Abstract

Most upper-management and sales force personnel, as well as workers in many other jobs, are paid based on performance, which is widely perceived as motivating effort and enhancing productivity relative to non-contingent pay schemes. However, psychological research suggests that excessive rewards can in some cases produce supra-optimal motivation, resulting in a decline in performance. To test whether very high monetary rewards can decrease performance, we conducted a set of experiments at MIT, the University of Chicago, and in rural India. Subjects in our experiment worked on different tasks and received performance-contingent payments that varied in amount from small to very large relative to their typical levels of pay. With some important exceptions, very high reward levels had a detrimental effect on performance.

The authors are grateful for the help of the faculty and students at Narayanan College in Madurai for their help in carrying out the experiment in India. Faculty: Dr. R. Srinivasan; Prof. A. Narasimhamurthy; Dr. K. Ramasamy; Dr. M. Jayakuma. Students: J. Moses Gnanakkan; P. Kalaignar; M. Ramesh; G. Selvakumar; K. Prabakara DOSS. Other thanks go to Christopher Simeone and Mark Porter for their help with the experiment at MIT, and to Jackie Zires for her help with the experiment at the University of Chicago. The authors are also indebted to Colin Camerer for his helpful comments.
I. Introduction

Workers in a wide variety of jobs are rewarded for their effort based on observed measures of performance. The intuitive logic for performance-based compensation is to motivate individuals to increase their effort, and hence their output. Some evidence suggests that payment for performance can indeed increase performance. For example, Lazear [2000] showed that when a large company, under new management, moved from hourly wages to piece-rate pay, productivity increased by a dramatic 44% [see Prendergast, 1999, for a survey].

The utilization of performance-based incentives can be observed not only in businesses, but also in other areas such as sports. Soccer federations, for example, offered rich rewards for success in the World Cup 2002, with bonuses rising progressively by each round, including an extra bonus for winning the title. Bonuses paid by national soccer federations have increased dramatically over the past years with those in the richest countries paying out millions of dollars for success and even poor nations catching up with substantial monetary incentives [Slam! Sports, May 14, 2002].

The expectation that increasing performance-contingent incentives will improve performance rests on two subsidiary assumptions: (1) that increasing performance-contingent incentives will lead to greater motivation and effort, and (2) that this increase in motivation and effort will result in improved performance.

The first assumption, that transitory performance-based increases in pay will produce increased motivation and effort, is generally accepted [Prendergast, 1999] although there are some notable exceptions. Gneezy and Rustichini [2000a], for example, have documented situations, both in laboratory and field experiments, in which people who were not paid at all exerted greater effort than those who were paid a small amount. In one of their experiment,
students who were collecting donations door to door actually visited fewer houses and collected less money when they were paid a small commission [see also Frey and Jegen, 2001; Heyman and Ariely, 2004; Gneezy and Rustichini [2000b]. Paying a small amount in such situations seems to risk squelching intrinsic motivation, and, if the amount of pay is not sufficient to compensate for the decline in intrinsic motivation, motivation and effort can decline.

Another situation in which effort may not respond in the expected fashion to a change in transitory wages is when workers have an earnings target that they apply narrowly. For example, Colin Camerer and coauthors [1997] found that New York City cab drivers quit early on days when their hourly earnings were high and worked longer hours when their earnings were low. They speculated that the cab drivers may have had a daily earnings target, beyond which their motivation to continue working dropped off.

Although there appear to be reasons to question the generality of the positive relationship between effort and pay, our focus in this paper is on the second assumption. The experiments we report address the question of whether increased effort necessarily leads to improved performance. Providing subjects with different levels of incentives, including incentives that were very high relative to their normal income, we examine whether, across different tasks, an increase in contingent pay leads to an improvement or decline in performance.

II. Prior Research on the Connection between Effort and Performance

Unlike the relationship between motivation and effort, the relationship between motivation and performance has not attracted much attention from economists, perhaps because the belief that increased motivation improves performance is so deeply held. However, research by
psychologists has documented situations in which increased motivation can result in a decrement in performance – a phenomenon known as “choking under pressure” [Baumeister, 1984].

The idea that excessive incentives could undermine task performance is embodied in the “Yerkes-Dodson law” [Yerkes and Dodson, 1908] which posits that there is an optimal level of arousal for executing tasks, and that departures from this level in either direction lead to a decrement in performance. The first demonstration of the effect by Yerkes and Dodson involved rats who were placed in a cage and forced to repeatedly choose between exploring either the left or right of two passages. On each trial, the experimenters randomly hung a white card in one passage and a black card in the other, and exploring the passage with the black card always resulted in a shock. For some rats the shock was always small, for some medium, and for a third group it was strong. The main finding was that the rats learned to avoid the shocks most quickly when the shocks were at an intermediate level of intensity. Similar results have been obtained with humans [Neiss, 1988]. Since arousal is tightly linked to motivation and performance, these findings imply that increases in motivation beyond an optimal level can, in some situations, produce supra-optimal levels of arousal and hence decrements in performance.

