Rule discovery performance unchanged by incentives

Anonymous CogSci submission

Abstract

Human behavior is modulated by financial incentives, but it is not well understood what types of behavior are immune to incentive and why. The cognitive processes underlying behavior appear to create restrictions on the effect an individual's motivation will have on their performance. We investigate a classic category learning task for which the effect of financial incentives is still unknown (Shepard, Hovland, & Jenkins, 1961). Across four renditions of the category learning experiment, we find no effect of incentive on performance. On a fifth experiment requiring category recognition but NOT learning, we find a large effect on response time and small effect on task performance. Humans appear to selectively apply more effort in valuable contexts, but the effort is disproportionate with the performance improvement. Taken together, the results suggest that performance in tasks which require novel inductive insights are relatively immune to financial incentive, while tasks that require rote perseverance of a fixed strategy are more malleable.

Keywords: categorization; inductive reasoning; motivation; learning; behavioral economics

Introduction

Human behavior is powerfully shaped by incentives. Financial rewards in particular have an integral role in culture and society, dictating punishments for improper behavior or motivating capitalizing behavior. In this paper we explore the consequences of changing incentives on cognitive performance.

While incentives are powerful at shaping behavior, humans are not only guided by incentives but are also limited by their cognitive capacities. High incentives thus do not correspond linearly with unaided performance, since behavior relies on many cognitive skills that a person may not be able to voluntarily augment. For example, a job offer to receive a million dollars to memorize the dictionary in a week would not encounter many successful applicants. Understanding what types of behavior are or are not easily modified with reward and incentives has several important implications at the level of policy and society.

Recent work suggests that cognitive processes like attention and effort can be modulated by incentive. DellaVigna and Pope (2018) show that effort, measured by number of button presses, increased substantially with higher probability and magnitude of payment. Additionally, Caplin, Csaba, Leahy, and Nov (2020) demonstrate how attention can be incentivized in simple perceptual tasks (e.g., counting the number of 7- versus 9-sided polygons in a crowded display). Performance and the time participants spend on the task scales with the point value of each trial.

However, there are also behavioral tasks that appear to be surprisingly immune to the effects of incentive. When testing whether people are less susceptible to cognitive biases when incentives are high, Enke et al. (2021) found that while response times increased substantially in a high stakes condition, there was only a weak effect on reducing cognitive biases. Similarly, van den Berg, Zou, and Ma (2020) found no effect of incentive on visual working memory performance. Even though these tasks rely on elements of cognition like attention and effort, it is evident that the reward modulation of higher-order cognitive function is potentially more nuanced.

Understanding these differences is an important frontier for cognitive research. First, to the degree that certain cognitive processes can be modulated by incentives, most models of human cognition do not precisely account for these effects. One exception is the recent work on "resource rational" theories of cognition which try to explicitly weight the cognitive cost of various more elementary cognitive operations against the benefits to task performance (Gershman, Horvitz, & Tenenbaum, 2015; Shenhav et al., 2017; Bhui, Lai, & Gershman, 2021). Second, understanding which types of tasks respond to incentives and which do not may help us understand ways to encourage better cognitive performance from athletes, students, and the general public.

Incentives and Rule Discovery

Rule discovery is a ubiquitous and ongoing human task (e.g., categorizing effective from ineffective COVID-19 masks), and also reflects the type of creative problem solving that is required in many professions. Successful rule learning, particularly for complex rules and patterns, requires coordination of several cognitive processes including attention, working memory, reasoning, and decision making.

In the current study, we investigate the effect of financial incentive on performance in a classic category learning rule discovery task known as the SHJ task for the initials of the lead authors (Shepard et al., 1961). In the task, participants must attend to relevant stimulus features to correctly distinguish two groups of objects. For the same eight stimuli (see Figure 1), different group assignments of the stimuli varies the difficulty of the task, as some categorizations are learned more quickly than others. Although participants could simply

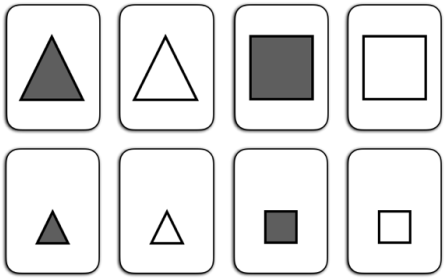

Figure 1: The eight cards differing in three feature dimensions: shape, size, and color. Across different conditions subject learn through experience that half of the stimuli belong to the same group and are asked to discover the rule that determines group membership.

