Cognitive Biases: Mistakes or Missing Stakes?*

Benjamin Enke Uri Gneezy Brian Hall David Martin
Vadim Nelidov Theo Offerman Jeroen van de Ven

March 9, 2021

Abstract

Despite decades of research on heuristics and biases, empirical evidence on the effect of large incentives – as present in relevant economic decisions – on cognitive biases is scant. This paper tests the effect of incentives on four widely documented biases: base rate neglect, anchoring, failure of contingent thinking, and intuitive reasoning in the Cognitive Reflection Test. In laboratory experiments with 1,236 college students in Nairobi, we implement three incentive levels: no incentives, standard lab payments, and very high incentives that increase the stakes by a factor of 100 to more than a monthly income. We find that response times – a proxy for cognitive effort – increase by 40% with very high stakes. Performance, on the other hand, improves very mildly or not at all as incentives increase, with the largest improvements due to a reduced reliance on intuitions. In none of the tasks are very high stakes sufficient to de-bias participants, or come even close to doing so.

JEL classification: D01

Keywords: Cognitive biases, incentives

*We thank three very constructive referees, Thomas Graeber and Florian Zimmermann for helpful comments. The experiments in this paper received IRB approval from Harvard's IRB. This study was funded using Hall's research funds from Harvard Business School. For excellent research assistance we are grateful to Tiffany Chang, Davis Heniford, and Karim Sameh. We are also grateful to the staff at the Busara Center for Behavioral Economics for dedicated support in implementing the experiments. Corresponding author: Enke: Harvard University and NBER; enke@fas.harvard.edu. Gneezy is at UC San Diego Rady, Hall at Harvard Business School, Martin at Harvard University, and Nelidov, Offerman, and van de Ven at the University of Amsterdam.

1 Introduction

Starting with Tversky and Kahneman (1974), the "heuristics and biases" program has occupied psychologists and behavioral economists for nearly half a century. In a nutshell, this voluminous and influential line of work has documented the existence and robustness of a large number of systematic errors – "cognitive biases" – in decision-making.

In studying these biases, psychologists often use hypothetical scenarios. Experimental economists criticize the lack of incentives, and use payments that amount to a couple of hours of wages for the students participating in order to motivate them to put effort into the task. Yet, non-experimental economists often raise concerns in response to findings based on such incentives, arguing that people will exert more effort in high-powered decisions, so that cognitive biases may be irrelevant for understanding real-world behavior. In other words, just like experimental economists criticize psychologists for not incentivizing at all, non-experimental economists often criticize experimental economists for using fairly small incentives. As Thaler (1986) states in his discussion of ways in which economists dismiss experimental findings: "If the stakes are large enough, people will get it right. This comment is usually offered as a rebuttal... but is also, of course, an empirical question. Do people tend to make better decisions when the stakes are high?"

This empirical question is relevant for two reasons. First, as noted by Thaler, a relevant issue is to understand whether systematic departures from the rational economic model are likely to appear only in the many small-stakes decisions that we make, or also in decisions with high-powered incentives and large financial implications. Such understanding can inform our modeling of important real-life decisions. Second, it is of interest to understand the mechanisms behind cognitive biases. For example, a very active recent theoretical and experimental literature attempts to identify the extent to which different biases are generated by micro-foundations such as incorrect mental models, memory imperfections, or limited attention, where low effort often features as one of the prime candidates.

Of course, documenting the relevance of the heuristics and biases program for high-powered real economic decisions has been on behavioral economists' to-do list for almost 40 years, and the empirical literature indeed contains many demonstrations that behavioral insights matter under high incentives (see references below). At the same time, perhaps somewhat surprisingly, systematic empirical evidence that carefully compares the presence of cognitive biases under small and very large incentives is scant.

The current paper targets this gap in the literature. We conduct systematic tests of the effects of incentive size, and in particular the effects of very large incentives, on four well-documented biases that are frequently studied by behavioral economists. Our design has three pay levels: no incentives, relatively small incentives that amount to standard laboratory pay, and very high incentives that are 100 times larger than the standard stake size and equivalent to more than one month's income for our participants.

