Journal of Experimental Psychology:
Learning, Memory, and Cognition
1991, Vol. 17, No. 4, 734-752

0278-7393/91/$3.00

Learning From Feedback: Exactingness and Incentives

Robin M. Hogarth
Graduate School of Business, Center for Decision Research
University of Chicago

Brian J. Gibbs
Graduate School of Business
Stanford University

Craig R. M. McKenzie
University of Chicago

Margaret A. Marquis
Department of Statistics
North Carolina State University

In a series of five experiments, exactingness, or the extent to which deviations from optimal decisions are punished, is studied within the context of learning a repetitive decision-making task together with the effects of incentives. Results include the findings that (a) performance is an inverted-U shaped function of exactingness, (b) performance is better under incentives when environments are lenient but not when they are exacting, (c) the interaction between exactingness and incentives does not obtain when an incentives function fails to discriminate sharply between good and bad performance, and (d) when the negative effects of exactingness on performance are eliminated, performance increases with exactingness.

The manner in which different types of feedback affect learning has long been of central concern in psychological studies of decision making (see, e.g., Balzer, Doherty, & O'Connor, 1989). In this article, we examine learning within the context of a repetitive decision-making task, and we examine a dimension of feedback that has received little attention to date, namely, the effects of differences in the severity with which performance is evaluated.

Central to our work is recognition of the inherent ambiguity of feedback. In particular, we note that feedback from the outcomes of decisions can serve two functions that are often confounded. One function is inferential. Feedback informs the decision maker about the structure of the underlying task. For example, when a student writes a paper, feedback in the form of a grade provides information about how to write a good paper. The second function is evaluative. Feedback provides information about the student's performance. It tells us whether the performance was good or bad. Note, however, that the feedback—in this case a grade—is confounded. To what extent does the grade reflect the student's ability to write papers, and to what extent does it reflect the teacher's grading policy?

Evaluation of decision-making performance can differ on a dimension that we term the *exactingness* of the environment and that reflects the severity of penalties imposed for errors. Tasks are exacting to the extent that deviations from optimal decisions are heavily punished and lenient to the extent that they are not.

This work was funded by a contract from the Office of Naval Research as well as by special funds from the Graduate School of Business, University of Chicago. We are grateful to Colin Camerer, Terry Connolly, Joshua Klayman, George Loewenstein, and Kenneth Hammond as well as several excellent referees for critical comments on earlier drafts.

Correspondence concerning this article should be addressed to Robin M. Hogarth, University of Chicago, Graduate School of Business, 1101 East 58th Street, Chicago, Illinois 60637.

In addition to exactingness, decision-making tasks can vary in the extent to which different levels of performance have consequences for the decision maker. In the case of the student essay, for example, the student may or may not perceive the grade as consequential (e.g., by affecting chances of admission to graduate school). In other words, tasks can vary in the extent to which decision makers have incentives to perform well.

There are several reasons for studying the effects of exactingness together with the effects of incentives. First, knowing when and how exactingness and incentives affect learning is important at a practical level. In business or in the military, for example, what levels of exactingness implied by different evaluation schemes promote efficient learning? Do real consequences in terms of money or lives help people learn to make decisions more effectively? If exactingness or incentives are detrimental, how can learning be structured to overcome these impediments? Second, despite the importance of exactingness in many real world tasks, little theoretical attention has been directed toward understanding its effects. Third, and also from a theoretical viewpoint, controversy exists as to whether incentives necessarily improve performance. For example, one could argue from naive behaviorist or economic viewpoints that incentives will always improve performance, and much evidence is consistent with this contention. However, there is also evidence that under some conditions incentives can be detrimental (see, e.g., Lepper & Greene, 1978).

The article is organized as follows. We first elaborate on the concept of exactingness and comment on the literature that has considered the link between incentives and performance. Next, we outline the theoretical framework, arguing that exactingness induces forces that both help and hinder learning as measured by performance and that incentives accentuate the effects of these opposing forces. This leads to predictions about how exactingness will affect learning and about the nature of interactions between exactingness and incentives. The theoretical framework is then tested in a series of five experiments. Finally, we discuss the results of our experimen-

tal work from both theoretical and practical perspectives and make suggestions for further research.

Evaluation and Incentive Schemes

Evaluation as Feedback

The task used in our experiments is similar to many real world situations in that subjects learn from outcome feedback. Although the ambiguity, and even the misleading nature of outcome feedback, has long been recognized (see, e.g., Brehmer, 1980; Einhorn & Hogarth, 1978; Hammond, Summers, & Deane, 1973), we wish to emphasize a specific aspect of this ambiguity, namely, that outcome feedback simultaneously conveys and confounds information concerning both the structure of the underlying task and how well the subject is performing. Thus, on receiving feedback a person may make inferences both about the structure of the task (e.g., how two variables are related), and the level of his or her performance (e.g., better than expected, better than a rival).

We conceive of feedback as being a function of three variables: (a) the specific action taken by the decision maker, (b) the nature of the underlying system governing outcomes, and (c) the manner in which these outcomes are evaluated. To illuminate the distinction between (b) and (c), note that if two otherwise identical tasks differed only in how outcomes were evaluated, a person making the same decisions in both tasks could receive different feedback. However, if the person was ignorant a priori of both the nature of the underlying tasks and how outcomes were evaluated, it would be difficult to attribute differences in feedback to the different evaluation functions as opposed to possible differences in the structures of the underlying tasks.

Incentives

It is common to classify incentives as *internal* or *external.* Internal incentives are any intrinsic motivations that people have to perform well in a task, the source of which can have various origins including, for example, a need to exhibit mastery (White, 1959), pride, or a wish to impress others (for a review, see Deci & Ryan, 1985). External incentives are explicit rewards, such as money, that depend on performance. Our major concern is with external incentives, although we do manipulate internal incentives in one study.

It would be naive to assert that incentives always improve performance. For example, when external incentives are removed for performing a task that people find intrinsically interesting, subsequent interest and performance in the task can decrease (Lepper, Greene, & Nisbett, 1973; Levine & Fasnacht, 1974). In addition, the presence of incentives has been found to reduce the amount of incidental learning that people acquire in cognitive tasks, presumably because attention is focused on the central task that is rewarded (Bahrick, 1954; Bahrick, Fitts, & Rankin, 1952).

The role of incentives has been examined in several different types of decision-making tasks. For our purposes, studies can be categorized as to whether subjects did or did not receive feedback after their decisions. Because the latter provide no opportunity for learning, we consider the former.

One well-studied task is the binary outcome prediction paradigm in which subjects are required to predict which of two signals will appear on each of a series of trials (for an overview, see Luce & Suppes, 1965). Incentives have produced mixed results. Siegel (1961) used two levels of monetary incentives and found that, with the greater level of incentives, the proportion of the time that subjects chose the more frequent signal became quite extreme (.95, which is still not the optimal value of 1). Edwards (1956) also found more extreme responses under incentives, and Tversky and Edwards (1966) found that although incentives changed behavior, behavior was still far from optimal. In general, the results of these and similar experiments are that payoffs affect subjects' behavior in the appropriate direction, but subjects still do not behave as the normative models prescribe.

Arkes, Dawes, and Christensen (1986) used a probabilistic task in which subjects were given a rule that would have enabled them to choose correctly 70% of the time. They found that, with incentives, subjects were more willing to abandon the rule. The result was that they performed worse than those without incentives. (See also Ashton, 1990.)

The literature does not reveal a simple relation between incentives and performance. For tasks that are understood, incentives appear to improve performance. For example, in summarizing many studies, McCullers (1978) pointed out that incentives enhance performance when the latter depends on making "simple, routine, unchanging responses and when circumstances favor the making of such responses quickly, frequently, and vigorously" (p. 14). He continued, however, by noting that the role of incentives is far less clear in tasks that require flexible, open-ended and creative responses. A similar distinction was made by McGraw (1978) between tasks requiring algorithmic or heuristic, problem-solving, mental strategies, on the one hand, and tasks that subjects find attractive or aversive, on the other. McGraw concluded that incentives are detrimental to performance in tasks that subjects find attractive and that require heuristic, problem-solving, mental strategies (cf. Amabile, 1982; McGraw & McCullers, 1979).

In reviewing work on processes of social facilitation, Zajonc (1965) offered the hypothesis that conditions of arousal tend to enhance the emission of dominant responses. Thus, although incentives should lead people to perform well at tasks with which they are familiar, they can accentuate the probability of producing incorrect responses in unfamiliar settings. Similarly, Easterbrook (1959) summarized a vast psychological literature showing that under high drive states people restrict attention to limited ranges of available cues and that this can inhibit performance in cognitive tasks (see also Kahneman, 1973).

More recently, several researchers have adopted a similar explanation as to why incentives may lead to worse performance when learning complex tasks (Humphreys & Revelle, 1984; Kanfer & Ackerman, 1989; Wood, Bandura, & Bailey, 1990). This is that, in the presence of incentives, complex tasks divert needed attention from inference to evaluation,

that is, from a concern about *how* to do the task to *how well* one is doing. In tasks that are understood, however, attention can be more profitably allocated to executing known strategies.

Theoretical Framework

The Specific Task

The structure of our task is similar to that used in many single- and multiple-cue probability learning studies (see, e.g., Klayman, 1988). Over a series of trials, subjects are presented with information in the form of cues or predictor variables and are asked to predict a criterion that is probabilistically related to the cues. Following each prediction, feedback is provided. In our task, subjects observed a value of a variable, W, and then chose a value of a decision variable, Q. Feedback, however, did not consist of observing the correct value of Q. Instead, feedback was provided in the form of evaluation points. Subjects were instructed that their objective was to maximize the number of evaluation points but were given no explanation as to how these were calculated. (More detailed information about the task is provided below in the experimental section of the paper.)

Evaluation points for each trial were calculated according to the formula

$$\text{Evaluation points} = 500 - \alpha(Q - D)^2, \tag{1}$$

in which Q was the subject's response and D was the correct value of the criterion. D had a strong but imperfect correlation with W, the variable observed by subjects before each trial, and can be characterized by the equation

$$D = \beta_0 + \beta_1 W + \epsilon, \tag{2}$$

in which β_0 (= −1,020) and β_1 (= 20) are parameters and ϵ is a random error term.

As illustrated by Equation 1, evaluation points for a given trial are a negative, linear function of the squared error of the subject's decision for that trial. The slope coefficient, α, makes operational the concept of exactingness, that is, as α increases, so does the penalty associated with erroneous decisions. Lenient environments are therefore characterized by small values of α, and exacting environments by large values.[1]

To compare performance (P) across different levels of exactingness, we measure observed performance (P) on a given trial i by

$$P_i = 500 - |Q_i - D_i|, \tag{3}$$

and consider mean performance, π, across a series of n trials, that is,

$$\pi = \left(\frac{1}{n}\right) \sum_{i=1}^{n} P_i. \tag{4}$$

Exactingness

We hypothesize that exactingness (α) induces forces that have both positive and negative effects on mean performance (π).

The positive aspect of increases in exactingness (α) lies in the opportunities that they provide for learning. To see this, imagine a situation where $\alpha = 0$. In this case, subjects always receive perfect scores of 500 no matter what values they select for Q (see Equation 1). They can therefore never learn what values of Q to associate with W. As α increases, evaluation points become more sensitive to differences between D and Q, thereby providing greater possibilities for learning the relation between W and Q. Learning, however, would not be expected to increase linearly with α. Instead, we hypothesize that the positive aspect of learning (as measured by performance, π) is an increasing, concave function of exactingness (α).