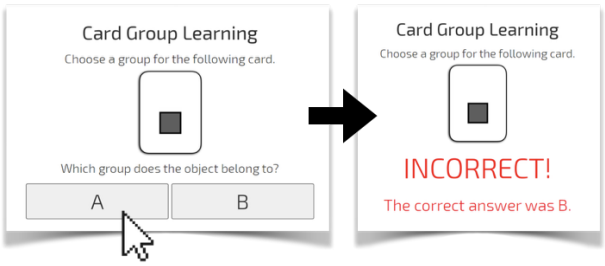

Figure 2: In the learning phase of the experiment, participants guess the category membership of each card and receive feedback on each trial.

memorize the category memberships for the eight stimuli no matter the grouping, a classic and highly replicated finding is that humans rely on inductive biases to find an explanation of the categorization based on simple descriptions of stimulus features (Kruschke, 1992; Nosofsky, Palmeri, & McKinley, 1994; Love, Medin, & Gureckis, 2004; Goodman, Tenenbaum, Feldman, & Griffiths, 2008).

Although this task has been replicated and tested many times, it is still unknown how incentive will change category learning behavior. Intriguingly though, one recent paper examining methods for obtaining high quality data from Amazon Mechanical Turk (mTurk) found no effect of changing the magnitude of incentives on category learning performance in a subset of the SHJ categorization tasks (Crump, McDonnell, & Gureckis, 2013). Here we aimed to perform a more systematic evaluation of incentives on category learning behavior that better controlled the rates and sizes of task payments inspired by recent work in economics on theories of rational inattention (Caplin et al., 2020).

Experiments

Stimuli

Figure 1 displays the eight cards used in the experiment, each of which contains an object with a particular shape (square or triangle), size (large or small), and color (white or black). The eight stimuli can be divided into two equal groups a total of 70 unique ways (calculated with $\binom{8}{4}$). Each of these 70 groupings fits into one of six rule types described by Shepard et al. (1961), which differ in the number of stimulus features required to define the rule. For example, a Type I rule varies the group along a single feature dimension, giving a rule like "Large objects are in group A and small objects are in group B." A Type II rule groups the stimuli along two feature dimensions; for example, "Black triangles and white squares are in group A." Rule Types III, IV, and V rely on all three features, and allow a "rule-plus-exception" type explanation such as "Large shapes are in category A, except for the white square." Rule Type VI applies to groupings that cannot be described by a simple feature-based rule. In these cases, the group membership of each stimulus must be memorized, making Type VI categorizations the hardest to learn.

Procedure

Experiments 1-4 derive their design from the classic category learning task introduced by Shepard et al. (1961). During a learning phase, subjects use trial-and-error to actively learn the assignment of eight stimuli into two groups. Immediately after this phase, subjects perform a test in which they report the group membership of each stimulus once. The experiments below all retain this same basic structure but differ in the number of trials in the learning phase, the modality of the financial incentive, and whether the manipulations were between- or within-subjects.

Participants were recruited via Amazon mTurk, and the experiment was restricted to users in the United States. The task was designed in JavaScript and delivered to the participants' browser via psiTurk (Gureckis et al., 2016). Subjects received a base payment that corresponded with the expected length of completing the task at a rate of $0.15 per minute, and could receive a performance-based bonus of up to $10. To determine whether or not they would receive the bonus, participants would see their local clock on screen with time shown to the milliseconds. They would click on a button to stop the clock, and the last two digits of the stopped time (milliseconds) would be compared to their bonus probability (explained in detailed in Methods, Exp. 1); if the two digits were at or below the bonus probability, they would receive the bonus. A page in the instructions allowed subjects to test stopping and starting the clock to demonstrate that the last two digits were random and could not be controlled. This emphasis on the random nature of the bonus served to minimize expectations of deception from the experimenters.

