Working Papers
Federal Reserve Bank of Boston

No. 05-11

Large Stakes and Big Mistakes

Dan Ariely, Uri Gneezy, George Loewenstein, and Nina Mazar

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 rural India. Subjects in our experiment worked on different tasks and received performance-contingent payments that varied in amount from small to large relative to their typical levels of pay. With some important exceptions, we observed that high reward levels can have detrimental effects on performance.

JEL Classifications: D00 microeconomics, general
Keywords: performance-based incentives

Dan Ariely is Luiz Alvarez Renta Professor of Management Science at MIT and a visiting scholar at the Federal Reserve Bank of Boston's Research Center for Behavioral Economics and Decision-Making. His email address is ariely@mit.edu. Uri Gneezy is Associate Professor of Behavioral Science at the University of Chicago Graduate School of Business. George Loewenstein is Professor of Economics and Psychology in the Department of Social and Decision Sciences at Carnegie Mellon University. Nina Mazar is Post-Doctoral Associate and Lecturer in Marketing at the Sloan School of Management at MIT.

This paper, which may be revised, is available on the web site of the Federal Reserve Bank of Boston at http://www.bos.frb.org/economic/wp/index.htm.

The views expressed in this paper are those of the authors and do not necessarily represent the views of the Federal Reserve Bank of Boston or the Federal Reserve System.

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. 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 Christoher 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.

This version: July 23, 2005

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 efforts and the output of their labor. Some recent evidence suggests that payment for performance can indeed increase performance. For example, Lazear (2000) showed that a large company that, under new management, moved from hourly wages to piece-rate pay, increased productivity by a dramatic 44 percent [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 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). While the ratcheting up of rewards seems to be premised on the belief that doing so will improve performance, it does not provide positive evidence that magnifying the rewards in this fashion actually has the intended effect.

The expectation that people will improve their performance when given high performance-contingent incentives rests on two subsidiary assumptions: (1) that increasing performance-contingent incentives will increase motivation and effort, and (2) that this increase in motivation and effort will result in improved performance.

The first assumption, that transitory performance-based increase in pay is increasing 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 experiments, students who were collecting donations door-to-door actually visited fewer houses and collected less money when they were paid a small commission [Gneezy and Rustichini (2000a); see also Frey and Jegen (2001); Heyman and Ariely (2004)]. 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 first assumption regarding the positive relationship between effort and pay, our focus in this paper is on the second assumption. The experiments we report, therefore, address the question of whether increased effort necessarily leads to improved performance. Providing subjects with different levels of incentives, including ones 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 / effort and pay, the relationship between motivation / effort and performance has not attracted much attention from economists, perhaps because the belief that 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 facing a test of discriminating safe from unsafe (that is, shock-inducing) areas in a cage. The results showed that the rats learned to discriminate most quickly when the shocks were at an intermediate level of intensity [for similar evidence in humans see Neiss (1988)]. Since arousal is tightly linked to motivation and performance, these findings imply that increases in motivation beyond an optimal level will tend to 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 a perverse relationship between motivation and performance. For example, one mechanism via which increased motivation can backfire is when it leads to greater self-consciousness. When performance on a task relies on highly practiced, automatic skills [Baumeister (1984); Langer and Imber (1979)], increasing awareness, competition, introducing a cash incentive or audience or ego-relevant threats (the belief that a task is diagnostic of something that one cares about, such

as intelligence) can cause people, involuntarily, to consciously think about the task, shifting control from 'automatic' to 'controlled' processes that are less effective [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 numerous 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 mechanisms of shifting from 'automatic' to 'controlled' processes can also account for why the presence of an audience, which tends to increase motivation to perform well, and hence conscious monitoring and control on the process of task-performance, can be so destructive [see also Zajonc (1965)].

A second mechanism by which increased motivation is also likely to have a negative effect on performance relates to a general focus of attention. Attentional focus can be detrimental for tasks that involve insight or creativity, since increased motivation tends to narrow individuals' focus of attention [Easterbrook (1959)], and creativity and insight require drawing 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 many sources [see Baumeister and Showers (1986)] that can lead to choking under pressure; among them are competition and

competitiveness, the introduction of monetary rewards, the presence of an audience, and ego-relevant threats.