The main negative aspect associated with increases in exactingness (α) lies in the interpretation of feedback and subsequent reactions to this. Specifically, as exactingness (α) increases, feedback in terms of evaluation points is increasingly liable to be negative and, in the absence of alternative points of reference, perceived as such.[2] For example, in a lenient environment with $\alpha = .01$, a difference of 50 between Q and D yields +475 evaluation points (see Equation 1). In an exacting environment with $\alpha = .50$, the same performance translates into −750 evaluation points. In learning environments, people are likely to react differently to positive and negative feedback. Whereas positive feedback reinforces maintaining and refining existing behavior or response strategies (cf. Schwartz, 1982), negative feedback encourages shifting strategies and seeking alternatives that may work better. Because the subset of response strategies that work in exacting environments is much smaller than those that do not, continual shifting of strategies results in lower performance (π)—at least in the short run.[3] We hypothesize that as exactingness (α) increases, the rate at which this negative factor affects learning does not decrease. Thus, the negative aspect of learning (as measured by performance, π) is a nonconcave decreasing function of exactingness (α).

Incentives

We propose that incentives accentuate both the positive and the negative forces of exactingness. More specifically, when feedback is generally positive, as in lenient environ-

---

[1] Although we have chosen to operationalize differences in exactingness by manipulating α within the context of a squared error loss function, there are other ways to do this. Exactingness could, for example, be represented by different forms of asymmetric loss functions.

[2] Whether feedback is negative also depends on the constant of 500 in Equation 1. It is important to note, therefore, that by choice of appropriate constants it would be possible to design lenient environments in which actual feedback is predominantly negative or exacting environments in which actual feedback is predominantly positive. What matters, however, is whether feedback is perceived by subjects to be positive or negative.

[3] Whereas shifting strategies can result in lower performance in the short run, we note that this behavior may often be generally adaptive. For example, if a subject's initial hypothesis about the nature of the underlying system is incorrect, then shifting strategies in attempts to test alternative hypotheses is quite appropriate.

ments, incentives will induce more consistent application of apparently successful response strategies, and performance will improve (cf. Hammond & Summers, 1972). When feedback is generally negative, as in exacting environments, incentives will induce a more intensive search for alternatives, and performance will degrade, at least in the short run.

A Formal Model

To clarify implications of the above arguments, we use the heuristic device of a simple, algebraic model. Let

$$\pi = k\,[b\alpha^{\lambda} - c\alpha], \tag{5}$$

where k is a constant of proportionality, b and c are coefficients ($b, c > 0$) representing the extent to which the presence of incentives accentuates, respectively, the positive and negative aspects of exactingness (α) on performance (π), and λ ($0 < \lambda < 1$) determines the degree of concavity of the function that represents the positive aspect of exactingness (α) on performance.[4]

We draw two general implications from this model. First, the form of Equation 5 is such that performance will be a single-peaked (inverted-U-shaped) function of exactingness (α) (cf. Coombs & Avrunin, 1977). This means that performance will be better when exactingness (α) is at intermediate rather than at extreme values.

Second, we can enquire about how incentives interact with exactingness. To do so, assume that Equation 5 represents performance with no incentives and denote performance with incentives by

$$\pi' = k[b'\alpha^{\lambda} - c'\alpha], \tag{6}$$

where $b' > b$ and $c' > c$. Next, ask when performance with incentives exceeds that without incentives, that is, when $\pi' > \pi$. Simple algebraic manipulation leads to the condition

$$(b' - b)/(c' - c) > \alpha^{1-\lambda}. \tag{7}$$

The general implication of Equation 7 is that there is a critical value of exactingness (α) below which incentives lead to superior performance but above which incentives are dysfunctional.[5]

Predictions

The model implied by Equation 5 and its underlying assumptions lead to several predictions concerning observed performance (π):

1. Environments characterized by intermediate levels of exactingness (α) will lead to better performance (π) than will lenient or exacting environments. (This is implied by the fact that π is a single-peaked or inverted-U-shaped function of α.)

2. There will be an interaction between incentives and exactingness. Whereas incentives will lead to improved performance in lenient environments, they will become less beneficial as exactingness increases.

Our third prediction is intended as a test of the assumption that exactingness (α) has both positive and negative effects on performance.

3. If the negative effects of exactingness (α) on performance (π) are eliminated, performance (π) should increase as a function of exactingness (α).[6]

In addition to these predictions, our theoretical model suggests other observable implications concerning the process by which exactingness and incentives affect performance. Of particular interest are the effects of these variables on the consistency with which subjects execute response strategies. First, because subjects in exacting environments are more likely to observe negative feedback than those in lenient environments, their strategies should exhibit more inconsistency as they search for strategies that work better. Second, there should be an interaction between exactingness and incentives. When feedback is positive (lenient environments), subjects should exhibit less inconsistency in the presence of incentives because they will be more motivated to take care in executing successful strategies. On the other hand, when facing negative feedback (exacting environments), the search for better strategies should be intensified in the presence of incentives, thereby resulting in even greater inconsistency.

Experimental Evidence

We conducted five experiments to test the above predictions and related issues. In Experiment 1, we investigated three levels of exactingness both with and without incentives. This study used what we call a *sharp* incentive scheme in which subjects were rewarded if their mean evaluation points over a series of trials were positive, but not otherwise. This allowed us to test Predictions 1 and 2.

In Experiment 2, we did not use explicit monetary rewards. Instead, arguing that self-determined aspirations are incentives, we manipulated subjects' aspirations of performance. This allowed us to test Prediction 2.

In Experiment 3, we tested the limits of our theoretical scheme by using an incentive scheme that did not make clear

---

[4] Although our model could be developed using general functional forms, we have used specific functions to make the presentation more concrete. In particular, whereas we have modeled the negative aspect of exactingness in Equation 5 by a linear function, the implications of our model would also hold for nonlinear functions that are nonconcave decreasing functions of exactingness.

[5] Recalling that $(b' - b)$ and $(c' - c)$ represent, respectively, the increases in the extent to which incentives accentuate the positive and negative aspects of exactingness, it is instructive to examine the precise implications of our model when α is restricted to the range from 0 to 1. First, note that if $(b' - b) > (c' - c)$, incentives will always improve performance regardless of exactingness (α). In other words, if the effect of incentives is to accentuate the positive aspects of exactingness (α) more than the negative, incentives will always increase performance. Second, if $(b' - b) < (c' - c)$, there is a critical point of exactingness (α) on the 0 to 1 range below which incentives are functional but above which they are dysfunctional. In other words, even if incentives have more influence on the negative as opposed to the positive aspects of exactingness, they can aid performance provided that exactingness is low. As exactingness increases, however, incentives become less beneficial.

[6] If there is no negative aspect of exactingness, the $c\alpha$ term in Equation 5 is assumed not to exist.

distinctions between successful and unsuccessful performance.

Because we were intrigued by the question of whether different mental sets during learning could lead to different levels of subsequent performance, we also ran a condition in which subjects were explicitly instructed to learn how to make decisions in the task as opposed to maximizing evaluation points. In Experiment 4, we compared the performance of these subjects with those in Experiments 1 and 2 under conditions in which all subjects' performance was evaluated by the same sharp evaluation scheme and when all had experience in the task. Finally, Experiment 5 tested Prediction 3 by having subjects perform the experimental task under conditions designed to mitigate the negative aspects of exactingness.

Experiment 1

Method

Subjects. The subjects in this and in our other experiments were all recruited in the same manner through advertisements placed around the University of Chicago. They were offered between $5 and $15 for participating in an experiment on decision making. Their mean age was 22.4 years, and their mean educational level was 2.9 years beyond high school. One hundred and twenty-one subjects participated in this experiment.

Task. The task, which was individually administered by microcomputer, involved making a series of decisions. As described above, subjects were shown a value of a predictor variable, W, and then required to give a response, Q. Immediate feedback on each decision was provided by way of a score labeled *evaluation points.* Subjects were told that the object of the game was to maximize evaluation points. Evaluation points were linearly related (negatively) to the squared difference between their response, Q, and the unobservable criterion, D (see Equation 1), though this was not known to subjects.

More specifically, subjects were told that they were to set a value of a "DECISION VARIABLE that can vary between 1 and 1000." Moreover, "At the time you make this decision you will see the value of another variable called W. Your performance in each period of the game will be measured by a variable called EVALUATION POINTS." As part of the feedback, subjects were also told that they would "see the values of 2 other variables that could be useful to you in your decision making. These are called A and B." A and B were variables that could have provided limited but useful information for subjects who achieved a more advanced understanding of the way the underlying system worked.[7] Subjects were permitted to take notes and were also given the ability to scroll back the computer screen and examine data from past decisions.

The relation between the unobservable criterion, D, and W was subject to a small random disturbance so that the same evaluation points would not necessarily be observed if subjects repeated a response to the same W—see Equation 2. (The correlation between W and D was high, $r = .99$.) W was normally distributed with a mean of 70 and a standard deviation of 7.

Design and Procedure

Each subject was allocated at random to one of six groups created by crossing two levels of incentives (incentives vs. no incentives) by three types of environment (lenient, intermediate, and exacting) so that there were 20 subjects in each group. (One group had 21 subjects.) Subjects in the no-incentives condition were informed, "Your pay for this part of the experiment will not depend on how well you do in the game." In contrast, subjects in the incentives condition were told that their pay would depend on how well they performed. Specifically, their pay would depend on the mean evaluation points achieved over 30 trials with one cent for each point above zero. Thus remuneration could vary between $0.00 and $5.00. Feedback concerning mean evaluation points earned to date was continually updated and displayed on the screen of the microcomputer used for administering the task for all subjects. We specifically maintained this information on the screen so that subjects would be aware of how well they were doing and whether they were likely to be paid for participating in this part of the experiment (i.e., whether their mean score was above or below zero).

Differences in exactingness of the environment were manipulated by changing the constant of proportionality, α, in Equation 1. For the lenient environment, $\alpha = .01$; for the intermediate environment, $\alpha = .05$; and for the exacting environment, $\alpha = .50$. We chose these values after observing outcomes associated with simulated strategies that differed in accuracy as measured by π (see Equation 4).

At the outset of the experiment, subjects were told that they would make 30 decisions. This was Round 1. After completing this task, they were first asked to rank themselves in percentile terms regarding how well they thought they had performed in the task relative to other University of Chicago students. They were then told that they were to play a second series of 30 trials under exactly the same conditions. This was Round 2. Next, subjects were asked to complete a questionnaire that quizzed them about their understanding of the model underlying the task (i.e., relations between variables, and so on). They were then asked to complete a further series of 30 trials for Round 3. For this round, however, subjects who had previously been in the no-incentives condition were required to make their decisions under the same incentives conditions as the other subjects. The question on self-ranking of performance was also repeated after Rounds 2 and 3. In this experiment we only consider responses for Rounds 1 and 2. Round 3 responses form part of the data for Experiment 4.

In short, the design of Experiment 1 involved two between-subject variables, one with two levels (incentives vs. no incentives), and the other with three (lenient, intermediate, and exacting environments). There were two rounds each involving 30 trials, and subjects completed a questionnaire about their understanding of the task after the second round.