Participants underwent a rigorous instructions phase followed by a comprehension check to ensure their complete understanding of the task and incentives before beginning the task. Although participants were asked not to take notes or pictures during the experiment, we also asked them to honestly report at the end if they had used any memory help, knowing that their payment would remain the same regardless of their response. In addition to these participants excluded due to admission of using externalized memory aids,

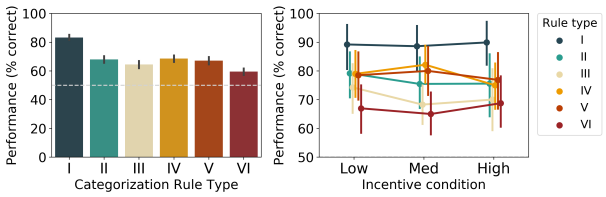

Figure 3: Left: Experiment 1 performance by rule type condition across all trials (both learning and test phase). Performance is percent correct across all learning and test phase trials. A gray dashed horizontal line shows performance at chance. Right: Experiment 1 test phase performance by incentive condition for the six rule types. Performance is measured only on test phase trials. Both: Error bars show 95% confidence intervals.

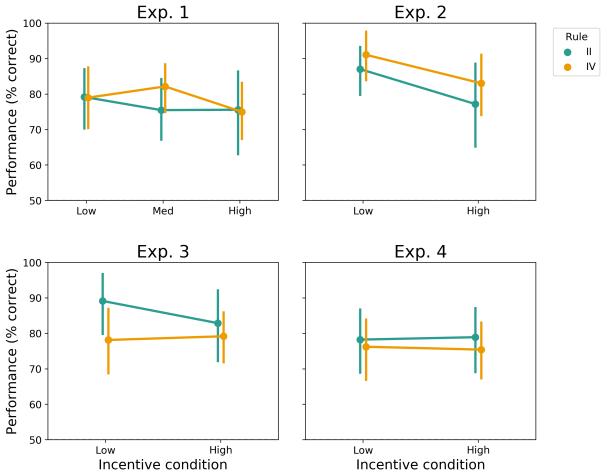

Figure 4: Performance on the test phase trials across conditions for the four category learning experiments.

subjects were also excluded if they encountered an error in the experiment that prevented completion.

Analyses

We use a Bayesian logistic regression model to predict the probability of a correct response as a function of incentive and rule type. The posterior was estimated with Markov Chain Monte Carlo (MCMC) sampling in the Bambi python package (Capretto et al., 2020). To fit each model, four MCMC chains each ran 2000 samples, the first half of which were discarded. The results sections report 94% Highest Density Intervals (HDIs) for the relevant model parameters.

Experiment 1 - Between Subjects Manipulation of Incentives

In Exp. 1, subjects learn one of six rule types at one of three incentive levels (low, medium, and high) with a completely between-subjects design. If learning performance in the task is modulated by incentive, we expected to see an increase in performance for subjects in higher incentive conditions at the same rule type.

Participants Exp. 1 tested 418 subjects across 18 conditions, not including 6 subjects who admitted to using external help. The task took approximately 15 minutes and subjects were paid a $2.25 base rate for their time, with the chance of earning a $10.00 bonus depending on their performance.

Methods and Design Exp. 1 tested all six rule types at three different incentive levels (18 conditions). The design was completely between-subjects to avoid learning effects across blocks. In Exp. 1, the learning phase consisted of 16 trials (two repeats of each stimulus in a random order). Participants were instructed that the purpose of the learning phase was simply to learn the groupings and did not determine their bonus. Their performance in the test phase determined the chance of winning the bonus.

The instructions explained how better performance on the test would increase their chance at winning a $10.00 bonus. To make the probabilistic nature of the incentive clear, we explained the probabilities in terms of pulling a marble out of a bag: "Imagine a bag full of red and blue marbles. We pull out a marble at random. If the marble is blue, you win the extra $10. If the marble is red, you receive only the base payment." Depending on the incentive condition the subject was assigned to, they would be shown a certain number of red marbles in the bag. Performance above chance on the test phase could turn some of the marbles in the bag blue, thus increasing their chance at winning the bonus. Since there were eight questions on the test, chance performance would mean getting four correct. Therefore, we replaced a red marble with a blue marble for each correct answer beyond chance performance, with a maximum possible four blue marbles to be earned if participants answered all eight questions correctly. If participants answered four or fewer of the eight test questions correctly, all of the marbles would remain red, meaning they had no chance to win the bonus.