For economics, however, the most interesting determinant of performance pressure is the level of performance-contingent monetary incentives, and in particular the effects of substantial incentives more common in the workplace. Our primary goal in the studies reported herein is to test, in experiments that satisfy the standard experimental economics criteria, the effects of relatively large incentives – examining whether increasing incentives beyond a certain point 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. However, all six tasks require at least some strategy and cognitive effort. 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 one task with and one task without any need for strategy or cognitive effort. As will be seen, the predicted differences emerged in this case. Finally, in our third experiment, we extend our 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.[i] The sample consisted of 26.4 percent females and 73.6 percent males. The majority of participants (90.8 percent) were Hindu, 5.7 percent were Christians, and 3.4 percent were Muslims. Their standard of living can be best described by our participants' level of education and their possessions. Participants in this experiment had, on average, 5.6 years of education, and 26 percent had no formal education. Approximately half of the participants reported that they owned a TV (M = 49.4 percent), and about half owned a bicycle for transportation (M = 51.7 percent). None owned a car, and only 6.9 percent 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 six different games in a random order and were promised a payment for each game, if they reached certain performance levels. The magnitude of the payment depended on the treatment and whether, for <u>each</u> game, they reached either of two specified performance levels which we labeled "good" and "very good." Participants received full payment (that is, 4, 40, or 400 Indian Rupees, 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.

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)].[ii] 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 nine metal pieces of quarter circles into a black wooden frame within a given time. It is easy to fit eight pieces, but, to fit all nine, 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 one 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 six consecutive lights. The "very good" performance level was defined by at least one repetition of eight 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 three digits. Participants had 14 trials in this task. The "good" performance level was defined by at least four correct trials. The "very good" performance level was defined by at least six 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 tilted 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 seventh hole. The "very good" performance level was defined by passing the ninth 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 five balls hit the center of the target. The "very good" performance level was defined by having at least eight 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 four balls hit the furthest hole. The "very good" performance level was defined by having at least six 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 a psychological perspective is the probability of reaching the “very good” performance level, since this is the performance level that represents the highest possible performance and payment. From an economics perspective, the most interesting measure is the fraction of possible earnings since it represents the measure that is most closely linked to the incentives that the subjects actually faced.[iii] In what follows, therefore, we present all results in terms of these two measures (“very good” performance and earnings).

••• Table I •••

To examine performance, we analyzed the data with a three (incentive levels) by six (games) mixed between subjects (incentive levels) and within subjects (games) repeated-measure analysis of variance (ANOVA). For both of our measures, this overall model revealed a significant effect for payment condition [Earnings: $F(2, 84) = 10.24$, $p < 0.001$; Very-Good: $F(2, 84) = 13.48$, $p < 0.001$], a significant effect for game [Earnings: $F(5, 420) = 9.22$, $p < 0.001$; Very-Good: $F(5, 420) = 4.35$, $p = 0.001$], and a nonsignificant interaction between them [Earnings: $F(10, 420) = 1.24$, $p = 0.263$; Very-Good: $F(10, 420) = 1.21$, $p = 0.28$]. The nonsignificant 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 six 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 percent), but higher and almost equal in the mid (M = 36.7 percent) and low payment conditions (M = 35.4 percent). Similarly, the average share of games (out of six) in which respondents reached the very-good performance level was lowest in the high payment condition (M = 6.3 percent), higher in the mid payment condition (M = 22.2 percent), and highest in the low payment condition (M = 25.6 percent).Post-hoc Fisher LSD tests for both measures revealed that the difference between the low and mid payment condition was not significant [earnings: $p = 0.768$; very good: $p < 0.396$], while the difference between the low and high condition was significant [earnings: $p < 0.001$; very good: $p < 0.001$] as was the difference between the mid and high condition [earnings: $p < 0.001$; very good: $p < 0.001$]. These findings support the main hypothesis that motivated the experiment – namely, that additional incentives can decrease performance.

••• Figure I •••

Somewhat contrary to our expectations, the pattern of results seems to hold across a wide range of 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 and for each of our two main dependent measures (very good and earnings), 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 contrasts of the low and mid levels of incentives revealed little difference in performance: For the earning-dependent measure, only one of the games (Labyrinth) showed a marginally significant effect. For the "very good" dependent measure, only one of the games (Labyrinth) showed a 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 for the earnings measure 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. The contrasts between the high payment condition and the low payment condition for the very-good measure were even more differentiated. The contrasts were significant at the 0.05 level for three of the six games (Simon, Roll-Up, Packing-Quarters), marginally significant at the 0.1 level for 2 games (Recall last-3 digits, Labyrinth), and not significant for one game (Dart-Ball).

••• Table II •••

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 – the same pattern as observed with the motor skill tasks and the creativity task.