Extending the empirical regularity of the Yerkes-Dodson law, research by psychologists has sought to identify the range of situations and psychological mechanisms that can produce an inverse relationship between motivation and performance. For example, when performance on a task relies on highly practiced, automatic skills [Baumeister, 1984; Langer and Imber, 1979], increasing the intensity of competition, cash incentives, the presence of an audience, or the belief that a task is diagnostic of something that one cares about, such as intelligence, can all cause people, involuntarily, to consciously think about the task, shifting control of behavior from ‘automatic’ to ‘controlled’ mental processes that are less effective for such tasks [see Camerer,
Loewenstein, and Prelec, 2004, for a detailed account of automatic and controlled processes]. Sports provide a prototypical example of such over-learned, automatic tasks. Thinking about how one is swinging the golf club or bat, or about how to get the basketball into the net, can have perverse effects on performance. In fact, there are some studies of choking under pressure in sports, including one Australian study which found that free-throw shooting performance among elite Australian basketball players was worse during games than during training [Dandy, Brewer, and Tottman, 2001]. The same shift from ‘automatic’ to ‘controlled’ processes can also help to explain why the presence of an audience, which tends to increase the motivation to perform well, can have detrimental effects [see also Zajonc, 1965].

A second mechanism by which increased motivation is likely to have a negative effect on performance involves the focus of attention. Increased motivation tends to narrow individuals’ focus of attention [Easterbrook, 1959], which can be detrimental for tasks that involve insight or creativity, since both require a kind of open-minded thinking that enables one to draw unusual connections between elements. McGraw and McCullers [1978] provided support for this mechanism by showing that the introduction of monetary rewards for tasks that involved problem-solving had detrimental effects on performance. In addition to the narrowing of attention, large incentives can simply occupy the mind and attention of the laborer, distracting the individual from the task at hand.

In summary, psychological research has identified several mechanisms that can produce choking under pressure, as well as diverse factors that create the type of pressure that produces choking [see Baumeister and Showers, 1986, for an in-depth discussion of these effects].

For economics, however, the most interesting determinant of performance pressure is the level of performance-contingent monetary incentives. Our primary goal in the studies reported
herein is to test, in experiments that satisfy standard experimental economics criteria, whether increasing incentives beyond some threshold may result in lower performance. A second goal is to examine the generality of any detrimental effect of incentives. Among the six tasks in the first experiment, therefore, we included some that drew primarily on motor skills, some that drew primarily on concentration, and some that drew primarily on creativity. Based on the literature showing detrimental effects of high incentives on motor skills and creativity, we anticipated that the high rewards might interfere with tasks that draw primarily on these skills, but not with those involving primarily concentration. As will be seen, however, no such differences emerged; the highest levels of rewards produced lower performance on all tasks in the first experiment. To examine this issue further in the second experiment, we then included a task that required only physical effort, which should not be subject to any of the mechanisms leading to choking under pressure that have been identified in the psychology literature. In this case the predicted differences between tasks did emerge. Finally, our third experiment extends the scope of investigation from financial to social incentives.

III. Experiment 1

Design

Eighty-seven residents of a rural town in India were recruited to participate in the experiment, which took place late in 2002. The sample consisted of 26.4% females and 73.6% males. The majority of participants (90.8%) were Hindu, 5.7% were Christian, and 3.4% were Muslim. Their standard of living of participants can be best described by their level of education and possessions. Participants in this experiment had, on average, 5.6 years of education, and 26% had no formal education. Approximately half of the participants reported that they owned a TV
(M = 49.4%), and about half owned a bicycle for transportation (M = 51.7%). None owned a car, and only 6.9% had a telephone in their house.

The experiment was conducted with one participant at a time. Participants were randomly assigned to one of three treatments in which they faced incentives (on all six games) that were either relatively small, moderate, or very large. In each treatment, participants played 6 different games in a random order and were promised payments for each game if they reached certain performance levels. The magnitude of the payment in each game depended on the treatment (small, medium or large incentive magnitude) and whether they reached either of two specified performance levels which we labeled “good” and “very good.” Participants received full payment (i.e. 4, 40 or 400 Indian Rupee depending on the treatment) if they reached the “very good” performance level, half of that if they reached the “good” performance level, and nothing if they failed to reach the “good” performance level (these two performance levels, as well as the games themselves, were selected based on pre-testing with MIT students).