The three incentive conditions differed by the maximal probability of winning the bonus, which we represented by changing the total number of marbles in the bag. The low, medium, and high incentive conditions showed 64, 8, and 4 marbles in the bag respectively. If a subject performed perfectly on the test and turned the maximum four marbles blue, this would correspond to a bonus probability of 6.25% in the low incentive condition, 50% in the medium incentive condition, and 100% in the high incentive condition. Correspondingly, each correct answer above chance was worth approximately 1.6%, 12.5%, or 25% in the low, medium and high incentive conditions respectively.

Results Our results replicate the main effect of rule on performance from Shepard et al. (1961) (Fig. 3). A logistic regression with the six rule types and the incentive condition as parameters confirmed that rule types II through VI had worse performance compared to type I, with all five rule type parameter means estimated to be negative and their upper tail

*HDI*s falling beneath -0.09. Rule Type VI had the greatest effect on performance, $HDI = [-.311, -.219]$. Across all incentive conditions and across learning and test trials, Type I was the easiest rule to learn and Type VI was the hardest rule to learn. The other four rules fell in between this range of difficulty. The model fit estimated a zero effect of incentive on performance, $HDI = [-.001, .001]$.

The right plot in Figure 3 shows participants' performance for the three different incentive conditions across the six unique rule types. Although the expected value of perfect performance in the high incentive condition was over fifteen times larger than that of the low incentive condition, performance stayed constant. Additionally, no meaningful patterns were found in the types of errors participants made, such as learning a simple type I rule even though they were assigned a type V rule (space prohibits an extensive report of those analyses). Response times on the test trials are reported in Figure 5. A logistic regression model fit on the response time data confirmed no effect of incentive on response time, $HDI = [-12.6, 12.3]$.

Experiment 2 - Additional Learning Trials

Given the null result of incentive in Exp. 1, we were concerned that 16 learning trials was too few chances to adequately learn the categories. The number of learning trials had been selected to avoid ceiling effects on easier rule types based on learning curves from Nosofsky et al. (1994), but it may have created a floor effect on the more difficult rule type conditions. Therefore, we doubled the amount of learning trials for Exp. 2. In addition, we focused on Rule Types II and IV, since these are both non-trivial but still incorporate rule discovery unlike Type VI. Rule Types II and IV are qualitatively different, and previous work has shown Type II performance to be better even when performance in Type III, IV, and V problems is indistinguishable (Nosofsky et al., 1994).

Participants In Exp. 2, we collected 97 participants across four conditions, not including 3 subjects who admitted to using external help. Because the length of the learning phase had increased, we increased the base payment to $2.50.

Methods and Design Exp. 2 replicated the design of Exp. 1 but with the number of learning trials increased to 32, so that participants saw four repeats of each of the eight stimuli during the learning phase. We gathered participants in four conditions; two rule type conditions (type II and type IV) crossed with two incentive conditions (low and high).

Results As in Exp. 1, we saw no effect of incentive on performance in either of the rule conditions tested. The average performance by condition in Exp. 2 is shown in the top right plot in Figure 4. A logistic regression fit on the Exp. 2 data showed a near-zero negative effect of incentive on performance, $HDI = [-.004, -.001]$. There was a marginal increase in response time as a function of incentive (top right of Fig. 5), with a regression model estimating the coefficient at a median of 14.6 ($MAD = 6.4$), $HDI = [-2.6, 31.4]$.

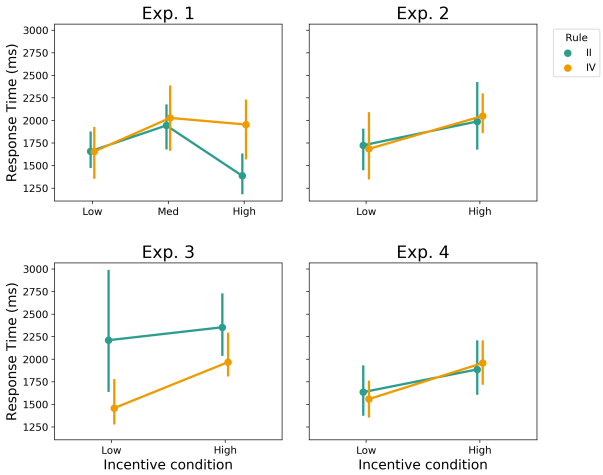

Figure 5: Median response times on the test phase trials across conditions for the four category learning experiments.