Results

As discussed above, the differential effects of positive and negative feedback are an important element of our theoretical model. A check on whether subjects did observe mainly positive, mixed, or negative feedback in the different environ-

[7] A and B were defined as follows (see Equation 1). If $Q < D$, then $A = Q$ and $B = 0$; if $Q \geq D$, then $A = D$ and $B = Q - D$.

ments, is provided by the proportions of subjects in each condition whose mean evaluation points were greater than zero in Round 2 (when subjects were more experienced in the task). These were .68, .46, and .05 for the lenient, intermediate, and exacting environments, respectively.

Performance (π). Table 1 and Figures 1 and 2 provide overviews of the results. For all six experimental conditions, the top two sections of Table 1 report means and standard deviations by rounds regarding performance (π) and evaluation points (i.e., the penalty functions actually experienced by the subjects).

Figure 1 shows performance (π) achieved by subjects in the three different environments for both rounds. The upward sloping lines indicate that performance (π) improved across rounds, that is, learning occurred. Subjects in the intermediate condition outperformed those in both the lenient and the exacting environments—Prediction 1.

Figure 2 displays the overall mean performance (π) across both rounds for each of the six experimental conditions and, in particular, the effects of incentives. There is clear evidence of an Incentives × Exactingness interaction—Prediction 2. Subjects in the lenient–incentives condition outperformed those in the lenient–no-incentives condition, mean of 359 versus 289; and subjects in the exacting–no-incentives condition outperformed those in the exacting–incentives condition, mean of 319 versus 301. Finally, there was essentially no difference between the mean scores of the incentives and the no-incentives groups in the intermediate environment, 351 versus 356.

As for formal tests of our predictions using performance (π) as the dependent variable, a repeated measures analysis of variance (ANOVA) (with round as the repeated measure) showed main effects for round, $F(1, 115) = 142.8$, $MS_e = 1{,}748$, $p < .001$, and exactingness, $F(2, 115) = 3.56$, $MS_e = 11{,}209$, $p = .03$, as well as a significant Incentives × Exactingness interaction, $F(2, 115) = 4.00$, $MS_e = 11{,}209$, $p = .02$. Round did not interact with any of the other variables.

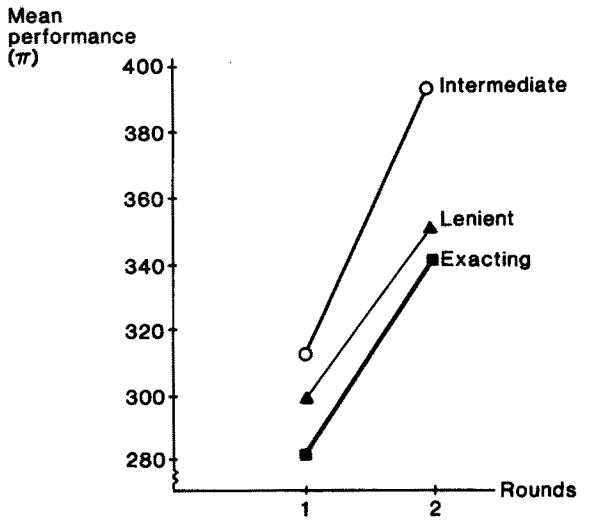

Figure 1. Experiment 1: Mean performance (π) by types of environment (lenient, intermediate, and exacting) across rounds.

Looking at the results in terms of planned comparisons, mean performance (π) of the intermediate condition in Round 1, 313, was not significantly greater than that of the lenient and exacting conditions, 297 and 281, respectively. However, the Round 2 performance (π) measures of 394 for intermediate versus 350 for lenient, and 339 for exacting, were significant, $t(79) = 2.31$, $p = .02$, for the former, and

Table 1
Experiment 1: Performance (π), Evaluation Points, and Inconsistency in Lenient, Intermediate, and Exacting Environments

Round	Incentives			No incentives		
	Lenient	Intermediate	Exacting	Lenient	Intermediate	Exacting
			Performance (π)			
1						
M	331	309	274	263	317	287
SD	73	89	77	77	67	71
2						
M	386	393	328	314	394	351
SD	74	98	93	84	66	87
			Evaluation points			
1						
M	−10	−2,812	−40,569	−390	−2,455	−36,809
SD	362	1,977	18,853	451	1,747	18,907
2						
M	230	−1,098	−27,065	−80	−787	−21,537
SD	267	2,137	21,719	427	1,340	20,393
			Inconsistency (σ_z)			
1						
M	167	209	229	229	206	222
2						
M	74	92	150	140	99	137

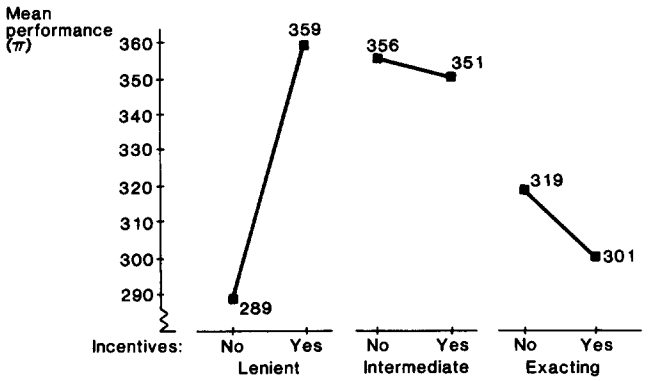

Figure 2. Experiment 1: Mean performance (π) for Rounds 1 and 2 by experimental conditions.

$t(79) = 2.82$, $p = .006$, for the latter. In addition, separate ANOVAs on the data in the lenient and exacting conditions showed that the predicted Incentives × Exactingness interactions were significant both in Round 1, $F(1, 76) = 5.81$, $MS_e = 5{,}574$, $p = .018$, and Round 2, $F(1, 76) = 6.29$, $MS_e = 7{,}202$, $p = .014$.

Further insight into these interactions can be gleaned by contrasts between the two incentives conditions within each of the three types of environment. Within the lenient environment, subjects in the incentives condition had superior performance (π) to the others in both Rounds 1 and 2, with respectively, means of 331 versus 263, $t(38) = 2.84$, $p = .007$, and 386 versus 314, $t(38) = 2.90$, $p = .006$. The differences between the mean performance (π) of incentives and no-incentives subjects in both the intermediate and exacting conditions were not statistically significant in either Round 1 or Round 2.

Parenthetically, we note that subjects in all experimental conditions were unbiased in that the average error of their decisions was not significantly different from zero in any of the rounds. This suggests that subjects responded appropriately to the nature of the symmetric penalty functions in their feedback.

Inconsistency of response strategies. Underlying our theoretical model is the notion that, after negative feedback, subjects shift response strategies, whereas after positive feedback, they persist with and refine the same strategies. Moreover, the impact of incentives is to accentuate these effects. Before examining this, we first define inconsistency in operational terms.

Imagine that subjects' response strategies can be modeled by the regression of their responses, Q, on W, that is,

$$Q = \beta'_0 + \beta'_1 W + z. \quad (8)$$

A measure of the inconsistency of response strategy implied by this model is σ_z, that is, the conditional standard deviation of Q given W. This measures the variance in subjects' responses that is not systematically related to W and can be thought of as indicating the extent to which subjects varied their response strategies.[8] We note, however, that observed inconsistency in strategies can have two sources. One is inconsistent use of valid strategies as a result of, for example, lack of attention. The second is the result of experimentation or deliberately trying out alternative strategies. Mean values of estimates of σ_z are reported at the foot of Table 1 by experimental condition.

Because the level of perceived negative feedback increases with exactingness (see above), strategies would be expected to be more inconsistent as exactingness increases. Analysis of variance on measures of σ_z for both rounds of Experiment 1 reveals a significant effect for exactingness in Round 2, $F(2, 115) = 3.15$, $MS_e = 7{,}894$, $p = .047$ but not for Round 1. For Round 2, mean estimates of inconsistency are 107, 96, and 144 for the lenient, intermediate, and exacting environments, respectively. Thus, inconsistency in the exacting environment ($\alpha = .50$) is greater than in the lenient ($\alpha = .01$) and intermediate ($\alpha = .05$) conditions.

It is important to note, however, that this pattern does not match performance (π), which, as reported above, was superior in the intermediate environment. Indeed, that performance in the intermediate condition is greater than in the lenient and, yet, inconsistency does not differ significantly between conditions, accords well with the assumption that exactingness exerts both positive and negative effects on performance (see Equation 5). In particular, even though subjects in the intermediate condition are just as inconsistent as subjects in the lenient condition, they are able to learn better strategies.

As for incentives, the pattern of inconsistency would be expected to mirror that of performance, that is, there should be an interaction between incentives and exactingness. This follows from the assumption that incentives accentuate the effects of exactingness by both reinforcing the consistent use of successful strategies (i.e., causing subjects to take more care in execution) and intensifying changes in strategy following the observation of negative feedback. To test this, we performed ANOVAs on both the 3 × 2 design of the experiment (i.e., 3 levels of Exactingness × 2 levels of Incentives) and a 2 × 2 design that omitted the intermediate level of exactingness. For the full design, a repeated measures analysis shows a main effect for round, $F(1, 115) = 217.4$, $MS_e = 2{,}493$, $p < .001$, indicating a significant decrease in inconsistency across rounds (see Table 1), but the predicted Incentives × Exactingness interaction is only marginally significant $F(2, 115) = 2.69$, $MS_e = 11{,}947$, $p = .07$. For the reduced design, however, in addition to the effect for round, $F(1, 76) = 108.4$, $MS_e = 2{,}744$, $p < .001$, the predicted interaction is significant, $F(1, 76) = 4.44$, $MS_e = 12{,}505$, $p = .038$.

Further results. Other sources of data shed light on the processes that gave rise to the observed effects. One datum

[8] The R^2 from the regression implied by Equation 8 is typically used to measure the consistency of response strategies within the lens model paradigm (Hammond & Summers, 1972) and is, of course, a function of σ_z (within each individual regression). However, we do not use R^2 here because (a) Conventional lens model analysis does not apply to single-cue situations, (b) the variance in Q is relevant to our theoretical analyses and differs across experimental conditions, and (c) W and D (the criterion) are so highly correlated that consistency as measured by R^2 would not provide an independent measure of the contribution of consistency to performance.

collected by the microcomputer was time (in minutes) taken by subjects to complete each round. These averaged 22.4 and 14.6 for Rounds 1 and 2, respectively. A repeated measures ANOVA (with round as the repeated measure) showed a significant main effect for round, $F(1, 115) = 96.7$, $MS_e = 38.5$, $p < .001$, but no significant effects for either incentives or environment and no Incentive × Environment interaction. In addition, round did not interact with the other variables. On the other hand, performance (π) was correlated at the individual level with time spent on the task, $r = .21$, $p < .05$, and .29, $p < .01$, for Rounds 1 and 2, respectively. We therefore reanalyzed performance (π) with time as a covariate using a repeated measures analysis (with round as the repeated measure and on the complete 3 × 2 design). Once again, main effects were observed for round, $F(1, 114) = 87.7$, $MS_e = 1{,}752$, $p < .001$, and for exactingness, $F(2, 114) = 3.30$, $MS_e = 10{,}545$, $p = .04$. In addition, the predicted Incentives × Exactingness interaction was significant, $F(2, 114) = 3.72$, $MS_e = 10{,}545$, $p = .03$.