There are a number of possible reasons that might explain why we did not observe different patterns of results for the two concentration tasks. One is that we may have inadvertently chosen only tasks for which excessive concentration is harmful; while this is possible, it is not very likely. As discussed in the introduction, psychologists have documented numerous situations in which trying to accomplish a task, such as carrying a full coffee mug, produces exactly the opposite of the intended result [Wegner, Ansfield, and Pilloff (1998)]. Another is that the incentives we chose may have simply been too high. Different tasks most likely have different optimal levels of arousal, and it is possible that the concentration tasks have a higher level of optimal arousal. Our choice of the levels of incentives in the three conditions, and particularly in the high-incentive condition could have produced arousal that exceeded even this optimal level – masking the relative advantage of arousal for these tasks.

Overall, the results point to two main conclusions: *First*, with the exception of one case (that is, Labyrinth) there was no (marginally) significant difference in the performance between the low- and mid-payment conditions. Thus, despite the relatively large difference in magnitude of reward across the treatments (that is, 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 already a level of performance that was the highest respondents could master, and, therefore, the increased reward had no incremental effect. *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 three of the six games[iv].

IV. Experiment 2

Design

Experiment 1 was conducted in India, enabling us to offer significant monetary incentives on a relatively modest budget. While the results suggest that very high incentives can be detrimental, there are a few experimental robustness checks that are in order. Experiment 2 was conducted at MIT with 24 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), using one task that required only effort and one that required mainly cognitive skills, and using a slightly more complex reward structure. The experiment was conducted toward the end of the semester, a time when the students have usually depleted their budgets and are thus more strapped for cash.

The two tasks were: adding and key-pressing. In the adding task, respondents were given a set of 20 matrixes 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. Performance was measured by the number of matrixes that were solved correctly in four minutes. In the key-

pressing task, respondents were asked to alternate between pressing the "v" and "n" keys on the keyboard. Performance was measured by the number of alternations done in four minutes. 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 resources and 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 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.

••• Figure III •••

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 key-pressing 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 he or she had not yet experienced).

The low incentive for the adding task was $0 if respondents solved 9 or fewer matrixes, $15 if respondents solved 10 matrixes, 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 key-pressing 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 alternations, which would equal a payment of $30. The high incentive for the adding task was ten times higher (0, $150, $300).

Results

In line with the analysis of Experiment 1, we examined the results once by the probability of reaching the threshold for getting any reward, and once by earnings as a fraction of total possible earnings in each task. Each type of data was analyzed in a two (incentive level: high and low) by two (task: adding and keypressing) by two (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 key-pressing task showed an increasing relationship between the level of incentives and performance. The analysis of whether participants reached the threshold for any payment (10 solved matrixes or 600 alternations) revealed a significant interaction between incentive level and task [$F(1, 91) = 19.08$, $p < 0.001$], a marginal effect for incentive level [$F(1, 91) = 2.68$, $p = 0.1$], and a nonsignificant effect for task [$F(1, 91) = 0.3$, $p = 0.59$], and a nonsignificant effect for

the order of the incentive levels [$F(1, 91) = 0.21$, $p = 0.65$]. Follow-up tests showed that in the key-pressing task, increasing incentives caused a significant increase in performance [$F(1, 91) = 18.19$, $p < 0.001$]; while in the adding task, increasing incentives caused a marginally significant decrease in performance [$F(1, 91) = 3.76$, $p = 0.056$]. The analysis of the fraction of earnings from the total possible earnings in each setting (percent from $30 in the low-incentive condition and from $300 in the high-incentive condition) revealed a significant interaction between incentive level and task [$F(1, 91) = 27.73$, $p < 0.001$], a marginal effect for incentive level [$F(1, 91) = 3.02$, $p = 0.086$], a nonsignificant effect for task [$F(1, 91) = 0.82$, $p = 0.37$], and a nonsignificant effect for the order of the incentive levels [$F(1, 91) = 0.21$, $p = 0.65$]. Follow-up tests showed that in the key-pressing 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. Adding to 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 typing), and even when they have 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.

••• Figure IV •••

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 noneconomic 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 consisted of 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 page 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 (where everyone solved two sets of three anagrams), and four public trials (where 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, leading to 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 found to be much higher in the private condition (M=1.16), compared with 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 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 existing 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 eight of the nine 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 six 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, one would think that, for example, having the home team win the world series would be more exciting than watching them choke under pressure.

It is also possible that, even if high incentives fail to improve the performance of those at the top of the income hierarchy, 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.

The fact that some of our tasks revealed nonmonotonic relationships between effort and performance of the exact type predicted by the "Yerkes-Dodson law" further cautions against generalizing results obtained with one level of incentives to levels of incentives that are radically different. 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. 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.

References

Baumeister, Roy F., "Choking under pressure: Self-Consciousness and Paradoxical Effects of Incentives on Skillful Performance," *Journal of Personality and Social Psychology*, XLVI (1984), 610-620.