The maximum possible payment for any one task in the high incentive treatment (Rs 400) was relatively close to the all-India average monthly per capita consumer expenditure (MPCE) in rural areas, which was Rs 495 [Rangachari, 2003]. Thus, in the unlikely event that a subject in the high payment treatment achieved “very good” performances on all six tasks, she would earn an amount approximately equal to half of the mean yearly consumer expenditure in the village. These stakes are effectively much larger than those that are typically offered in experimental settings.

The Games. The six games fell into three broad categories based on whether they required primarily: creativity, concentration, or motor skills.
The game that was used as a creativity task was “Packing Quarters.” In this game participants were asked to fit 9 metal pieces of quarter circles into a black wooden frame within a given time. It is easy to fit 8 pieces, but, to fit all 9, the pieces have to be packed in a particular way. The good performance level was defined by a completion of the task within 240 seconds. The very good performance level was defined by a completion of the task within 120 seconds. Participants had only 1 trial to reach these goals.

The concentration tasks included two games: “Simon” and “Recall last 3-digits.” “Simon” is an electronic game that requires memory and repetitions. The game flashes a sequence of colored lights accompanied by the light-specific sounds, and the task is to repeat the sequence by pushing the corresponding light-buttons in the same order. The good performance level was defined by at least one repetition of 6 consecutive lights. The very good performance level was defined by at least one repetition of 8 consecutive lights. Participants had 10 trials to reach these goals. The second concentration game was “Recall last 3-digits” in which the experimenter reads a sequence of digits, stops at an unannounced point, and the participant is asked to recall the last 3-digits. Participants had 14 trials in this task. The good performance level was defined by at least 4 correct trials. The very good performance level was defined by at least 6 correct trials.

Finally, there were three different motor skill tasks: “Labyrinth”, “Dart Ball”, and “Roll-Up”. “Labyrinth” is a game with a playing surface on top of a box that can be tilt in either of two planes. The playing surface shows a pathway from the “start” position along which the player has to advance a small steel ball to the “finish” position, while avoiding the traps (holes in the board). The good performance level was defined by passing the 7th hole. The very good performance level was defined by passing the 9th hole. Participants had 10 trials to reach these goals. “Dart Ball” is similar to Darts but instead of throwing sharp metal arrows, the game uses
tennis balls thrown at an inflated target with Velcro patches. Participants had 20 trials in this task. The good performance level was defined by having at least 5 balls hit the center of the target. The very good performance level was defined by having at least 8 balls hit the center of the target. “Roll-Up” is a game in which one attempts to drop a ball into the highest possible slot by deftly spreading apart then pushing together two rods [Baumeister’s 1984]. Participants had 20 trials in this task. The good performance level was defined by having at least 4 balls hit the furthest hole. The very good performance level was defined by having at least 6 balls hit the furthest hole.

Results

There are four possible ways to treat the dependent measures in this experiment: One would be to look at the raw scores, but this is not ideal since it does not directly relate to the compensation participants received. A second way is to examine the probability of reaching at least the “good” performance level. Yet another would be to examine the probability of reaching the “very good” performance level, and the final would be to examine the fraction of earnings from the total possible earnings. As is evident from Table I, the general pattern of conclusions was the same regardless of how we analyzed the data. The most interesting measure from an economics perspective is the fraction of possible earnings, since it represents the measure that is most closely linked to the incentives that the subjects actually faced. In what follows, therefore, we present all results in terms of this measure.

*** Table I ***
To test for the significance of observed differences, we analyzed the data with a 3 (incentive levels) by 6 (games) mixed between subjects (incentive levels) and within subjects (games) repeated-measure ANOVA. This overall model revealed a significant effect for payment condition \[F(2, 84) = 10.24, p < 0.001\], a significant effect for game \[F(5, 420) = 9.22, p < 0.001\], and a non significant interaction between them \[F(10, 420) = 1.24, p = 0.263\]. The non-significant interaction suggests that the effect of incentive level on the different games was generally similar.

As can be seen in Figure 1, the aggregated performance levels across all 6 games supported the hypothesis that relatively high monetary incentives can have perverse effects on performance. The average share of earnings relative to maximum possible earnings was lowest in the high payment condition (M = 19.5%), but higher and almost equal in the mid (M = 36.7%) and low payment conditions (M = 35.4%). A post-hoc Fisher LSD test revealed that the difference between the low and mid payment condition was not significant \[p = 0.768\], while the difference between the low and high condition was significant \[p < 0.001\] as was the difference between the mid and high condition \[p < 0.001\]. These findings support the main hypothesis that motivated the experiment – namely that additional incentives can decrease performance.