Recall that at the end of each round, subjects were asked to rank their performance in percentile terms vis-à-vis other University of Chicago students. Overall, the mean rankings were at the 45.1 and 54.9 percentiles for Rounds 1 and 2, respectively. A repeated measures ANOVA showed the difference between Rounds 1 and 2 to be significant, $F(1, 115) = 49.8$, $MS_e = 114.8$, $p < .001$, as well as a Round × Exactingness interaction, $F(2, 115) = 5.59$, $MS_e = 114.8$, $p = .005$. (Mean rankings increased more between rounds for subjects in the intermediate as opposed to other conditions.) At the individual level, it is of interest to note that whereas there was essentially no relation between self-assessed rank and feedback (i.e., evaluation points) for Round 1, $r = .10$, *ns*, this was not the case for Round 2, where the analogous correlation was .42, $p < .001$. Experience with the task apparently helped subjects assess their own performance more accurately in relative terms.

The questionnaire completed after Round 2 contained two kinds of questions. The first were direct questions concerning which variables subjects deemed most important as well as whether they thought that "the outcomes of the game (i.e., evaluation points) are determined according to some systematic set of rules." With regard to the latter, there was an interesting effect for exactingness. Subjects in the intermediate condition (who performed best) rated outcomes as being determined by a more systematic set of rules than subjects in the other conditions, mean of 5.41 on a 7-point scale versus 4.40 for lenient and 3.89 for exacting, $F(2, 113) = 6.96$, $MS_e = 3.49$, $p = .001$. The contrast of intermediate versus lenient was significant, $t(79) = 2.58$, $p = .01$, as was the contrast of intermediate versus exacting, $t(77) = 3.66$, $p < .001$.

Subjects were also asked to write "How does the game work?" by specifying the roles played by the different variables and their interrelations, and while imagining having "to explain to an agent how to play the game in your behalf," to give "a simple description of the system to convey a general sense of how it works" as well as "any specific tips you might have to achieve high evaluation points." The answers to these questions were graded like an examination using a preestablished checklist of criteria. Of particular interest was whether subjects articulated both the sign and slope of the critical relation between W and the decision variable. Each subject's questionnaire was scored on 4-point scales for both variables. We also gave each subject a total understanding score which, in addition to the scores for sign and slope, took into account their understanding that there were two types of error (i.e., setting the decision variable too high as well as too low), recognizing an identity between the decision variable and the sum of A and B (the two secondary feedback variables, see above), and whether they gave any valid tips to an "agent." The total score was calculated by summing the scores of the components (Einhorn & Hogarth, 1975). The ratings of the questionnaires were made independently by two of the authors and their judgments were averaged. As an indication of reliability, the correlation between the scores of the two judges on the total index was .85, $p < .001$.

Statistical analyses of the various indices revealed effects due to exactingness whereby subjects in the intermediate condition showed superior understanding to those in the lenient and exacting environments in a manner paralleling Prediction 1. The appropriate means for the slope index were 0.91 for intermediate versus 0.53 and 0.26 for the lenient and exacting, respectively, $F(2, 115) = 5.46$, $MS_e = 0.79$, $p = .005$; the analogous figures for the sign index were 1.40 versus 0.95 and 0.78, *ns*; for appreciation of errors, 0.88 versus 0.22 and 0.28, $F(2, 115) = 11.72$, $MS_e = 0.47$, $p < .001$; and for the total index, 4.03 versus 2.49 and 2.22, $F(2, 115) = 6.42$, $MS_e = 6.04$, $p = .002$.

The indices were also related to performance (π) at the individual subject level. Across all subjects, the correlations between performance (π) and the sign index in Rounds 1 and 2 were .38, and .51, respectively, $p < .001$ for both. The corresponding figures for the slope index were .40, and .55, $p < .001$ for both. In addition, the correlations between performance across rounds and the index of total knowledge were .47, and .64, $p < .001$ for both.

Finally, we found no significant differences when we analyzed results by demographic variables (e.g., age, gender, mathematics, and science background vs. nonmathematics and science background).

Discussion

In short, the results of Experiment 1 validate two of our theoretical predictions. First, performance was seen to have an inverted-U-shaped relation with exactingness, that is, performance was better in the intermediate as opposed to the lenient or the exacting environments (Prediction 1).

Second, incentives interacted with exactingness in their effects on performance (Prediction 2). In lenient environments, incentives improved performance; however, incentives were ineffective in intermediate and exacting environments. Indeed, the data suggest deleterious effects of incentives in exacting environments.

In addition to performance, other data supported the main results. First, a measure of inconsistency in strategy use showed that this varied with exactingness. In addition, inconsistency mirrored the Incentives × Exactingness interaction.

Second, subjects in the intermediate condition reported finding the task as having been generated by a more systematic set of rules. Their written analyses also showed greater understanding of the task. Related to this were the strong correlations between measures of understanding and performance at the individual level. These are important findings and speak to the issue of whether and when verbalizable knowledge and performance are related. In some studies, Broadbent and his colleagues have found no relation between the ability to verbalize understanding of relations between variables learned through taking decisions and performance (Berry & Broadbent, 1984; Broadbent, 1977; Broadbent & Aston, 1978). However, more recent work suggests that when the relation between decision and action is salient (as in the case of simple tasks), performance and verbalizable knowledge could well be related (Berry & Broadbent, 1988; Broadbent, FitzGerald, & Broadbent, 1986).

Finally, this study employed a sharp incentive scheme whereby subjects in the incentives condition were only remunerated if they had positive mean scores for evaluation points. Thus, it is tempting to argue that the lack of a positive effect for incentives in the exacting environment was due to subjects giving up when they realized that they had little chance of achieving a positive score. (In the exacting condition, discerning subjects who had made bad errors could realize that it would be impossible for them to break even by the end of the round.) Two arguments mitigate against this explanation. First, as reported above, we found no significant time differences by experimental conditions. Second, we reanalyzed the data to check for trends in learning by experimental conditions, both across the 30 trials and blocks of 5 trials within rounds, paying particular attention to behavior toward the end of each round (where giving up would be most likely to occur). We found no evidence of giving up.

Experiment 2

Rationale. As noted in our review of the literature, incentives can take many forms. In this experiment we consider the incentives implicit in targets or aspirations that people set for themselves. Our argument is the following: Aspirations are a form of incentives and therefore, like monetary payments, should accentuate both the positive and the negative aspects of exactingness. Thus, in terms of the model in Equation 5, high aspirations imply larger b and c coefficients than low aspirations. This, in turn, implies an interaction between aspirations and exactingness: In lenient environments, high aspirations should lead to better performance than low aspirations; as exactingness increases, however, this difference should reverse. In other words, the use of aspirations permits a further test of Prediction 2.

Method

Subjects. There were 80 subjects recruited in the same manner and from the same population as Experiment 1.

Task. The task was identical to the no-incentives condition used in Experiment 1 except that we modified the instructions and the information display to manipulate aspirations.

Design and procedure. Each subject was allocated at random to one of four groups created by crossing two levels of aspirations (high vs. low) by two types of environment (lenient vs. exacting) so that there were 20 subjects in each group. (We did not include an intermediate exactingness condition in this experiment.) Assuming that subjects like to compare their performance with that of similar others, we manipulated aspirations by informing subjects of the median average score achieved by other University of Chicago students who had participated in the experiment before them. This figure was provided in the instructions (on the microcomputer screen) before subjects began Rounds 1 and 2 and remained on the screen throughout each round. In fact, the actual median was not used. Instead, subjects in the low-aspirations condition were given the score corresponding to the 0.10 fractile of the empirical distribution of scores in the appropriate lenient or exacting conditions of Experiment 1. For subjects in the high-aspirations condition, it was the score corresponding to the 0.90 fractile. Consistent with the data from Experiment 1, these medians showed improvement for Round 2 over Round 1.

As in Experiment 1, there were two rounds consisting of 30 decisions each. After each round, subjects were asked to rank themselves in percentile terms regarding how well they thought they had performed the task relative to other University of Chicago students. After Round 2, subjects completed the same questionnaire used in Experiment 1. After this, subjects completed a third round that differed from the others in that no aspirations were provided, and all subjects played under an incentives condition in exactly the same manner as Round 3 of Experiment 1. The data from this third round are reported below as part of Experiment 4.

Because of the deception involved in the experimental manipulation, subjects were contacted individually after the experiment was completed and were thoroughly debriefed. This involved providing full details concerning the underlying rationale for the study and the nature of the deception. Subjects were also offered a written summary of the study and results.

Results

Because aspirations were manipulated by providing the median performance of other University of Chicago students, subjects' rankings of their own performance relative to the same population provide a check on the effectiveness of the manipulation. Specifically, subjects in the high-aspirations condition would be expected to assess their relative rank lower than those in the low-aspirations condition. This was the case. Mean percentile ranks, for high versus low, were 34 versus 64, and 41 versus 69 for Rounds 1 and 2, respectively. These differences were statistically significant, $F(1, 76) = 69.7$, $MS_e = 267$, $p < .001$ for Round 1, and $F(1, 76) = 41.2$, $MS_e = 423$, $p < .001$ for Round 2.

Performance (π). Table 2 reports means and standard deviations by rounds for performance (π) and evaluation points.

A repeated measures ANOVA on performance (π) (with round as the repeated measure) reveals a single main effect for round (thereby indicating learning), $F(1, 76) = 96.0$, $MS_e = 1{,}419$, $p < .001$, an Incentives × Round interaction, $F(1, 76) = 8.77$, $MS_e = 1{,}419$, $p = .004$, and a marginally significant Aspiration Level × Exactingness interaction, $F(1, 76) = 3.46$, $MS_e = 10{,}947$, $p = .067$. These two interactions are illuminated by separate ANOVAs for each round that show a significant Aspiration Level × Exactingness interaction for Round 1, $F(1, 76) = 4.02$, $MS_e = 6{,}924$, $p = .049$, but not for Round 2. Moreover, the nature of the interaction supports Prediction 2, that is, in the lenient environment, high aspira-

Table 2
Experiment 2: Performance (π) and Evaluation Points Among Subjects With High and Low Aspirations in Lenient and Exacting Environments

Round	High		Low	
	Lenient	Exacting	Lenient	Exacting
	Performance (π)			
1				
M	299	250	280	298
SD	64	67	73	87
2				
M	364	337	315	345
SD	60	93	84	91
	Evaluation points			
1				
M	−180	−47,196	−247	−35,148
SD	339	18,249	394	22,589
2				
M	152	−26,390	−55	−22,326
SD	245	21,192	484	19,023

tion subjects outperform those with low aspirations; in the exacting environment, it is the reverse.

Inconsistency of response strategies. There were neither significant main effects nor interactions involving the measure of inconsistency of response strategies.

Further results. Time spent on the task was similar to that in Experiment 1 with means of 21.0 and 15.0 min for Rounds 1 and 2, respectively. Across experimental conditions, the only statistically significant difference on time was between rounds, $F(1, 76) = 48.3$, $MS_e = 29.7$, $p < .001$. As in Experiment 1, time spent on the task was significantly correlated at the individual level with performance (π) for both Round 1, $r = .30$, $p < .01$, and Round 2, $r = .38$, $p = .001$.

With respect to the questionnaire administered after Round 2, there were no significant main effects or interactions for the sign and slope indices. (Recall that unlike Experiment 1 there was no intermediate exactingness condition.) However, the indices did correlate significantly with performance (π) at the individual level. For sign, the correlations for Rounds 1 and 2 were, respectively, .51, and .62; for slope, .48 and .54; and for the index of total knowledge, .58 and .71, $p < .001$ for all.