Baumeister, Roy F., and Carolin J. Showers, "A review of paradoxical performance effects: Choking under pressure in sports and mental tests," *European Journal of Social Psychology*, XVI (October-December 1986), 361-383.

Camerer, Colin F., Linda Babcock, George Loewenstein, and Richard H. Thaler, "Labor Supply of New York City Cab Drivers," *The Quarterly Journal of Economics*, CXII (May 1997), 407-441.

Camerer, Colin F., George Loewenstein, and Drazen Prelec, "Neuroeconomics: How neuroscience can inform economics," *CMU Working Paper*, April 2004.

Dandy, Justine, Neil Brewer, and Robin Tottman, "Self-consciousness and performance decrements within a sporting context," *Journal of Social Psychology*, CXLI (February 2001), 150-152.

Easterbrook, J.A., "The effect of emotion on cue utilization and the organization of behavior," *Psychological Review*, LXVI (1959), 183-201.

Federal Reserve Statistical Release, "Foreign Exchange Rates (monthly)," *Website: http://www.federalreserve.gov/Releases/G5/20030102/* (January 2, 2003).

Frey, Bruno S., and Reto Jegen, "Motivation crowding theory," *Journal of Economic Surveys*, XV (December 2001), 589-611.

Gneezy, Uri, Niederle, Muriel, and Aldo Rustichini, "Performance in competitive environments: Gender differences," *The Quarterly Journal of Economics*, CXVIII (August 2003), 1049-1074.

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.

Langer, Ellen J., and Lois G. Imber, "When practice makes imperfect: the debilitating effects of overlearning," *Journal of Personality and Social Psychology*, XXXVII (1979), 2014-2024.

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

McGraw, Kenneth .O., and John C. McCullers, "Evidence of a Detrimental Effect of Extrinsic Incentives on Breaking a Mental Set," *Journal of Experimental Social Psychology*, XV (1979), 285-294.

Neiss, Rob, "Reconceptualizing Arousal: Psychological States in Motor Performance," *Psychological Bulletin*, CIII (May 1988), 345-366.

Prendergast, Canice, "The Provision of incentives in firms," *Journal of Economic Literature*, XXXVII (March 1999), 7-63.

Rangachari, Dilip, "Poverty down but urban-rural divide sharp. The Times of India," *Website: http://timesofindia.indiatimes.com/articleshow/40894515.cms* (March 20, 2003).

Slam! Sports, "From England to Korea, soccer federations offering rich reward for success," *Website: http://cgi.canoe.ca/Slam020514/soc_bonus-ap.html* (May 14, 2002).

Wegner, Daniel M., Matthew E. Ansfield, and Daniel Pilloff, "The putt and the pendulum: Ironic effects of the mental control of action," *Psychological Science*, IX (May 1998), 196-199.

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.

Zajonc, Robert B., "Social Facilitation" *Science*, CXLIX (July 1965), 269-27.

Table I

Performance by Game and Treatment Presented as Raw Scores, Percent of Individuals who Reached at Least the Good and the Very Good Performance Levels, and Percent of Maximal Earnings

	Mean raw score (Std.)			% at least good			% very good			% earnings		
Games	Low	Mid	High	Low	Mid	High	Low	Mid	High	Low	Mid	High
Packing Quarters	202.0 (65.4)	185.7 (70.5)	235.9 (12.9)	28.6	43.3	10.3	25.0	33.3	0	26.8	38.3	5.2
Simon	6.5 (2.1)	6.3 (1.4)	5.2 (1.4)	64.2	76.7	44.8	32.1	16.7	3.5	48.2	46.7	24.2
Recall last-3 digits	4.9 (2.7)	5.5 (2.8)	4.6 (2.4)	64.3	73.3	58.6	42.9	36.7	20.7	53.6	55	39.7
Labyrinth	5.9 (2.5)	4.6 (1.8)	4.1 (1.8)	64.3	50.0	27.6	21.4	3.3	3.5	42.9	26.7	15.6
Dart Ball	2.8 (2.0)	3.6 (2.6)	2.9 (1.7)	25.0	40.0	37.9	10.7	23.3	6.9	17.9	31.7	22.4
Roll up	1.8 (2.1)	1.8 (3.1)	1.2 (1.5)	25.0	23.3	17.2	21.4	20.0	3.5	23.2	21.7	10.4

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.

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

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

Figure captions

Figure I: Means of the Two Main Dependent Measures for the Three Payment Levels, Averaged Across the Six Games.

Figure II: Means of the Two Main Dependent Measures 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 Two Main Dependent Measures for Key-pressing and Adding.

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

Endnotes

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

[ii] 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].

[iii] 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.

[iv] 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.