Somewhat contrary to our expectations, the pattern of results held across tasks differing both in terms of difficulty and the types of skills they require (see Figure II). To examine performance in each of the six games, we carried out four sets of simple contrasts: one for each of the pairwise comparisons of the three incentive levels (low-mid; low-high; mid-high) and one
that compared performance in the high payment condition to performance in the low and mid payment conditions combined. This final contrast was based on the post-hoc analysis of overall effects across games, which revealed that performance in the high incentive condition was often below that of the low and mid incentive level conditions and that performance in these two conditions was similar.

*** Figure II ***

The contrast of the low and mid levels of incentives revealed little difference in performance: Only one of the games (Labyrinth) showed a marginally significant effect (see table II). Comparisons between the high payment condition and either the low, mid, or both payment conditions together, however, revealed a number of statistically significant differences (see table II). For example, the contrast between the high payment condition and the low and mid payment conditions together was significant at the 0.05 level for Simon, Labyrinth, and Packing Quarters, marginally significant at the 0.1 level for Roll-Up, and not significant for Dart-Ball and Recall last 3 digits.

*** Table II ***

**Summary**

Overall, the results point to three main conclusions: First, with the sole exception of the Labyrinth task there was no significant difference in the performance between the low and mid payment conditions. Thus, despite the relative large difference in magnitude of reward across the treatments (i.e. 10 times higher for the mid payment condition relative to the low payment
condition), performance did not seem to increase. One interpretation of this result is that the incentives in the low payment condition (which were not altogether that low) created a level of performance that was already at a peak.

Second, and more importantly, the performance of participants was always lowest in the high payment condition when compared with the low and mid payment conditions together, although this pattern was significant only for 3 of the 6 games.

Third, and contrary to our expectations, we did not observe any obvious difference in the effect of incentives on performance for different categories of games. We included, for example, “Simon” and “Recall last 3 digits” because these tasks require tiresome concentration, and we thought that subjects who were more highly motivated might be more likely to maintain high levels of concentration. We did not, however, observe any such difference; both games generally displayed declining performance as a function of incentives – same as the motor skill tasks and the creativity task.

There are a number of possible reasons for why we did not observe different patterns of results for the two concentration tasks. One is that, while these tasks did involve concentration, they may have also involved other cognitive skills that are subject to choking. For example, sometimes memorizing information is actually easier if one doesn't pay too much attention. Another is that the incentives we chose may have simply been too high. Perhaps the concentration tasks had higher threshold levels of motivation at which performance started to decline, but our choice of incentive levels in the three conditions, and particularly in the high incentive condition could have produced arousal that exceeded this level for all tasks, masking any difference between them.
IV. Experiment 2

Design

Experiment 1 was conducted in India, which enabled us to offer significant monetary incentives on a relatively modest budget. While the results suggest that very high incentives can be detrimental, it does suffer from some limitations that we sought to address in a follow-up study.

One limitation was that subjects in experiment 1 were unfamiliar with most of the tasks, raising the question of whether incentives would have different effects for more familiar tasks or tasks on which people had had an opportunity to practice. Second, the experimental setup in experiment 1 was based on a between-subject design, so that all subjects completed all six tasks under the same incentive conditions. Clearly, it would be interesting to see if the same subjects would exhibit different levels of performance when confronted with different levels of incentives. A within-subject design is also useful in examining whether the results of experiment 1 could be attributed to inferences. It is possible that subjects relied on the level of incentive to infer the difficulty of the task. A within-subject design would alleviate concerns related to such inferences. Third, the unusual nature of the subject population raises questions about whether the results would generalize to people who were more used to conditions in an advanced capitalist country. Responding to all of these issues, experiment 2 was conducted at MIT with twenty-four undergraduate students, using two tasks that are more familiar to the participants, with practice trials for both tasks before the start of the experiment, using a within-subject design (in which each subject received both the high and the low levels of both treatments).
A surprise from the first experiment was the failure to observe different effects of high incentives for different types of tasks, contrary to the rationale that guided the inclusion of the different tasks. In the second experiment, we attempted to address this issue again, by including one task that requires only physical effort and another that required mainly cognitive skills. The two tasks were adding and keypressing. In the adding task respondents were given a set of 20 matrices one at a time, with 12 numbers in each matrix (see Figure III for a sample), and were asked to find the two numbers in that matrix that would add to 10. In the key-pressing task respondents were asked to alternate between pressing the “v” and “n” keys on the keyboard. We used these tasks because they are based on simple elementary aspects of performance: adding two numbers and typing - tasks that are very familiar to our respondents. One other important aspect of these tasks is that, while the adding task requires cognitive effort, the key pressing one requires only pure physical effort without any need for cognitive resources. Thus, we should be able to examine the first postulate – that high performance-contingent incentives increase pure effort and, as a consequence, improve performance that is based solely on pure effort – as well as the second postulate – that high performance-contingent incentives can decrease performance that is based on cognitive skills. We therefore expected an improvement in performance for the key pressing task when the stakes were high. However, because the addition task required cognitive resources and effort, we predicted that increased incentives would lead to a decrement in performance on this task.