We apply these stake-size variations to the following well-established biases: base rate neglect (BRN); anchoring; failure of contingent thinking in the Wason selection task; and intuitive reasoning in the Cognitive Reflection Test (CRT). Our interest in this paper is not so much in these biases per se, but rather in the effects of varying the stake size. We therefore selected these particular biases subject to the following criteria: (i) the tasks that underlie these biases have an objectively correct answer; (ii) the biases are cognitive in nature, rather than preference-based; (iii) standard experimental instructions to measure these biases are short and simple, which helps rule out confusion resulting from complex instructions; and (iv) these biases all have received much attention and ample experimental scrutiny in the literature.[1] An added benefit of including the CRT in our set of tasks is that it allows us to gauge the role of intuitions in generating cognitive biases: if it were true that higher stakes and effort reduced biases in the CRT but not

[1]Base rate neglect is one of the most prominent and widely studied biases in belief updating (Grether, 1980, 1992; Camerer, 1987; Benjamin, 2019). Anchoring has likewise received much attention, with widely cited papers such as Ariely et al. (2003); Epley and Gilovich (2006); Chapman and Johnson (2002). Contingent reasoning has been studied in the psychology of judgment for many decades (e.g., Bazerman and Samuelson, 1983; Johnson-Laird, 1983; Cheng and Holyoak, 1985; Cosmides, 1989) and is a very active subject of study in the current literature, as it appears to manifest in different errors in statistical reasoning (Esponda and Vespa, 2014, 2016; Enke, 2020; Enke and Zimmermann, 2019; Martínez-Marquina et al., 2019). Finally, intuitive reasoning in the CRT is a widely implemented cognitive test in behavioral economics, at least partly because it is strongly correlated with many behavioral anomalies from the heuristics and biases program (Frederick, 2005; Hoppe and Kusterer, 2011; Toplak et al., 2011; Oechssler et al., 2009).

otherwise, then other biases are less likely to be primarily generated by intuitions and a lack of deliberative thinking.

Because there is a discussion in the literature about the frequency of cognitive biases in abstractly vs. intuitively framed problems (Cheng and Holyoak, 1985; Gigerenzer and Hoffrage, 1995), we implement two cognitive tasks (base rate neglect and the Wason selection task aimed at studying contingent thinking) in both a relatively abstract and a relatively intuitive frame. Entirely abstract frames present only the elements of a problem that are necessary to solve it, without further context. More intuitive frames present a problem with a context intended to help people to relate it to their daily life experiences. In total, we implement our three incentive conditions with six types of tasks: abstract base rate neglect, intuitive base rate neglect, anchoring, abstract Wason selection task, intuitive Wason selection task, and the CRT.

We run our experiments with a total of $N = 1,236$ college students in the Busara Center for Behavioral Economics in Nairobi, Kenya. We selected this lab to run our experiments because of the lab's ability to recruit a large number of analytically capable students for whom our large-stakes treatment is equal to more than a month's worth of income. Participants are recruited among students of the University of Nairobi, the largest and most prestigious public university in Kenya. While this sample is different from the types of samples that are typically studied in laboratory experiments, the average CRT scores of these participants are similar to those reported in a large meta-study with predominantly U.S. and European-based populations (Brañas-Garza et al., 2019).

The focus of our paper is the comparison between high stakes and standard stakes. At the same time, ideally we would also like to gather meaningful information on participants' behavior without any financial incentives. To achieve this objective while at the same time maintain high statistical power with a given budget, we implemented three payment levels (no, standard and high stakes) but only two randomized treatment conditions. In the first part of the experiment, each subject completes the questions for a randomly selected bias without any incentives. Then, the possibility of earning a bonus in the second part of the experiment is mentioned. In this second part, subjects are randomized into high or standard incentives for a cognitive bias that is different from the one in the first part. Thus, treatment assignment between standard and high stakes is random, yet we still have a meaningful benchmark for behavior without incentives from the first part of the experiment.

In the two financially incentivized conditions, the maximum bonus is 130 KSh (\$1.30) and 13,000 KSh (\$130). Median monthly income and consumption in our sample are in the range of 10,000–12,000 KSh, so that the high stakes condition offers a bonus of more than 100% of monthly income and consumption. As a second point of comparison, note that our standard and high incentive levels correspond to about \$23.50 and \$2,350 at purchasing power parity in the United States. We chose experimental procedures that make these incentive payments both salient and credible. We deliberately selected the Busara lab for implementation of our experiments because the lab follows a strict no-deception rule. In addition, both the written and the oral instructions highlight that all information that is provided in the experimental instructions is true and that all consequences of subjects' actions will happen as described. Finally, the computer screen that immediately precedes the main decision tasks reminds subjects of the possibility of earning a given bonus size.

We find that, across all of our six tasks, response times – our proxy for cognitive effort – are virtually identical with no incentives and standard lab incentives. On the other hand, response times increase by about 40% in the very high incentive condition, and this increase is similar across all tasks. Thus, there appears to be a significant effect of incentives on cognitive effort that could in principle translate into substantial reductions in the frequency of observed biases.

There are at least two *ex ante* plausible hypotheses about the effect of financial incentives on biases. A first is that cognitive biases are largely driven by low motivation, so that the increase in effort that we observe should go a long way towards debiasing people. An alternative hypothesis is that cognitive biases reflect the high difficulty of rational reasoning and / or high cognitive effort costs, so that even very large incentives will not dramatically improve performance.