Experiment 3 - Manipulation of Reward Magnitude

Considering the null result from the first two experiments, we were concerned about the complexity of the probabilistic nature of the incentive. Perhaps subjects were uninfluenced by the incentive manipulation because they did not understand it or it did not seem meaningful. Not only is it possible that the participants do not trust the legitimacy of the randomness determining their winnings, but it is also challenging to convey such probabilistic information about the gambles.

Participants In Exp. 3, we collected 93 subjects across four conditions, not including 7 subjects who admitted to using external help. As in Exp. 2, the base payment was $2.50, and subjects could earn a bonus of up to $0.64 if they were randomly assigned to the low incentive condition, or $10.00 in the high incentive condition.

Methods and Design The methods for Exp. 3 are identical to those of Exp. 2, with the conditions tested being a cross of two rule types (II and IV) and two incentive levels (low and high). However, the incentive was not represented as a number of marbles in a bag but instead as a magnitude value of tickets that participants would earn for their correct answers above chance on the test. Since random/chance performance would generally get four of eight answers correct, subjects earned one ticket if their test score was five, two tickets if their test score was six, three tickets if their test score was seven, and four tickets if their test score was eight. In the low incentive condition, tickets were worth $0.16 each - the expected value of the 1.6% chance at $10 that correct answers above chance were worth in Exp. 1 and 2. Subjects assigned to the low incentive condition could earn a maximum bonus of $0.64. In the high incentive condition, tickets were worth $2.50 each, allowing a maximum bonus of $10.00.

Results Even though there was a direct, certain relationship between participants' performance in the task and the magni-

tude of the bonus they would receive, performance did not differ between the low and high incentive groups as shown in the bottom left plot in Figure 4. The logistic regression model results showed the average effect of incentive on performance to be nearly zero with a mean of -.001, $SD = .001$, $HDI = [-.003, .001]$. As in Exp. 2, the small increase in response times in the high incentive condition for Rule Type IV was not robust (bottom left, Fig. 5). The model fit on the response time data confirmed the lack of effect of incentive on response time, $HDI = [-9.0, 35.4]$.

Experiment 4 - Within-Subjects Design

Previously, we had avoided any within-subjects designs in order to prevent the learning effects that could come with performing multiple rounds of the category learning task. However, we were also concerned that modulating incentive between subjects possibly increased variance due to between subject differences in sensitivity to small magnitude payments. If we were able to manipulate incentive within subject, a single participant would be able to see that incentive as a signal to devote more effort and attention to the more valuable portions of the experiment. The experiment that we used as a basis for our design, which had found effects of incentive on attention and effort, also manipulated incentive within-subject (Caplin et al., 2020). As a result, we expected that if the SHJ task was sensitive to a voluntary change in cognitive approach, this experiment would finally reveal such an effect.

Participants Exp. 4 had 31 subjects, each of whom completed four game blocks. This total does not include 5 subjects who admitted to using external help. The study took about 30 minutes to complete and participants received a base rate payment of $4.50 for their time. Participants were eligible to earn up to a $10.00 bonus based on their performance.

Methods and Design In Exp. 4, both rule type and incentive were varied within subject. The four game blocks each had a unique rule type and incentive pairing, crossing rule types II and IV with a low and high incentive. As in Exp. 2 and Exp. 3, there were 32 learning trials in each game. Performance was incentivized by increasing magnitude rather than probability of reward, as in Exp. 3.

Participants took part in four consecutive games. In two of the games, participants could earn "blue bonus tickets" worth $0.02 each. In the other two games, participants could earn "gold bonus tickets" worth $1.23 each. Each game followed the procedure of the task in Exp. 3, such that participants could earn up to four tickets in each game depending on their performance in the test phase. This corresponded to maximal earnings of $0.08 on the low incentive "blue ticket" games, and $4.92 on the high incentive "gold ticket" games. Correct answers in high incentive games were therefore over sixty times more valuable than those in low incentive games. The order of the games was randomized for each subject.

