incentives, Hierarchy, and Internal Labor Markets," *Journal of Political Economy,* 92(3): 486-507.

Malcomson, James M. (1999), "Individual Employment Contracts," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics,* volume 3, number 3.

Mas, Alexandre (2006), "Pay, Reference Points, and Police Performance," *Quarterly Journal of Economics,* 121(3): 783-821.

Mas, Alexandre (2008), "Labour Unrest and the Quality of Production: Evidence from the Construction Equipment Resale Market," *Review of Economic Studies,* 75(1): 229-58.

Mas, Alexandre and Enrico Moretti (2009), "Peers at Work," *American Economic Review,* 99(1): 112-45.

McCabe, Kevin A., Mary L. Rigdon, and Vernon L. Smith (2003), "Positive Reciprocity and Intentions in Trust Games," *Journal of Economic Behavior and Organization,* 52(2): 267-275.

McMillan, John, John Whalley and Lijing Zhu (1989), "The Impact of China's Economic Reforms on Agricultural Productivity Growth," *Journal of Political Economy,* 97(4): 781-807.

Moore, Don A. and George Loewenstein (2004), "Self-Interest, Automaticity, and the Psychology of Conflict of Interest," *Social Justice Research,* 17(2): 189-202.

Nagin, Daniel, James Rebitzer, Seth Sanders, and Lowell Taylor (2002), "Monitoring and Motivation: An Analysis of a Field Experiment" *American Economic Review,* 92(4).

Neal, Derek (2006), "Why Has Black-White Convergence Stopped?" *Handbook of the Economics of Education,* vol. 1, edited by Eric Hanushek and Finis Welch. Amsterdam: Elsevier.

Niederle, Muriel and Lise Vesterlund (2007), "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics,* 122(3), 1067-1101.

Oyer, Paul (1998), "The Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality," *Quarterly Journal of Economics,* 113(1), 14985.

Oyer, Paul and Scott Shaefer (2010), "Personnel Economics: Hiring and Incentives," this handbook.

Prendergast, Canice (1999), "The Provision of Incentives in Firms," *Journal of Economic Literature,* 37(1): 7-63.

Prendergast, Canice (2007), "The Motivation and Bias of Bureaucrats," *American Economic Review,* 97(1): 180-96.

Prendergast, Canice and Robert H. Topel (1996), "Favoritism in Organizations," *Journal of Political Economy,* 104(5): 958-78.

Rabin, Matthew (1993), "Incorporating Fairness into Game Theory and Economics," *American Economic Review,* 83: 1281-1302.

Rabin, Matthew and Richard H. Thaler (2001), "Anomalies: Risk Aversion," *Journal of Economic Perspectives,* 15(1): 13.

Rebitzer, James B. (1995), "Job Safety and Contract Workers in the Petrochemical Industry," *Industrial Relations,* 34(1): 40-57.

Rebitzer, James B., and Lowell J. Taylor (1991), "A Model of Dual Labor Markets When Product Demand is Uncertain," *Quarterly Journal of Economics,* 106(4): 1373-1383.

Rebitzer, James B., and Lowell J. Taylor (1995a), "Efficiency Wages and Employment Rents: The Employer Size Wage Effect in the Job Market for Lawyers," *Journal of Labor Economics,* 13(4): 678-708.

Rebitzer, James B., and Lowell J. Taylor (1995b), "Do Labor Markets Provide Enough Short-Hour Jobs? An Analysis of Work Hours and Work Incentives," *Economic Inquiry,* 33(2): 257-273.

Rebitzer, James B., and Lowell J. Taylor (1995c), "The Consequences of Minimum Wage Laws: Some New Theoretical Ideas," *Journal of Public Economics,* 56(2): 245-255.

Ritter, Joseph A., and Lowell J. Taylor (1994), "Workers as Creditors: Efficiency Wages and Performance Bonds," *American Economic Review,* 84(3): 694-704.

Ritter, Joseph A., and Lowell J. Taylor (forthcoming), "Racial Disparity in Unemployment," *Review of Economics and Statistics.*

Ritter, Joseph A., and Lowell J. Taylor (2009), "Low-Powered Incentives and the Motivation of Critical Workers," manuscript.

Rousseau, Denise M. (1995), *Psychological Contracts in Organizations: Understanding Written and Unwritten Agreements,* Thousand Oaks: Sage Publications.

: Rudman, P. and L.A. Glick (1999), "Feminized Management and Backlash toward Agentic Women: The Hidden Costs to Women of a Kinder, Gentler Image of Middle Managers, *Journal of Personality and Social Psychology,* 77(5): 1004-1010.

Ryan, Richard M., and Edward L. Deci (2000), "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions," *Contemporary Education Psychology,* 25(1): 54-67.

Sally, David (2002), "Two Economic Applications of Sympathy," *Journal of Law, Economics, and Organization,* 18(2): 455-87.

Shapiro, Carl and Joseph Stiglitz (1984), "Involuntary Unemployment as a Worker Discipline Device," *American Economic Review,* 74(3): 433-44.