Looking at the frequency of biases across incentive levels, our headline result is that cognitive biases are largely, and almost always entirely, unresponsive to stakes. In five out of our six tasks, the frequency of errors is statistically indistinguishable between standard and very large incentives, and in five tasks it is statistically indistinguishable between standard and no incentives. Given our large sample size, these "null results" are relatively precisely estimated: across the different tasks, we can statistically rule out performance increases of more than 3–18 percentage points (based on 95% confidence intervals). In none of the tasks did cognitive biases disappear, and even with very large incentives the error rates range between 40% and 90%. We further document that high incentives generally do not reduce the frequency of specific well-known decision heuristics.

The only task in which very large incentives produce statistically significant performance improvements is the CRT. We also find some mildly suggestive evidence that stakes matter more in the intuitive versions of base rate neglect and the Wason task. A plausible interpretation of these patterns is that increased incentives reduce reliance on intuitions, yet some problems are sufficiently complex for people that the binding constraint is not low effort and reliance on intuitions but instead a lack of conceptual problem solving skills. Our correlational follow-up analyses are in line with such an interpretation: the within-treatment correlations between cognitive effort and performance are always very small, suggesting that it is not only effort per se but at least partially "the right way of looking at a problem" that matters for cognitive biases. In addition, participants appear to exhibit some awareness that increased effort does not necessarily translate into better performance: in non-incentivized confidence questions at the end of the experiment, participants indicated almost identical levels of confidence across treatment conditions.

Our results contrast with the predictions of a sample of 68 researchers, drawn from professional experimental economists and Harvard students with exposure to graduate-level experimental economics. These researchers predict that performance will improve by an average of 25% going from no incentives to standard incentives, and by another 25% going from standard to very high incentives. While there is some variation in projected performance increases across tasks, these predictions are always more bullish about the effect of incentives than our experimental data warrant.

Our paper ties into the large lab experimental literature that has investigated the role of the stake size for various types of economic decisions. In contrast to our focus on very high stakes, prior work on cognitive biases has considered the difference between no and "standard" (small) incentives, or between very small and small incentives. Early experimental economists made a point of implementing financially incentivized designs to replicate biases from the psychological literature that were previously studied using hypothetical questions (e.g., Grether and Plott, 1979; Grether, 1980). In Appendix A, we review papers that have studied the effect of (no vs. small) incentives in the tasks that we implement here; while the results are a bit mixed, the bottom line is that introducing small incentives generally did not affect the presence of biases. Indeed, more generally, in an early survey of the literature, Camerer and Hogarth (1999) conclude that "...no replicated study has made rationality violations disappear purely by raising incentives." Yet despite the insights generated by this literature, it remains an open question whether very large stakes – as present in many economically relevant decisions – eliminate or significantly reduce biases.

Investigating the effect of very large stakes on biases appears relevant also in light of literatures that show that behavior in preferences-based tasks or strategic games often dramatically changes in the presence of higher stakes (Binswanger, 1980; Holt and Laury, 2002). For example, high-stakes behavior in the ultimatum game reverts back close to predictions based on selfishness and rationality (Slonim and Roth, 1998; Cameron, 1999; Andersen et al., 2011). Likewise, a literature in experimental game theory highlights that raising the stakes often significantly increases the fraction of equilibrium play (Smith and Walker, 1993; Cooper et al., 1999; Rapoport et al., 2003; Parravano and Poulsen, 2015). Ariely et al. (2009) study the effect of large incentives on "choking under pressure" in creativity, motor skill and memory tasks such as fitting pieces into frames, throwing darts, or memorizing sequences. An important difference between our paper and theirs is that we focus on established tasks aimed at measuring cognitive biases. In summary, existing experimental work on stake size variations has either compared no (or very small) with "standard" incentives, or it has studied high-stakes behavior in tasks and games that do not measure cognitive biases.[2]

Finally, a related literature investigates the effects of incentives on students' performance on standardized tests and academic test performance. We review this literature in detail in Appendix A, also see Gneezy et al. (2011) for an early review. In general, this literature reports mixed or small positive results of explicit financial incentives on test performance (e.g. Fryer Jr, 2011; Bettinger, 2012; Levitt et al., 2016; O'Neil et al., 1995, 2005; Baumert and Demmrich, 2001). In cases where the literature does identify positive effects on performance, the effect sizes correspond to performance improvements of about 0.10–0.15 standard deviations (Levitt et al., 2016; Bettinger, 2012). While it is difficult to directly compare this effect size to our study given differences in participant pools, size of incentives and local purchasing power, a useful comparison may be that, in our CRT task (the only task in which we observe a significant improvement in performance), the score improves about 0.20 standard deviations when going from standard stakes to very high stakes.