Finally, we also incorporated a slightly more complex reward structure. For the addition task, performance was measured by the number of matrices that were solved correctly in 4 minutes. For the key-pressing task, performance was measured by the number of alternations done in 4 minutes – a deliberately numbingly boring task. The low incentive for the adding task
was $0 if respondents solved 9 or fewer matrices, $15 if respondents solved 10 matrices, and an additional $1.50 for each additional matrix solved to a maximum of $30. The high incentive for the adding task was ten times higher (0, $150, $300). The low incentive for the keypressing task was $0 if respondents pressed 599 alternations or less, $15 if respondents pressed 600 alternations, and an additional $0.10 for each additional alternation (based on pilot testing we expected the maximum to be 750 alternation, which would equal a payment of $30. The high incentive for the adding task was ten times higher (0, $150, $300).

The experiment was conducted toward the end of the semester, a time when the students have usually depleted their budget and are thus more strapped for cash. When respondents first came to the lab they were given instructions for the adding task, and were given four minutes to perform this task without any incentives. Next, they were given instructions for the keypressing task and were given four minutes to perform this task without any incentives. After this initial practice with the tasks, half of the respondents were given the same two tasks in the same order with low incentives, and the other half were given the same two tasks in the same order with high incentives. After finishing the first set of tasks-for-pay, each respondent was given the same two tasks in the same order for the other level of incentives (the level they had not yet experienced).

*Results*
In line with the analysis of Experiment 1, the main dependent variable in our analysis was, for each task, the subject's earnings as a fraction of total possible earnings for that task (% of $30 in the low incentive condition and % of $300 in the high incentive condition). The results were analyzed with a 2 (incentive level: high and low) by 2 (task: adding and keypressing) by 2 (order of the two incentives: low-high and high-low) mixed between subjects (order) and within subjects (incentive level and task) repeated-measure ANOVA.

As can be seen from Figure IV, the results for the adding task replicated the basic results from Experiment 1, with performance decreasing as a function of stakes, while the results from the keypressing task showed an increasing relationship between the level of incentives and performance. The analysis revealed a significant interaction between incentive level and task [F(1, 91) = 27.73, p < 0.001], a marginal effect for incentive level [F(1, 91) = 3.02, p = 0.086], a non-significant effect for task [F(1, 91) = 0.82, p = 0.37], and a non-significant effect for the order of the incentive levels [F(1, 91) = 0.21, p = 0.65]. Follow-up tests showed that in the keypressing task increasing incentives caused a significant increase in performance [F(1, 91) = 24.73, p < 0.001], while in the adding task increasing incentives caused a significant decrease in performance [F(1, 91) = 6.28, p = 0.014].

These findings provide additional support for the main hypothesis that motivated the current work – namely that additional incentives can decrease performance. Combined with the results from Experiment 1, these results also show that such negative returns to incentives can appear in tasks that respondents are generally familiar with (adding numbers), and even when they had some practice with the specific tasks. The results also show that the order of the two incentive levels gives rise to the same basic pattern of results – suggesting that the effects are not due to inferences respondents draw based on the level of reward. Finally, the increased
performance with the high incentive level in the key pressing task shows an important boundary condition for the applicability of these results. Tasks that involve only effort are likely to benefit from increased incentives, while for tasks that include a cognitive component, there seems to be a level of incentive beyond which further increases can have detrimental effects on performance.

V. Experiment 3

Design

Experiments 1 and 2 demonstrated that large contingent financial incentives can sometimes decrease performance. In Experiment 3 we extend the scope of investigation to examine social, as opposed to financial, incentives. Specifically, we examine the impact on performance of having an audience watch one work on a cognitive task. Although audience effects might seem at first glance to be non-economic in nature, in fact there are many tasks of great economic significance that are performed under conditions of public scrutiny. Determining whether the increased motivation brought by an audience improves or detracts from performance, therefore, not only provides more basic evidence on the relationship between performance and motivation, but could also have ramifications in applied settings.