Simon, Herbert A. (1951), "A Formal Theory of the Employment Relationship," *Econometrica,* 19(3), 293-305.

Sliwka, Dirk (2007), "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes," *American Economic Review,* 97(3): 999-1012.

Small, Deborah, Michele Gelfand, Linda Babcock, and Hilary Gettman (2007), "Who Goes to the Bargaining Table? The Influence of Gender and Framing on the Initiation of Negotiation," *Journal of Personality and Social Psychology,* 93(4): 600-13.

Spier Kathryn E. (1992), "Incomplete Contracts and Signalling," *RAND Journal of Economics,* 23(3) 432-43.

Stevenson, Betsy and Justin Wolfers (2008), "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox," *Brookings Papers on Economic Activity.*

Taylor, Lowell J. (2007), "Optimal Wages in the Market for Nurses: An Analysis Based on Heyes' Model," *Journal of Health Economics,* 26(5), 1027-30.

Thaler, Richard H. and Cass R. Sunstein (2008), *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Yale University Press.

Williamson, Oliver E. (1985), *The Economic Institutions of Capitalism.* New York: Free Press.

Valley, Kathleen L., Joseph Moag, and Max H. Bazerman (1998), "'A Matter of Trust:' Effects of Communication on the Efficiency and Distribution of Outcomes," *Journal of Economic Behavior and Organization,* 34(2): 211-38.

FIGURE 1. Equilibrium Wage and Employment

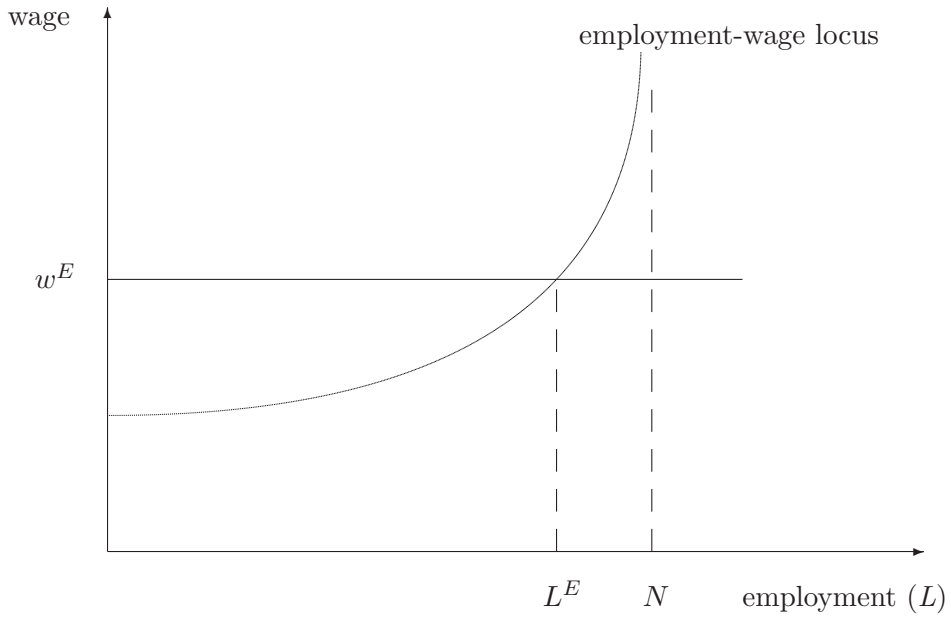
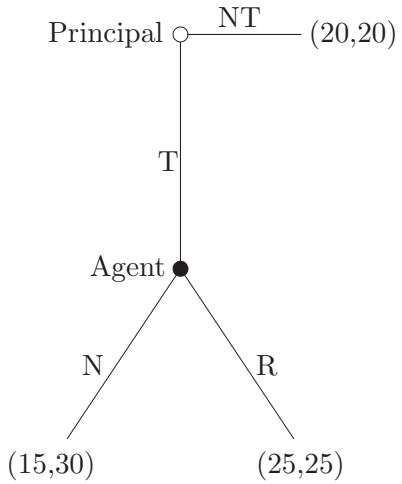
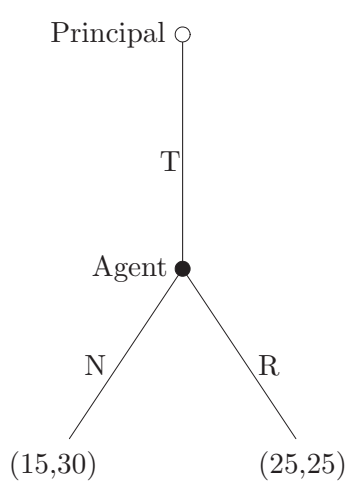


FIGURE 2. A Principal Agent Game of Trust



Case 1. "Trust" Can Serve as a Signal of the Principal's Pro-Social Inclination



Case 2. Principal Must Play "Trust"

FIGURE 3. The Effects of an Extrinsic Reward on the Pool of Agents Providing Effort