Discussion

These data provide further support for Prediction 2 in that there was a significant Incentives × Exactingness interaction when incentives were operationalized by aspirations as opposed to cash rewards. In addition, there were significant correlations between the ability to articulate understanding of the task and performance.

The main result, however, is weaker than that of Experiment 1 in that although the predicted Incentives × Exactingness interaction was significant in Round 1, it was not significant in Round 2. A possible reason for the weaker effect is that subjects might have been uncertain as to whether feedback should be interpreted as positive or negative. On the one hand, one can imagine subjects coding outcomes as successes or failures relative to the aspiration levels implicit in the task instructions. On the other hand, subjects could also have been responding to whether outcomes yielded positive or negative scores for evaluation points. For example, whereas 29 (of 40) and 2 (of 40) subjects achieved positive mean evaluation point scores for Round 2 in the lenient and exacting environments, respectively, only 1 (out of 20) subjects in the lenient–high-aspirations condition achieved a score above the aspiration level, and 18 (out of 20) achieved a score above the aspiration level in the exacting–low-aspirations condition. This weaker effect of the Incentives × Exactingness interaction, and the possible confusion in interpreting feedback as positive or negative, was further evidenced by the fact that there were no significant effects for inconsistency in strategy use.

Our manipulation of aspirations inevitably leads to comparisons with work on goal setting. One of the most consistent results from this literature is the comparison between situations where people are given high, explicit goals as opposed to being told to do their best. High, explicit goals, it is claimed, lead to better performance than more vague "do your best" goals (Locke, Shaw, Saari, & Latham, 1981). In a recent study, Earley, Connolly, and Ekegren (1989) demonstrated limits to this empirical regularity. Specifically, in a prediction task subjects who were given high, explicit goals in terms of stringent conditions on allowable prediction error performed worse than subjects who were given "do your best" goals. If we equate Earley et al.'s stringent conditions with our exacting environment, the results are consistent and support the notion that when subjects fail to reach high goals, subsequent changes in strategy can lead to lower performance. Earley et al. further point out that in the tasks on which the empirical regularity summarized by Locke et al. is based, performance is typically measured by quantity of outputs (e.g., number of judgments) as opposed to quality (e.g., mean predictive accuracy). Whereas a direct relation may be expected between effort and quantity of outputs, it is not clear that effort and quality would be related in the same way.

In fact, by combining experimental conditions from Experiments 1 and 2, it is possible to test directly the differential effects of high, explicit versus "do your best" goals. Specifically, consider the comparison between subjects in the high-aspirations group of Experiment 2 and the no-incentives condition of Experiment 1. Using a repeated measures ANOVA (with round as the repeated measure), there was a significant Instructions (no incentives vs. high aspirations) × Exactingness interaction, $F(1, 76) = 4.79$, $MS_e = 9{,}802$, $p = .03$. In the lenient environment, high aspirations led to better performance in accordance with the goal-setting literature, mean of 332 versus 289; in the exacting environment, however, the relation was reversed, mean of 294 versus 319. In addition, although there was a main effect for round (indicating improvement from Round 1 to Round 2), $F(1, 76) = 97.1$, $MS_e = 1{,}832$, $p < .001$, round did not interact with the other variables. Finally, in terms of contrasts within conditions of exactingness, the only statistically significant main effect was within the lenient environment, where the high-aspirations group outperformed the no-incentives subjects in Round 2, mean performance (π) of 364 versus 314, $t(38) = 2.17$, $p = .037$.

Experiment 3

Rationale. The incentive scheme used in Experiment 1 contained an important discontinuity in that subjects were only remunerated if their scores on mean evaluation points were positive. In Experiment 2, where incentives took the form of induced aspirations, effects were weaker. Moreover, we speculated that this might have been due to ambiguity concerning the positive or negative nature of feedback. Taken together, these results (as well as our theoretical framework) suggest limitations on the types of incentives that are likely to have an impact. Specifically, if feedback from incentive schemes does not clearly distinguish between positive and negative outcomes, incentives are far less likely to affect performance in the form of an Incentives × Environment interaction (Prediction 2).[9] Experiment 3 was designed to test this issue.

Method

Subjects. There were 80 subjects recruited in the same manner and from the same population as Experiments 1 and 2.

Task. The task was identical to that used in Experiment 1 except that we modified the payment scheme in the incentives condition. On the basis of the distribution of mean evaluation scores in Experiment 1, we created a function linking mean evaluation points to remuneration so that even though the function always indicated greater rewards for better performance, poor performance would always receive some reward. To make the reward associated with a perfect score the same as in Experiment 1, we restricted payment from \$0 to \$5. The range of evaluation points remunerated was from −58,900 to +500, and the function was nonlinear in that it was flatter for low scores (e.g., a 100-point improvement at the high end of the evaluation points scale brought greater incremental rewards than at the low end). So that subjects would clearly understand the relation between performance and payment, each was provided with a chart that showed how remuneration varied with mean evaluation points.

Design and procedure. Each subject was allocated at random to one of four groups created by crossing two levels of incentives (incentives vs. no incentives) by two types of environment (lenient vs. exacting) so that there were 20 subjects in each group. Subjects completed two rounds of 30 trials and completed a questionnaire after Round 2. (As in the previous experiments, subjects also completed a third round of 30 trials in which all subjects were in the same incentives condition.)

Results

Performance. There were no statistically significant effects in terms of performance (π). Averaged across both rounds, mean scores (π) in the lenient environment were 325 and 316 for the incentives and no-incentives conditions, respectively. For the exacting environment, the corresponding figures were 316 and 324. A repeated measures ANOVA (with round at the repeated measure) showed main effects only for round, mean of 289 for Round 1 and 352 for Round 2, $F(1, 77) = 102.13$, $MS_e = 1{,}543$, $p < .001$, and no significant interactions.

Inconsistency of response strategies. The data revealed a main effect within both rounds whereby subjects in the lenient environment exhibited less inconsistency, mean of 188 versus 222 for Round 1, $F(1, 77) = 4.60$, $MS_e = 5{,}152$, $p = .039$; and 95 versus 138 for Round 2, $F(1, 77) = 5.17$, $MS_e = 7{,}235$, $p = .026$.

Further results. Despite the lack of effects in performance (π) between experimental conditions, subjects' experiences with the task differed. First, subjects in the incentives condition spent more time, means of 28.7 min versus 21.1 min in Round 1, $F(1, 77) = 6.26$, $MS_e = 186.8$, $p = .015$; and 16.9 min versus 13.4 min in Round 2, $F(1, 77) = 4.33$, $MS_e = 55.5$, $p = .041$. However, using time as a covariate did not alter the results for performance (π). Second, subjects in the lenient condition ranked their own performance higher after both Round 1 and Round 2, means of 53.8 versus 41.2, $F(1, 77) = 7.78$, $MS_e = 392$, $p = .007$, and 64.9 versus 47.6, $F(1, 77) = 13.99$, $MS_e = 430$, $p < .001$.

On the basis of the questionnaire administered after Round 2, subjects in the lenient condition also rated outcomes as being determined by a more systematic set of rules with a mean of 5.16 versus 4.18 on a 7-point scale, $F(1, 77) = 6.27$, $MS_e = 3.13$, $p = .014$. Relative to the questionnaire administered in Experiment 1, this questionnaire contained some additional questions (also on 7-point scales) about the process. Mean scores on this questionnaire revealed that, relative to the exacting environment, subjects in the lenient condition found the task more enjoyable, 3.95 versus 3.08, $F(1, 77) = 7.59$, $MS_e = 2.05$, $p = .007$; less frustrating, 4.05 versus 5.35, $F(1, 77) = 13.7$, $MS_e = 2.50$, $p = .004$; less challenging, 4.77 versus 5.80, $F(1, 77) = 13.6$, $MS_e = 1.58$, $p < .001$; and less discouraging, 3.40 versus 5.25, $F(1, 77) = 27.7$, $MS_e = 2.51$, $p < .001$. They also stated that they put less effort into the task, 4.17 versus 4.96, $F(1, 53) = 5.74$, $MS_e = 1.52$, $p = .020$, and fewer reported thinking about giving up, 25% versus 60%, $F(1, 77) = 11.5$, $MS_e = .022$, $p = .001$. However, there was no difference between the proportions of subjects reporting having given up.

Analyses of subjects' descriptions of how the game worked revealed only one statistically significant result. This was an interaction between incentives and environment concerning subjects' appreciation that there were two types of error in the task. Within the lenient environment, this was better understood by subjects in the incentives condition, mean of .80 versus .50, whereas the relation reversed in the exacting environment, .24 versus .65, $F(1, 77) = 2.56$, $MS_e = 0.60$, $p = .040$.

Once again, indices of understanding were correlated with performance (π) at the individual level. Across all subjects, the correlations between performance (π) and the sign index in Rounds 1 and 2 were, respectively, .22, $p = .05$, and .42, $p < .001$. The corresponding figures for the slope index were .21, *ns*, and .38, $p = .001$. In addition, the correlations between performance across rounds and the index of total knowledge were .38, $p = .001$, and .53, $p < .001$.

Discussion

Experiment 3 was similar to Experiment 1 except that (a) there was no intermediate exactingness condition, and (b) we

[9] In terms of Equation 7, the fact that feedback from incentives schemes does not distinguish clearly between positive and negative outcomes can be modeled by setting $b = b'$ and $c = c'$, that is, incentives have no differential effect on the positive and negative aspects of exactingness.

changed the nature of the incentives scheme so that there was no longer a discontinuity between performance that was rewarded and performance that was not rewarded. Underlying this manipulation was the hypothesis that incentives are more likely to have an effect on performance (π) if they clearly discriminate between positive and negative feedback, a hypothesis that is consistent with these results.

Despite the lack of an Incentives × Environment interaction on performance (π) with this incentive scheme, supplementary evidence indicated that subjects in the different experimental conditions did not have the same experience of the task. Under incentives, subjects took longer to perform the task. In the lenient environment, subjects' use of strategies showed less inconsistency, and they generally found the task less aversive than in the exacting environment.

Experiment 4

Rationale

There are two major, practical questions for the present line of research. First, what combinations of incentives and exactingness are most effective for learning? Second, once people have acquired a certain level of expertise, does performance reflect carryover effects from conditions experienced during learning (cf. Schwartz, 1982)? Recall that an important conceptual consideration underlying our work is the notion that outcome feedback confounds two kinds of information. One is information concerning the structure of the underlying system in which decisions are being made (i.e., how variables are related). The second relates to how well the decision maker is performing the task. Given this ambiguity, it is reasonable to assume that concentrating attention only on the structure of the task during learning should improve subsequent performance. Experiment 4 was designed to test this hypothesis. During two 30-trial rounds subjects were instructed to concentrate only on learning the task used in Experiments 1 and 2 (instead of maximizing evaluation points). They were then switched to an incentives condition in a third round, and their performance was compared with the Round 3 incentives performance of the subjects in Experiments 1 and 2 who had learned the task under different conditions.

Method

Subjects. There were 201 subjects, 80 of whom had participated in Experiment 1 and 80 of whom had participated in Experiment 2. The 41 subjects who participated only in this experiment came from the same population as the other experiments and were recruited in the same manner.