The experiment took about 30 minutes and was conducted in five sessions at the University of Chicago. Four of the sessions had eight participants, and one session had seven participants. Upon arriving, each student received instructions in which he/she was told that they would be participating in an experiment of problem solving, and that the task in the experiment was to solve anagrams. It was explained that anagrams are jumbled letters that can be made into one,
and only one, very common word. Following the instructions participants had a one-minute trial in which they were asked to solve three examples of anagrams. At the end of the practice trial the correct answers were revealed.

The experiment was based on 26 trials, each consisting of one minute to solve three anagrams. The important feature of the design is that in the 10 private trials all participants worked without being observed by anyone, while in the 16 public trials, one participant chosen at random worked in plain sight of the other participants. In the public trials, a random number was drawn and the corresponding participant stood next to the experimenter and attempted to solve the anagrams in front of the entire group, using a larger version of the same page that was used when anagrams were solved in private. While that participant was solving the anagrams, the other members of the groups observed the anagrams, the participant who was trying to solve them, and his/her success.

The sequence of trials alternated between two private trials (in which everyone solved two sets of three anagrams), and four public trials (in which four different participants, got up one at a time and each solved one set of three anagrams). Payment was 33 cents for every anagram successfully solved, whether in a private or public round. In addition, each participant received a flat $5 for showing up.

Results

The main interest in this experiment is the number of solved anagrams across the two conditions. Because the anagram task involves creativity, and because we thought that solving the anagrams in front of others would produce high levels of motivation, we predicted that the
public condition would lead to the anticipated choking under pressure on this task. In addition, prior results by Gneezy, Niederle and Rustichini [2003] suggest that men are much more responsive to competitive incentives than women, raising the question of whether there might be a gender difference in the tendency to choke under these conditions.

To test both of these questions we analyzed the average number of correctly solved anagrams per trial type (private vs. public) and the respondent’s gender in a mixed design ANOVA, with the type of trial (private vs. public) as a within subjects factor, and gender as a between subjects factor. As can be seen in Figure V, the results showed a significant effect for the type of trial \([F(1,37) = 10.14, p = 0.003]\), with the average number of anagrams solved correctly higher in the private condition \((M=1.16)\) than in the public condition \((M=0.67)\). There was, however, no evidence of any gender difference in ability to solve anagrams, nor any evidence of differential tendency for the two genders to be differentially influenced by the social pressure. The average number of anagrams solved per trial was 1.17 for men and 1.15 for women in the private condition, and 0.64 for men and 0.69 for women in the public condition.

Figure V

VI. General Discussion

Many institutions provide very large incentives for exactly the types of tasks we used here – those that require creativity, problem solving, and concentration. Our results challenge the assumption that increases in motivation necessarily lead to improvements in performance. In 8 of the 9 tasks we examined across the three experiments, higher incentives led to worse performance. In fact, we were surprised by the robustness of the effect; we had expected some of
the 6 tasks included in the first experiment to respond in a positive monotonic fashion to level of incentive.

Do administrators who are in charge of setting compensation have greater insight into such effects? The prevalence of very high incentives contingent on performance in many economic settings raises questions about whether administrators base their decisions on empirically derived knowledge of the impact of incentives or whether they are simply assuming that incentives enhance performance.

One possible interpretation of our results is that incentives may not always be implemented optimally. However, it is possible that there may be reasons for such incentives other than the desire to elicit maximum levels of performance. For example, in athletic competitions, it is possible that the negative effects of high payments on performance are widely recognized, and that this negative effect of incentives on performance actually creates excitement on the part of audiences. However, you would think that, for example, having your home team win the world series would be more exciting than watching them choke under pressure.

It is also possible, even if high incentives fail to improve the performance of those at the top of the income hierarchy, that they could still increase motivation for rank and file workers who are not actually facing high incentives but are motivated by the prospect of doing so. However, again there would seem to be better solutions to this problem, such as simply paying top workers a higher fixed wage.

Our results also have implications for the debate between proponents and opponents of behavioral economics. One of the common criticisms of behavioral economics is that observed anomalies are unlikely to occur when the stakes are high [Thaler, 1986]. Although people's performance undoubtedly improves in some situations as the stakes increase, the results of the
experiments reported here suggest, at a minimum, that high incentives cannot be relied upon to produce optimal behavior.