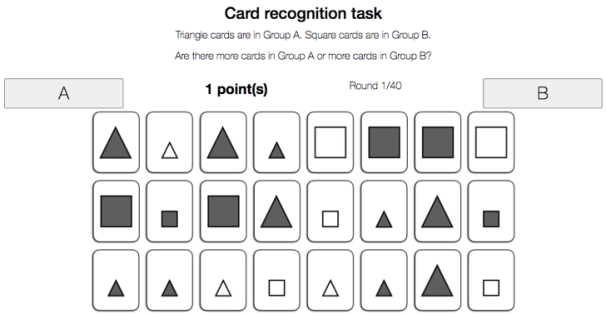

Figure 6: The computer display on each trial of the card category recognition task of Experiment 5. The relevant rule is shown at the top, as well as the number of points the participant will earn for a correct answer.

Results Although subjects now received more directly comparable signals of the relative value of each trial, we were surprised to see yet again no effect of the incentive manipulation on performance (bottom right, Fig. 4). The null result was substantiated by the model fit on the data, showing no effect of incentive on performance, $HDI = [-.044, .040]$. If subjects did apply more effort only on higher value portions of the experiment, this effort was not visible in their performance. There was a marginal increase in response time in the high incentive condition (bottom right, Fig. 5), with the model parameter estimate overlapping zero, $HDI = [-2109.0, 1282.2]$.

Experiment 5 - No Rule Discovery, Just Rule Following

After this streak of null results for incentive (with multiple robust replications of the original categorization effects from SHJ), we wanted to restrict our design even further to be as similar as possible to studies that found effects of incentive on performance, such as Caplin et al. (2020). Our results so far suggested that rule inference and category learning are not influenced by incentive, in spite of many experiment variations. If we removed the requirement for subjects to execute the complex strategy of inductive inference, would we then see an effect of incentive on performance?

Therefore, we conducted a task that relied on simply *using* a provided rule to categorize cards and determine whether there were more instances of cards in Group A or cards in Group B on the screen on a trial. The task would be more difficult for more complex rules, so we expected to see an effect of rule type on performance as in the previous experiments. However, the relevant rule would be provided to participants, removing their reliance on inference and memory.

Oprea (2020) investigates what characteristics make rules complex, like the number of states a rule requires and the amount of redundancy it permits. However, the implementation context is also a large factor in the complexity of a rule. Oprea (2020) emphasizes that implementing rules mentally is much more costly than implementing a rule physically; that

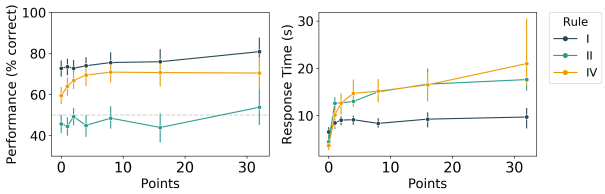

Figure 7: Left: Exp. 5 mean performance by incentive condition - the number of points a trial is worth - for the three rule types tested. A gray dashed horizontal line at 50% shows expected chance performance. Right: Exp. 5 median response times by incentive for the three rule types tested.

is, making an inference is significantly more challenging than executing a strategy. The complexity of the rule will still affect task difficulty, but this version of the task removes the requirement to learn or discover a categorization rule.

Participants Exp. 5 collected data from 200 participants across three rule type conditions. This total did not include 10 subjects excluded due to admitting to using outside help.

Methods The experiment was designed very similarly to the polygon identification task in Caplin et al. (2020). Subjects performed 40 trials in which their task was to identify whether there were more Group A cards or Group B cards on the screen (Fig. 6). Before each trial, the text describing the rule was shown, and this text was also at the top of the screen during each trial. Each trial showed 24 cards that were sampled randomly from the eight stimuli such that 11 cards were in Group A and 13 cards were in Group B or vice versa. Thus the difference between the card group counts was kept constant at 2 for every trial.

Participants could earn up to 200 points total across all trials, with 2 trials offering 32 points for a correct answer, 3 trials offering 16 points, 5 trials offering 8 points, 6 trials offering 4 points, 8 trials offering 2 points, 8 trials offering 1 point, and 8 trials offering 0 points. This distribution implemented the design of Caplin et al. (2020) to examine a spread of incentive values within each participant. Subjects were instructed that their chance at winning a $10.00 bonus would be their final score minus 100, since chance performance would correspond with a score of 100 points. For example, if a subject scored 165 points, they had a 65% chance at winning the bonus. As in Caplin et al. (2020), subjects were not told how many points they had earned until the end of the experiment.