Task. For the 41 subjects participating uniquely in Experiment 4, the task was the same as Experiment 1 with two exceptions. First, instructions differed in that subjects were told, "The object of this game is to maximize EVALUATION POINTS. However, in playing the game you should not be concerned with how well you do. Instead, your objective is to learn how the game works." In addition, half of the subjects were specifically told to expect to be asked how the game worked and to make their understanding explicit. Second, after Rounds 1 and 2 subjects did not rank their own performance. Instead they ranked how well they thought that they had understood the task relative to other University of Chicago students.

Design and procedure. The design of the study involved two between-subjects variables. These were prior learning conditions (with six levels) and exactingness of the environment (with two levels, lenient and exacting). Two levels of the prior learning conditions involved the 41 Experiment 4-only subjects, whom we shall refer to as the inference group and who were allocated at random to four subgroups. These were level of instructions (explicitly told to expect to have to explain their understanding of the game vs. not explicitly told) and exactingness of the task environment (lenient vs. exacting, using the same parameters as Experiments 1 and 2). Thus, inference group subjects had two rounds of 30 trials in which their task was to discover how the system worked; they then completed the same questionnaire used in Experiments 1 and 2 before being switched in Round 3 to the same incentives condition experienced by subjects in those experiments. Apart from the differences in the task noted above, procedures for these subjects were exactly the same as in Experiment 1.

Two other levels of prior learning conditions were for incentives and no incentives and involved the 80 Experiment 1 subjects who had been exposed to the lenient and exacting environments. The final two levels of prior learning were the high- and low-aspirations conditions from Experiment 2.[10]

Results

We first note that there were no significant main effects or interactions involving the difference in the levels of the instructions given to inference group subjects concerning whether they would be asked later on to explain their understanding of the game. We therefore ignore this experimental manipulation and analyze the results as arising from a 5 × 2 design (i.e., 5 prior Learning Conditions × 2 Levels of Exactingness).

Before discussing the results of performance in Round 3, it is important to note that subjects in the inference group took, on average, 54% longer than the others to complete the experimental tasks, $t(199) = 5.64$, $p < .001$. Mean times were 33.2 min versus 21.5 min in Round 1, 23.6 versus 14.2 in Round 2, and 18.0 versus 13.0 in Round 3. This result is particularly interesting because subjects in all experiments were given the same expectations concerning remuneration for participation and had identical incentives in Round 3. In addition, neither group was told how much time to spend on the experimental tasks. Apparently giving subjects a set to learn induced a more careful approach (evidenced by time spent) that also carried over to the incentives condition in Round 3.

The first panel of Table 3 summarizes data on mean accuracy scores for Round 3 in which all subjects were in the same incentives condition. The other panels report mean indices of understanding in respect of the sign and slope of the important predictive relation determining outcomes. These means are based on the questionnaire completed at the end of Round 2.

For performance (π), Table 3 shows little difference for the effect of exactingness of the environment, 366 versus 372. However, differences due to prior experimental treatments

[10] We do not include subjects from Experiment 3 in these comparisons because these subjects faced a different incentive scheme in Round 3.

Table 3
Selected Results From Experiment 4

Outcome	Prior condition					
				Aspirations		
	Inference	Incentives	No incentives	High	Low	*M*
Mean performance (π) for Round 3						
Lenient environment	389	399	347	384	313	366
Exacting environment	405	372	372	363	350	372
M	397	385	359	374	332	—
Indices of understanding						
Sign						
Lenient environment	1.74	1.43	0.48	0.93	0.80	1.08
Exacting environment	1.33	0.90	0.65	0.76	1.20	0.97
M	1.54	1.17	0.57	0.35	1.00	—
Slope						
Lenient environment	0.93	0.83	0.23	0.33	0.18	0.50
Exacting environment	0.80	0.23	0.30	0.34	0.88	0.51
M	0.87	0.53	0.27	0.34	0.53	—

are large. Overall, the inference condition has the highest mean score, 397, compared with the poor showing of the low-aspirations group, 332. A 5 × 2 ANOVA only reveals a significant main effect for prior learning condition, $F(4, 191) = 3.37$, $MS_e = 7{,}680$, $p = .011$, and no significant Prior Learning Condition × Exactingness interaction.

Contrasts between prior learning conditions permit more refined analyses. The inference group outperformed the no-incentives subgroup, 397 versus 359, $t(79) = 1.96$, $p = .053$, but there was no significant difference between the inference and incentive groups, 397 versus 385, $t(79) = 0.619$, $p = .538$. Similarly, the inference group significantly outperformed the low-aspirations group, 397 versus 332, $t(79) = 3.15$, $p = .002$, but not the high-aspirations group, 397 versus 374, $t(79) = 1.35$, $p = .181$.

In other words, averaged across both lenient and exacting environments, subjects who either learned under incentives in Rounds 1 and 2 or who were given high aspirations performed as well in Round 3 as the inference subjects who had been given a set to learn despite the fact that the latter took much longer to perform the task. The inference subjects did, however, perform better than subjects who, during the first two rounds were either in the no-incentives condition or who were given low aspirations.

Concerning the prior-aspirations conditions (high vs. low), there is a large difference in Round 3 performance of subjects who were previously exposed to high aspirations, 374, as opposed to those who were previously exposed to low aspirations, 332, $F(1, 76) = 4.00$, $MS_e = 8{,}692$, $p = .049$, but no main effect for exactingness nor a significant interaction between exactingness and prior-aspirations conditions. Subjects in the low-aspirations group showed relatively poor performance in Round 2 (mean score of 330) and did not improve when switched to incentives in Round 3 (mean score of 332).

The inference group had the highest scores on the indices of understanding for both sign and slope. An ANOVA shows main effects for prior learning in respect of sign, $F(4, 190) = 3.64$, $MS_e = 1.45$, $p = .007$, and slope, $F(4, 190) = 3.28$, $MS_e = 0.67$, $p = .013$. In terms of contrasts, the mean score achieved by the inference group for sign of 1.54 was significantly greater than that achieved by the others, 0.89, $t(198) = 3.04$, $p = .003$. For slope, the analogous figures were 0.87 vs. 0.41, $t(198) = 3.10$, $p = .002$. Thus, in addition to more time spent on the task, the set to learn was accompanied by a greater ability to articulate the appropriate predictive relation.

Finally, correlations between individual scores on the understanding indices and performance were also high for the inference group. For Round 3, these were .69 for sign, .59 for slope, and .73 for the index of total knowledge, $p < .001$ for all. In addition, there was a relation between how well inference group subjects thought they had performed in the task after Round 3 and actual performance in evaluation points, $r = .51$, $p < .001$.

Discussion

The results of Experiment 4 show that performance in Round 3 (in which all subjects were in an incentives condition) reflects subjects' prior exposure to the decision-making task. The inference subjects outperformed those in the no-incentives and low-aspirations conditions but did no better on average than the incentives and high-aspirations groups. On the other hand, the inference subjects took on average 54% longer to complete the experimental tasks, which suggests that gains in performance should be measured against additional costs in time. Subjects who had previously been provided with low aspirations performed at a lower level than those who were given high aspirations. Finally, subjects in the inference condition were more capable of articulating an accurate understanding of the task.

Although the inference subjects were instructed to learn the game in Rounds 1 and 2 and thus ignore the evaluative dimension of feedback, it is unclear whether people can ignore the evaluative implications of any feedback. Two pieces of evidence support this notion. First, if exacting feedback has greater potential for learning, one would expect subjects in the exacting condition to have learned more effectively in the absence of evaluation. However, mean performance in Round 3 between inference subjects in the lenient and the exacting environments did not differ significantly, 389 versus 405.

Second, whereas from our viewpoint scoring performance of the inference subjects lacks meaning for Rounds 1 and 2, these subjects still observed the evaluation points that they would have achieved. Moreover, their performance (π) was comparable with subjects in the other conditions, 283 versus 285 for Round 1, and 359 versus 342 for Round 2, thereby suggesting that they had not been penalized for experimenting more than subjects who had been instructed to maximize evaluation points.

A further interesting Round 3 comparison can be made between the level of performance (π) obtained by the inference group and subjects in the intermediate exactingness environment of Experiment 1. These were 397 for the former and 415 for the latter but did not differ significantly. In addition, there were no significant differences between the scores that both groups achieved on the sign and slope indices. In short, there were no significant differences in either performance or understanding between the inference subjects, averaging over lenient and exacting environments, and subjects in an environment of intermediate exactingness, averaging over conditions of incentives and no incentives. The data show that there are different paths to the same levels of performance and understanding.

Experiment 5

Rationale

Prediction 3 states that if the negative effects of exactingness (α) on performance (π) are eliminated, performance (π) should increase as a function of exactingness (α)—see Equation 5. Experiment 5 was designed to test this hypothesis. Further, because eliminating the negative effects of exactingness reduces the impact of negative feedback, inconsistency of response strategies should not be expected to increase with exactingness.

Method

Subjects. Subjects were recruited in the same manner and from the same population as in the other experiments. The plan was to have 80 subjects in the experiment but, because after running several subjects, it became clear that there were "outliers," a total of 90 subjects were finally recruited to participate in the task (see below).

Task. The task was identical to that used in Experiment 1 (using the same discontinuous incentive function) except that subjects were provided with feedback both in the form of evaluation points and the correct value of the decision variable, that is, D—see Equations 1 and 2. The rationale was that correct outcome feedback would eliminate the negative effects of exactingness because there would no longer be any ambiguity concerning the meaning of feedback expressed in evaluation points, that is, subjects could also measure performance by differences between Q, the decision variable, and D, the outcome.

Design and procedure. Each subject was allocated at random to one of four groups created by crossing two levels of incentives (incentives vs. no incentives) by two types of environment (lenient vs. exacting). As in previous experiments, there were three rounds each involving 30 trials. After Round 2, subjects completed a questionnaire, and in Round 3 all subjects faced the same incentives condition.

Results

Distributions of scores on individual performance (π) indicated several outliers. To eliminate outliers, we used two criteria based on performance achieved in Round 3 and time taken on the task. Subjects were eliminated if their performance was greater than five standard deviations from the mean of the distributions of their experimental conditions (excluding outliers), or if total time taken for the whole task was either under 20 min or over 3 hr. This left the data of 79 subjects for subsequent analysis, 20 in each of the lenient–incentives and exacting–no-incentives conditions, 18 in the exacting–incentives conditions, and 21 in the lenient-no-incentives conditions. Ten of the 11 subjects eliminated were in the exacting conditions.

Performance (π). Results by experimental conditions and rounds are presented in Table 4. Subjects in the exacting conditions outperformed those in the lenient conditions in all three rounds. Moreover, this difference in performance (π) is statistically significant in both Rounds 2 and 3, $F(1, 75) = 5.05$, $MS_e = 1{,}987$, $p = .027$, and $F(1, 75) = 5.34$, $MS_e = 1{,}547$, $p = .023$. The only other statistically significant effects for performance (π) are in respect of round, between Rounds 1 and 2, overall means of 427 and 471, $F(1, 73) = 128.9$, $MS_e = 62{,}381$, $p < .001$, but not between Rounds 2 and 3, means of 471 versus 476.

Inconsistency of response strategies. Subjects in the exacting environment were not more inconsistent in their use of response strategies than those in the lenient condition. If anything, their measures of inconsistency were smaller, with means averaged over Rounds 1 and 2 of 64 versus 84 for those in the lenient condition. This difference, however, was not statistically significant.