The fact that some of our tasks revealed non-monotonic relationships between effort and performance of the exact type predicted by the “Yerkes-Dodson law” does caution against generalizing results obtained with one level of incentives to levels of incentives that are radically different (see, e.g., Parco, Rapoport and Stein, 2002). For many tasks, introducing incentives where there previously were none or raising small incentives on the margin is likely to have a positive impact on performance. This could be true even when the level of incentives is high (Ehrenberg and Boganno, 1990; Lazear, 2000). Our experiment suggests, however, that one cannot assume that introducing or raising incentives always improves performance. It now appears that beyond some threshold level, raising incentives may increase motivation to supra-optimal levels and result in perverse effects on performance. Given that incentives are generally costly for those providing them, raising contingent incentives beyond a certain point may be a losing proposition. Perhaps there is good reason why so many workers continue to be paid on a straight salary basis.


Gneezy, Uri, and Aldo Rustichini, “Pay enough or don't pay at all,” The Quarterly Journal of Economics, CXV (August 2000a), 791-810.

Gneezy, Uri, and Aldo Rustichini, “A fine is a price,” Journal of Legal Studies, XXIX (January 2000b), 1-18.

Heyman, James, and Dan Ariely, “Effort for Payment: A Tale of Two Markets,” Psychological Science, Forthcoming.


Yerkes, Robert M., and John D. Dodson, ”The relationship of strength of stimulus to rapidity of habit-formation,” *Journal of Comparative Neurology of Psychology*, XVIII (1908), 459-482.

<table>
<thead>
<tr>
<th>Games</th>
<th>Mean raw score (Std.)</th>
<th>% at least good</th>
<th>% very-good</th>
<th>% earnings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Mid</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Packing Quarters</td>
<td>202.0 (65.4)</td>
<td>185.7 (70.5)</td>
<td>235.9 (12.9)</td>
<td>28.6</td>
</tr>
<tr>
<td>Simon</td>
<td>6.5 (2.1)</td>
<td>6.3 (1.4)</td>
<td>5.2 (1.4)</td>
<td>64.2</td>
</tr>
<tr>
<td>Recall last 3-digits</td>
<td>4.9 (2.7)</td>
<td>5.5 (2.8)</td>
<td>4.6 (2.4)</td>
<td>64.3</td>
</tr>
<tr>
<td>Labyrinth</td>
<td>5.9 (2.5)</td>
<td>4.6 (1.8)</td>
<td>4.1 (1.8)</td>
<td>64.3</td>
</tr>
<tr>
<td>Dart Ball</td>
<td>2.8 (2.0)</td>
<td>3.6 (2.6)</td>
<td>2.9 (1.7)</td>
<td>25.0</td>
</tr>
<tr>
<td>Roll up</td>
<td>1.8 (2.1)</td>
<td>1.8 (3.1)</td>
<td>1.2 (1.5)</td>
<td>25.0</td>
</tr>
</tbody>
</table>
Table II
Planned Contrasts Across Treatments for Each of the Six Games: Marginally significant differences (p < 0.1) are underlined and significant differences (p < 0.05) are bold.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Measure</th>
<th>Earnings</th>
<th>Simon</th>
<th>Dart Ball</th>
<th>Recall</th>
<th>Roll-Up</th>
<th>Labyrinth</th>
<th>Packing Quarters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low - Mid</td>
<td>t-value</td>
<td></td>
<td>0.171</td>
<td>1.372</td>
<td>0.132</td>
<td>0.142</td>
<td>1.963</td>
<td>0.968</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td></td>
<td>0.865</td>
<td>0.176</td>
<td>0.896</td>
<td>0.887</td>
<td>0.053</td>
<td>0.337</td>
</tr>
<tr>
<td>Low - High</td>
<td>t-value</td>
<td></td>
<td>2.631</td>
<td>0.524</td>
<td>1.272</td>
<td>1.408</td>
<td>3.288</td>
<td>2.453</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td></td>
<td><strong>0.010</strong></td>
<td>0.602</td>
<td>0.207</td>
<td>0.166</td>
<td><strong>0.001</strong></td>
<td><strong>0.020</strong></td>
</tr>
<tr>
<td>Mid - High</td>
<td>t-value</td>
<td></td>
<td>2.505</td>
<td>0.951</td>
<td>1.427</td>
<td>1.295</td>
<td>1.364</td>
<td>3.681</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td></td>
<td><strong>0.014</strong></td>
<td>0.346</td>
<td>0.157</td>
<td>0.202</td>
<td>0.176</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Low &amp; Mid - High</td>
<td>t-value</td>
<td></td>
<td>2.966</td>
<td>0.304</td>
<td>1.557</td>
<td>1.704</td>
<td>2.695</td>
<td>4.136</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td></td>
<td><strong>0.004</strong></td>
<td>0.762</td>
<td>0.123</td>
<td><strong>0.092</strong></td>
<td><strong>0.008</strong></td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Measure</th>
<th>Very-Good</th>
<th>Simon</th>
<th>Dart Ball</th>
<th>Recall</th>
<th>Roll-Up</th>
<th>Labyrinth</th>
<th>Packing Quarters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low - Mid</td>
<td>t-value</td>
<td></td>
<td>1.364</td>
<td>1.281</td>
<td>0.474</td>
<td>0.132</td>
<td>2.111</td>
<td>0.689</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td></td>
<td>0.178</td>
<td>0.206</td>
<td>0.638</td>
<td>0.896</td>
<td><strong>0.042</strong></td>
<td>0.493</td>
</tr>
<tr>
<td>Low - High</td>
<td>t-value</td>
<td></td>
<td>2.981</td>
<td>0.500</td>
<td>1.814</td>
<td>2.087</td>
<td>2.087</td>
<td>3.000</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td></td>
<td><strong>0.005</strong></td>
<td>0.619</td>
<td>0.075</td>
<td><strong>0.044</strong></td>
<td><strong>0.044</strong></td>
<td><strong>0.006</strong></td>
</tr>
<tr>
<td>Mid - High</td>
<td>t-value</td>
<td></td>
<td>1.710</td>
<td>1.787</td>
<td>1.357</td>
<td>2.021</td>
<td>0.024</td>
<td>3.808</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td></td>
<td><strong>0.095</strong></td>
<td>0.080</td>
<td>0.180</td>
<td>&lt;0.050</td>
<td>0.981</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Low &amp; Mid - High</td>
<td>t-value</td>
<td></td>
<td>3.157</td>
<td>1.474</td>
<td>1.895</td>
<td>2.688</td>
<td>1.624</td>
<td>4.826</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td></td>
<td><strong>0.002</strong></td>
<td>0.145</td>
<td>0.062</td>
<td><strong>0.009</strong></td>
<td><strong>0.089</strong></td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Figure captions