Rule type was varied between subject, with each trial generating a randomized instance of that rule type so that the exact rule was not repeated across trials. For instance, a participant assigned to the Rule Type I condition might see a "Small vs. large" rule in one trial and "Triangle vs. square" rule in the next.

Results Figure 7 shows performance increasing with the point value of the trials. This effect is largest for rule type 4 and smaller for Rule Types I and II. Our interpretation of this result is that Rule Type I is so easy that marginal effort generates suitable performance, and Rule Type II is so difficult that effort only has a small impact on performance. As with the other experiments, we fit a logistic model to the performance data and a separate model to the response time data. In the model predicting performance, the incentive parameter estimation was greater than zero, with a mean of .002 ($SD = .001$), $HDI = [.001, .004]$. From this result we conclude that incentive has a positive effect on performance. Compared to Rule Type I, the Rule Type II condition had a substantial negative impact on performance with an average of -.274 ($SD = .013$),$HDI = [-.301, -.253]$; meanwhile Rule Type IV was associated with a milder average decrease in performance of 0.08 ($SD = .013$),$HDI = [-.102, -.055]$. The finding that Rule Type II is more challenging than Rule Type IV opposes traditional results in learning contexts that Type IV problems are more difficult than Type II.

In Exp. 5, we found a substantial effect on trial value on response time (Fig. 7). The incentive parameter in the logistic regression model fit on response time had an average value of 297.1 ($SD = 52$), $HDI = [202.1, 398.0]$. Compared to the Rule Type I condition, participants spent longer on each trial in the Rule Type II condition ($\bar{x} = 5554.2$, $SD = 976.7$, $HDI = [3614.0, 7266.4]$) and Rule Type IV condition ($\bar{x} = 9102.4$, $SD = 991.6$, $HDI = [7272.3, 10938.7]$). However, the more difficult Type II condition still retains near-chance performance even when participants spend up to ten seconds longer deliberating on the trial.

Discussion

We find no evidence that category learning is modulated by incentive, suggesting that rule discovery is not a cognitive process that can be voluntarily manipulated by humans. However, a task that removes the requirement of rule discovery and instead demands only rule implementation does show sensitivity to incentives. As discussed briefly in the Exp. 1 results, participants did not follow consistent trends in the errors they made across conditions. However, the experiment was not designed to specifically investigate errors in category beliefs and it is possible participants do err on the side of simplicity when formulating a rule to apply to the stimuli. Further work may benefit from using transfer stimuli - e.g., stimuli not seen before - to assess the specific category beliefs that participants acquired during training.

The results from our five experiments shed light on different contexts in which human behavior changes with financial incentives. One possible explanation of these results are that strategies that rely on simple cognitive routines that can be easily compiled and repeated are mutable, and can be voluntarily controlled in the face of increasing incentives. Meanwhile, humans are unable to improve the performance of complex strategies such as rule induction that involve a large space of possible hypothesis testing and learning steps and must be implemented anew each time.

References

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15–21.

Caplin, A., Csaba, D., Leahy, J., & Nov, O. (2020). Rational inattention, competitive supply, and psychometrics. *The Quarterly Journal of Economics*, *135*(3), 1681–1724.

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. A. (2020). *Bambi: A simple interface for fitting bayesian linear models in python.*

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, *8*(3), e57410.

DellaVigna, S., & Pope, D. (2018). What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, *85*(2), 1029–1069.

Enke, B., Gneezy, U., Hall, B., Martin, D. C., Nelidov, V., Offerman, T., & van de Ven, J. (2021). *Cognitive biases: Mistakes or missing stakes?* (Tech. Rep.). National Bureau of Economic Research.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological review*, *111*(2), 309.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*(1), 53.

Oprea, R. (2020). What makes a rule complex? *American economic review*, *110*(12), 3913–51.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience*, *40*, 99–124.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

van den Berg, R., Zou, Q., & Ma, W. J. (2020). No effect of monetary reward in a visual working memory task. *bioRxiv*, 767343.