Table 4
Experiment 5: Performance (π) and Evaluation Points in Lenient and Exacting Environments

Round	Incentives		No incentives	
	Lenient	Exacting	Lenient	Exacting
	Performance (π)			
1				
M	423	430	405	449
SD	62	79	93	38
2				
M	466	481	454	484
SD	46	32	66	18
3				
M	468	483	463	489
SD	50	22	54	6
	Evaluation points			
1				
M	278	−10,075	210	−5,397
SD	257	16,914	373	5,919
2				
M	442	−594	398	−471
SD	151	2,823	210	1,633
3				
M	445	−171	432	196
SD	146	1,614	141	606

Further results. One important difference between subjects in the lenient and the exacting conditions was that the latter took more time to complete the task. Mean times by rounds were, for lenient and exacting, respectively, 30.1 versus 37.8 for Round 1, *ns*; 17.0 versus 21.0 for Round 2, $F(1, 75) = 4.74$, $MS_e = 66.6$, $p = .03$; and 15.0 versus 20.0 for Round 3, $F(1, 75) = 7.45$, $MS_e = 67.5$, $p = .008$. Using time as a covariate, reanalysis of the data showed no main effect for exactingness on performance (π) in any of the rounds. Time, however, was only weakly correlated with performance (π) at the individual level, $r = .21$, *ns*, .29, $p < .05$, and .12, *ns*, for Rounds 1, 2, and 3, respectively.

In terms of the questionnaire administered after Round 2, the only statistically significant differences revealed for either attitudes toward the task or understanding of how the underlying system worked were that, in the exacting environment, subjects found the experience more discouraging than those in the lenient environment, means of 4.04 versus 2.43 on a 7-point scale, $F(1, 75) = 19.4$, $MS_e = 2.63$, $p < .001$. This was also true of subjects in the incentives as opposed to the no-incentives condition, means of 3.74 versus 2.73, $F(1, 75) = 7.64$, $MS_e = 2.63$, $p = .007$.

At the individual subject level, correlations between the indices of understanding and performance (π) were significant. For example, the correlation between the index of total understanding and performance was .38, $p = .001$, .53, $p < .001$, and .50, $p < .001$ for Rounds 1, 2, and 3, respectively.

Discussion

The main result from Experiment 5 was that subjects in the exacting environment outperformed those in the lenient environment, thereby validating Prediction 3. Another difference was that subjects in the exacting environment took more time to complete the task.

In addition to superior performance, further evidence that the provision of correct outcome feedback mitigated the negative aspect of exactingness was that inconsistency in strategy use was not significantly greater in the exacting condition.

Total time taken on the task by subjects in this experiment exceeded that in Experiments 1, 2, and 3, mean of 70.5 min versus 50.0, 49.0, and 53.4 min, respectively, $t(198) = 5.16$, $t(157) = 4.77$, and $t(159) = 3.94$, $p < .001$, for all. In fact, the total time taken in this experiment was similar to that in the inference condition of Experiment 4, mean of 74.8 min. That subjects in this experiment took more time than those in Experiments 1–3 might be rationalized by the fact that they had more information to consider, that is, revealed values of the D variable. On the other hand, this information also made the task easier to understand. Perhaps what these data suggest is that the presence of correct outcome feedback gave subjects more to think about when they were planning their decisions than in situations in which this information was absent. In the latter, one can imagine that because of ignorance, subjects put more emphasis on learning through taking action and observing, as opposed to thinking more about outcomes that were already observed.

We believe that caution should be exercised in interpreting our results because most of the outliers excluded from the data analysis were in the exacting conditions. For the most part, these subjects rushed through the task (in under 20 min) and their performance (π) showed little or no improvement from Rounds 1 to 3. Their data were so different from the majority of subjects in the exacting conditions that we had few scruples in eliminating them from analysis. On the other hand, they do also suggest that, for a minority of subjects, provision of the "correct" amount to be predicted, D, did not eliminate the negative aspect of exactingness on performance.

The astute reader will have noticed an additional implication of the model in Equation 5 with respect to the effects of incentives. When there are no negative effects of exactingness, incentives should improve performance (because $b' > b$). Why, then, was there no effect for incentives in the present experiment? The reason, we believe, is that because most subjects received positive feedback following their decisions, they did not experience a discontinuity between being rewarded and not being rewarded. In other words, for subjects the incentive function was experienced as being more like that in Experiment 3 than Experiment 1. For example, in Round 2 the percentages of subjects who received positive mean evaluation points were 85% and 74% in the lenient and exacting conditions, respectively. Thus, given the levels of performance achieved, the presence or absence of incentives did not differentially accentuate exactingness.

General Discussion

We first review the major findings of our experiments, distinguishing between effects of exactingness, effects of incentives, and other issues. Subsequently, we discuss these results from both theoretical and practical perspectives. We also suggest topics for further study.

Effects of Exactingness

Our experiments demonstrated that exactingness (α) has both positive and negative effects on performance. First, consistent with our theoretical model, we showed in Experiment 1 that an intermediate level of exactingness ($\alpha = .05$) resulted in superior performance to lenient ($\alpha = .01$) and exacting ($\alpha = .50$) environments—Prediction 1. Second, when the negative effects of exactingness were mitigated by using correct outcome feedback in addition to evaluation points (Experiment 5), performance (π) was better in the exacting environment—Prediction 3.

Critical to our underlying model was the assumption that, during learning, people react differently to positive and negative feedback. Positive feedback reinforces the use of existing strategies, negative feedback encourages the search for other strategies that might work better. Because positive feedback is likely in lenient environments, and negative feedback is likely in exacting environments, we postulated greater inconsistency in strategy use in exacting environments. This implication was validated in Experiments 1 and 3 but not in Experiment 2, in which incentives took the form of manipulated aspirations. On the other hand, as noted in the discussion of that experiment, whether feedback was encoded as positive or negative was ambiguous in that subjects could attend to

aspiration levels, the sign of evaluation points achieved, or both. In Experiment 5, we provided correct outcome feedback to mitigate the negative effects of exactingness. Thus, in this case the lack of an effect for exactingness on inconsistency in strategy use was consistent with our theoretical expectations.

Based on work within the lens model tradition (Hammond & Summers, 1972), one might imagine that performance (π) would always be inversely related to inconsistency in strategy use so that measures of these variables would be redundant. However, it is important to state that this is not the case. In Experiment 1, for example, subjects in both the lenient (α = .01) and intermediate (α = .05) conditions did not have significantly different levels of inconsistency in strategy use, and yet, performance in the intermediate was better than in the lenient. In other words, although subjects in both conditions were equally inconsistent, those in the intermediate environment learned better strategies. Note also that in this experiment, although subjects in the lenient condition were less inconsistent than those in the exacting environment, their performance was not significantly different.

Questionnaires constructed to determine subjects' attitudes toward the experimental task as well as their understanding of the underlying task structure revealed some differences that were caused by exactingness. For example, subjects found the lenient task to be more enjoyable than the exacting (Experiment 3), and less discouraging (Experiment 5). Measures of understanding (e.g., the sign and slope indices) were largest in the intermediate environment (Experiment 1). An interesting finding was that, in all of our experiments, subjects' ability to articulate their understanding of the task was highly correlated with performance.

Effects of Incentives

Our theoretical model stated that incentives would accentuate the effects of exactingness, thereby leading to an Incentives × Exactingness interaction on performance—Prediction 2. We investigated this prediction under three conditions. In Experiment 1, we employed an incentive function with a discontinuity in which subjects could easily discriminate successful (cash reward) versus unsuccessful (no cash reward) performance. In Experiment 3, on the other hand, we employed an incentive function that effectively guaranteed that all subjects would receive some cash reward (although the better they did, the more they earned). The predicted interaction was validated in Experiment 1 but not in Experiment 3, thereby suggesting that, to have an effect, external incentive schemes need to be "sharp." We also induced differential incentive conditions in Experiment 2 by manipulating subjects' aspirations and, although we obtained the predicted interaction, the effect was weaker than in Experiment 1. As noted above, we have reason to believe that the aspiration-level manipulation did not necessarily lead subjects to make clear distinctions between successful and unsuccessful performance, and this may have induced the weaker effect.

We obtained no effects for incentives in Experiment 5 (with correct outcome feedback), but we noted that almost all subjects received positive feedback following their decisions. Thus, incentives had little opportunity to accentuate the effects of exactingness.

Other Issues

Experiment 4 addressed the issue of how different conditions experienced during learning affect subsequent performance. This was achieved by seeing how all subjects performed in a third round of 30 trials administered under incentive conditions. Results showed no effects between lenient and exacting environments, but there were differences caused by prior incentive conditions. Subjects who learned either under no incentives or with low aspirations (Experiment 2) performed less well. Of particular interest were subjects who, during the first two rounds, were instructed to learn the task as opposed to maximize evaluation points. Their performance was at the same level as that of subjects who learned under incentives. They differed, however, in that they took more time and were capable of expressing a better understanding of the underlying task. On the other hand, the performance of these inference subjects on both the task and indices of understanding was matched by the subjects in Experiment 1, who had been exposed to the intermediate level of exactingness (α = .05). In other words, in terms of performance, training in the intermediate environment (with or without incentives) was just as effective as training in a learning mode in the lenient and exacting environments. The effectiveness of an intermediate level of exactingness, however, can be measured by the fact that it took one third less time to achieve the same measures of performance.

Other differences in time were also noted. In Experiment 5 (with correct outcome feedback), subjects took more time in the exacting as opposed to the lenient conditions. Moreover, we interpreted these findings as indicating that the differential levels of exactingness induced these differences. Of additional interest was the finding that the total time taken by the subjects in Experiment 5 matched that taken by the inference subjects in Experiment 4. In one sense, because of the additional feedback, subjects in Experiment 5 had an easier task than did those in the other experiments (their performance was also better). We speculate that this additional feedback led to a more reflective style of decision making as opposed to learning "through doing" that was more characteristic of subjects who only received feedback in the form of evaluation points.

Theory

From a theoretical viewpoint, our studies break new ground in that we explicitly consider the possible effects of exactingness on learning in a decision-making task. Central to our model is the notion that there are positive and negative effects of exactingness that trade off so that learning (as measured by performance, π) is an inverted-U-shaped function of exactingness. On the positive side, as exactingness increases, so do opportunities for learning. The negative aspect reflects how people react to feedback. If feedback is interpreted as negative (which is increasingly likely with greater levels of exactingness), we argued that inconsistency in responses induced by

trying out alternative strategies would lead to lower levels of performance. Exactingness therefore influences learning because people react differently to positive and negative feedback.

An issue that we have not addressed in this article is the relation between exactingness and task complexity (for discussions of the latter concept, see Hammond, 1988; Wood, 1986). Conceptually, we propose that these concepts be treated as distinct in that whereas exactingness reflects the evaluative dimension of feedback (how well one is performing), complexity reflects the difficulty of the inferential dimension (how the underlying system works). However, because feedback is confounded, it is likely that exactingness and complexity are also confounded in the mind of the decision maker. This suggests, therefore, that because different levels of task complexity will affect the extent to which decision makers receive positive or negative feedback, task complexity and exactingness might have similar effects on performance (see, e.g., Wood, Bandura, & Bailey, 1990). Future research should explicitly address the effects of both complexity and exactingness and the extent to which they might have independent or interactive effects on performance.