Figure I: Means of the Share of Earnings Relative to the Maximum Possible Earnings for the Three Payment Levels, Averaged Across the Six Games.

Figure II: Means of the Share of Earnings Relative to the Maximum Possible Earnings for the Three Payment Levels, Plotted Separately by Game. Games Are Indicated by Their Category: Motor Skills (ms), Concentration (co), and Creativity (cr).

Figure III: Sample Screen with Matrix in Adding Task

Figure IV: Means of the Share of Earnings Relative to the Maximum Possible Earnings for Keypressing and Adding.

Figure IV: Frequency Distribution of Average Correct Anagrams For the Public and Private Conditions.
Footnotes

i Some evidence that in sport increased incentives will improve performance are reported in Ehrenberg and Boganno (1990), who find that tournaments have incitves effect. In particular, they show that in golf tournaments the level and structure of prizes positively influences performance. This effect occurs primarily in the later parts of the tournament when fatigue has set in and it is more difficult for players to maintain concentration. In the case of other sports, it is possible that sport federation administrators hold incorrect theories about how rewards affect performance, that audiences get more excited about games when players are highly rewarded, or that the increases in rewards result from competition to attract the best players.

ii The experiment was conducted by local research assistants from Narayanan College at Madurai, India, who were naïve to the hypotheses.

iii The conversion is based on the average exchange rate in December 2001 of Indian Rupee Rs 47.93 = US $1 [see Federal Reserve Statistical Release, 2003].

iv We also tested models, which included socio-demographic variables and their interactions with the payment condition. In no case were the socio-demographic variables significant, and, as a consequence, they are not considered in the analyses we report.

v In another study we gave 60 participants all the information about experiment 1 and asked them to predict the results of the Simon and Packing quarters games. The predictions of the respondents indicated that they expected performance to be positively and monotonically linked to level of contingent reward.

vi In fact, firms could also exploit this phenomenon (rather than acting suboptimally on it). See Della Vigna and Malmendier (forthcoming) and Gabaix and Laibson (2004) for discussion of ways in which firms may exploit consumer biases.
Figure 2
Figure 3

This is Matrix 2 out of 20

no. of correctly solved matrixes: 1

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9.38</td>
<td>6.74</td>
<td>8.17</td>
</tr>
<tr>
<td>5.15</td>
<td>6.61</td>
<td>3.06</td>
</tr>
<tr>
<td>9.71</td>
<td>.91</td>
<td>4.88</td>
</tr>
<tr>
<td>3.58</td>
<td>4.87</td>
<td>6.42</td>
</tr>
</tbody>
</table>

Next