As to incentives, our model suggests that these accentuate both the positive and the negative aspects of exactingness. Interestingly, this leads to an implication similar to data observed in several studies—albeit with respect to tasks that vary in complexity. This is that for tasks that are not well understood, incentives can be dysfunctional. As noted earlier, the main theoretical argument in the literature is that, in the presence of incentives, more complex tasks divert needed attention from inference to evaluation, that is, from a concern about how to do the task to how well one is doing (Humphreys & Revelle, 1984; Kanfer & Ackerman, 1989; Wood, Bandura, & Bailey, 1990).

These theories and findings are consistent within our conceptual scheme if one makes the distinction between the subtasks of developing an appropriate response strategy, on the one hand, and executing it, on the other. When tasks are relatively simple or understood, more effort can be devoted to execution, and this leads to improved performance. However, for more complex tasks, the chances of having or developing the appropriate strategy are smaller. Thus, if because of negative feedback, incentives induce greater inconsistency in execution, performance is degraded. The key, therefore, lies in how people react to positive and negative feedback.

When we talk about incentives, it is easy to overlook the fact that these can take many forms. For example, even within the class of external incentives, schemes for rewarding performance can vary greatly. One hint provided by our data is that incentive functions have to be sharp enough so that people can distinguish between good and bad performance. At the theoretical level, therefore, this point links nicely to the fact that people react differently to positive and negative feedback.

Practical Implications

Our results raise many practical issues. First, however, it is appropriate to consider the limitations of our experimental paradigm and the extent to which the findings might be expected to generalize to a wider range of situations. In many ways, our experiments provided almost ideal opportunities for learning compared with more realistic settings. Feedback after decisions was immediate. Subjects could take notes and consult their histories of past decisions. The task did not involve a large number of variables, and there was a limited number of relations between variables in the system that were important. Moreover, the system that generated observations did not change over time.

There are many real world tasks that exhibit similar characteristics, such as production and inventory scheduling decisions, predictions of economic and financial indicators, and weather forecasts. These tasks may differ from ours, however, in that people would typically not make so many decisions in such a short period of time. (Experiments tend to collapse experience in terms of time.) In one sense, real world tasks may also be more inferentially complex than ours; on the other hand, this complexity may be offset by having more time to think through issues before making decisions. Moreover, whereas our task involved abstract variables, the context of real world tasks engages knowledge that facilitates inference. Nonetheless, there are other real world tasks that are similar to ours and in which people do experience much feedback within fairly short periods of time. These tasks include learning how to handle mechanical or electronic devices that require frequent decisions and provide almost immediate feedback. Word processing systems are a good example.

Two important dimensions of real world tasks are whether people are aware of the exactingness of the environment and whether they or others have the ability to control or manipulate it. In many situations in which outcomes and rewards are the same (as in financial transactions), people are typically ignorant of the effects of exactingness. Thus, incentives may or may not promote effective learning. In this case, it would be advisable to learn to make decisions within an inference set (as in Experiment 4) before having to deal with real payoffs. On the other hand, in situations in which it is possible to control how decisions are evaluated (as in our experiments), this may be used deliberately in training decision makers. The implications from our results are clear. Intermediate environments induce more effective learning than do lenient or exacting ones, and in this kind of environment incentives make little difference. If one is forced into using a lenient evaluation function, however, use incentives; with an exacting function, do not use incentives.

Finally, in our task subjects were not told how they were evaluated, that is, how decisions and outcomes were translated into evaluation points. An argument could be made that learning would be fostered if people were aware of the exact nature of the evaluation function because this would reduce one source of ambiguity in feedback and, indeed, this was shown to be the case by the results of Experiment 5. On the other hand, because different evaluation functions induce different rates of learning, it is not clear that it would always be advantageous to reveal these functions to learners. We believe that this issue should be explored further in future studies.

Issues for Further Study

Because, to the best of our knowledge, the effects of exactingness and incentives have not previously been studied together, the present research suggests many issues for further investigation. We mention a few.

First, the penalty functions used in our tasks were symmetric. Subjects received the same penalty if they overshot or undershot the appropriate setting of the decision variable. It would also be interesting to investigate different types of asymmetric penalty functions. In particular, with highly skewed functions subjects would experience large variations in penalties that might be similar in effect to exacting environments with symmetric functions. However, they would probably also learn to adjust responses to avoid the larger penalties. In our work, we adopted a simple mechanism to model exactingness in the environment. It is possible that this could be achieved in other ways.

Second, although we examined the effects of different types of incentive functions (i.e., sharp vs. continuous) the present work ignored the fact that there could be many different levels of incentives. Thus, although the level of incentives used with the sharp function was sufficient to induce effects, we have no information concerning the relation between size of incentives and effects. We suspect that in a laboratory task relatively small differences in real money paid to subjects do have motivational effects (see also Arkes et al., 1986; Edwards, 1956; Hogarth & Einhorn, 1990), but we are uncertain about how this might generalize outside the psychological laboratory.

Third, our studies have been based on a single experimental task. We believe that it is important to investigate the effects of incentives over a wider range of experimental tasks but, in doing so, we see the necessity of developing an appropriate taxonomy of tasks. In particular, whereas our studies explored the effect of different penalty functions, feedback could also prove more or less frustrating to subjects, depending on the complexity of the underlying causal model generating outcomes. We therefore need to understand how complexity interacts with exactingness. Recently, Hammond et al. (1987) have elaborated a theory of how characteristics of tasks map into different modes of cognition that vary on a continuum from analysis to intuition. Hammond et al. would classify our task as *analysis inducing* so that it would best be handled by an analytic mode of cognition. Whether our theoretical framework and results would also apply in tasks that could be defined as *intuition inducing* is an open and important issue.

Fourth, we noted above that by informing people of the nature of penalty functions, one should, in principle, reduce the ambiguity of outcome feedback. However, because feedback still implies an evaluation, it is not clear that people are able to separate the informational content of feedback concerning the inferential structure of the task from its evaluative component (cf. Experiment 5). This suggests conducting studies similar to those reported above in which the nature of the penalty function is made explicit to subjects. The question asked is whether it is necessarily better to inform people how they are being evaluated.

Fifth, a central premise of this work is that feedback is ambiguous. Given this ambiguity, it is legitimate to ask whether people might learn more effectively if they received less rather than more information about the effectiveness of past decisions. For example, instead of providing feedback for each decision, would subjects perform better by the end of the experimental session if they only received feedback in the form of average statistics over small blocks of trials? Advantages are that subjects might be forced to experiment with particular strategies over specific blocks of trials and the effects of random error would be mitigated by the averaging process.

To conclude, we have demonstrated that changes in the parameter of the function that evaluates outcomes of decisions can induce significant changes in performance as well as reverse the sign of the effects of incentives. Such sensitivity to a single task feature merits more detailed attention.

References

Amabile, T. M. (1982). Children's artistic creativity: Detrimental effects of competition in a field setting. *Personality and Social Psychology Bulletin, 8,* 573–578.

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes, 37,* 93–110.

Ashton, R. H. (1990, April). *Paradoxical effects of incentives, feedback, and justification in accounting decision settings.* Unpublished manuscript, Fuqua School of Business, Duke University.

Bahrick, H. P. (1954). Incidental learning under two incentive conditions. *Journal of Experimental Psychology, 47,* 170–172.

Bahrick, H. P., Fitts, P. M., & Rankin, R. E. (1952). Effects of incentives upon reactions to peripheral stimuli. *Journal of Experimental Psychology, 44,* 400–406.

Balzer, W. K., Doherty, M. E., & O'Connor, R., Jr. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106,* 410–433.

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology, 36A,* 209–231.

Berry, D. C., & Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology, 79,* 251–272.

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica, 45,* 223–241.

Broadbent, D. E. (1977). Levels, hierarchies, and the locus of control. *Quarterly Journal of Experimental Psychology, 29,* 181–201.

Broadbent, D. E., & Aston, B. (1978). Human control of a simulated economic system. *Ergonomics, 21(12),* 1035–1043.

Broadbent, D. E., FitzGerald, P., & Broadbent, M. H. P. (1986). Implicit and explicit knowledge in the control of complex systems. *British Journal of Psychology, 77,* 33–50.

Coombs, C. H., & Avrunin, G. S. (1977). Single-peaked functions and the theory of preference. *Psychological Review, 84(2),* 216–230.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* New York: Plenum Press.

Earley, P. C., Connolly, T., & Ekegren, G. (1989). Goals, strategy development, and task performance: Some limits on the efficacy of goal setting. *Journal of Applied Psychology, 74(1),* 24–33.

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological Review, 66(3),* 183–201.

Edwards, W. (1956). Reward probability, amount, and information as determiners of sequential two-alternative decisions. *Journal of Experimental Psychology, 52,* 177–188.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13,* 171–192.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85,* 395–416.

Hammond, K. R. (1988). Judgment and decision making in dynamic tasks. *Information and Decision Technologies, 14,* 3–14.

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-17* (5), 753–770.

Hammond, K. R., & Summers, D. A. (1972). Cognitive control. *Psychological Review, 79,* 58–67.

Hammond, K. R., Summers, D. A., & Deane, D. H. (1973). Negative effects of outcome-feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance, 9,* 30–34.

Hogarth, R. M., & Einhorn, H. J. (1990). Venture theory: A model of decision weights. *Management Science, 36,* 780–803.

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review, 91(2),* 153–184.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74(4),* 657–690.

Klayman, J. (1988). On the how and why (not) of learning from outcomes. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 115–162). Amsterdam: North-Holland.

Lepper, M. R., & Greene, D. (Eds.). (1978). *The hidden costs of reward.* Hillsdale, NJ: Erlbaum.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology, 28,* 129–137.

Levine, F. M., & Fasnacht, G. (1974). Token rewards may lead to token learning. *American Psychologist, 29,* 816–820.

Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance: 1969–1980. *Psychological Bulletin, 90,* 125–152.

Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.) *Handbook of mathematical psychology, Vol. III* (pp. 249–410). New York: Wiley.

McCullers, J. C. (1978). Issues in learning and motivation. In M. R. Lepper & D. Greene (Eds.), *The hidden costs of reward* (pp. 5–18). Hillsdale, NJ: Erlbaum.

McGraw, K. O. (1978). The detrimental effects of reward on performance: A literature review and a prediction model. In M. R. Lepper & D. Greene (Eds.), *The hidden costs of reward* (pp. 33–60). Hillsdale, NJ: Erlbaum.

McGraw, K. O., & McCullers, J. C. (1979). Evidence of a detrimental effect of extrinsic incentives on breaking a mental set. *Journal of Experimental Social Psychology, 15,* 285–294.

Schwartz, B. (1982). Reinforcement-induced behavioral stereotypy: How not to teach people to discover rules. *Journal of Experimental Psychology: General, 111(1),* 23–59.

Siegel, S. (1961). Decision making and learning under varying conditions of reinforcement. *Annals of the New York Academy of Sciences, 89,* 766–783.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology, 71,* 680–683.

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review, 66,* 297–333.

Wood, R. E. (1986). Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes, 37,* 60–82.

Wood, R., Bandura, A., & Bailey, T. (1990). Mechanisms governing organizational performance in complex decision-making environments. *Organizational Behavior and Human Decision Processes, 46,* 181–201.

Zajonc, R. B. (1965). Social facilitation. *Science, 149,* 269–274.

Received October 2, 1989
Revision received October 26, 1990
Accepted December 14, 1990 ■