[2]Non-experimental work on behavior under high incentives includes a line of work on game shows (Metrick, 1995; Berk et al., 1996; Levitt, 2004; Belot et al., 2010; Van den Assem et al., 2012) and a line of work on biases in real market environments (e.g., Beggs and Graddy, 2009; Pope and Schweitzer, 2011; Graddy et al., 2014; Chen et al., 2016; Jetter and Walker, 2017). These studies generally document the existence of cognitive biases under high incentives, but do not make a careful comparison between standard and high incentives.

2 Experimental Design and Procedures

2.1 Tasks

2.1.1 Base Rate Neglect

A large number of studies document departures from Bayesian updating. A prominent finding is that base rates are ignored or underweighted in making inferences (Kahneman and Tversky, 1973; Grether, 1980; Camerer, 1987).

In our experiments, we use two different questions about base rates: the well-known "mammography" and "car accident" problems (Gigerenzer and Hoffrage, 1995). Motivated by a long literature that has argued that people find information about base rates more intuitive when it is presented in a frequentist rather than probabilistic format, we implement both probabilistic ("abstract") and frequentist ("intuitive") versions of each problem. The wording of the abstract and intuitive versions of the mammography problem is presented below, with the wording of the conceptually analogous car accidents problems provided in Appendix B.

> ***Abstract mammography problem:*** *1% of women screened at age 40 have breast cancer. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will get a positive mammography. A 40-year-old woman had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?*

In the abstract version of the mammography problem, participants are asked to provide a scalar probability. The Bayesian posterior is approximately 7.8 percent, yet research has consistently shown that people's subjective probabilities are too high, consistent with neglecting the low base rate of having cancer. The intuitive version of the base rate neglect task adds a figure to illustrate the task to subjects, and only works with frequencies.

Intuitive mammography problem: *10 out of every 1,000 women at age 40 who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will get a positive mammography. A diagram presenting this information is below:*

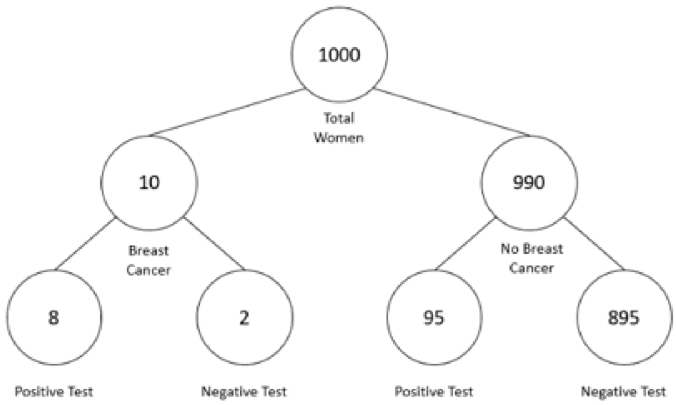


Figure 1: Diagram used to illustrate the intuitive mammography base rate neglect task

In a new representative sample of 100 women at age 40 who got a positive mammography in routine screening, how many women do you expect to actually have breast cancer?

Subjects who complete the base rate neglect portion of our study (see below for details on randomization) work on two of the four problems described above. Each participant completes one abstract and one intuitive problem, and one mammograpy and one car accidents problem. We randomize which format (abstract or intuitive) is presented first, and which problem is presented in the intuitive and which one in the abstract frame.

For each problem, participants can earn a fixed sum of money (that varies across treatments) if their guess g is within $g \in [x-2, x+2]$ for a Bayesian response x. To keep the procedures as simple as possible, the instructions explain that subjects will be rewarded relative to an expert forecast that relies on the same information as they have. We implement a binary "all-or-nothing" payment rule rather than a more complex, continuous scoring rule such as the binarized scoring rule both to keep the payout procedures similar to the other tasks, and because of recent evidence that subjects appear to understand simpler scoring rules better (Danz et al., 2019).

2.1.2 Contingent Reasoning: The Wason Selection Task

Contingent reasoning has been studied in the psychology of judgment for many decades (e.g., Bazerman and Samuelson, 1983; Johnson-Laird, 1983; Cheng and Holyoak, 1985; Cosmides, 1989) and more recently in behavioral economics (e.g. Giglio and Shue, 2014; Esponda and Vespa, 2016; Enke, 2020; Barron et al., 2019). While the experimental tasks in this literature differ across studies depending on the specific design objective, they all share in common the need to think about hypothetical contingencies. The Wason selection task is a well-known and particularly simple test of such contingent reasoning.

In this task, a participant is presented with four cards and a rule of the form "if P then Q." Each card has information on both sides – one side has either "P" or "not P" and the other side has either "Q" or "not Q" – but only one side is visible. Participants are asked to find out if the cards violate the rule by turning over some cards. Not all cards are helpful in finding possible violations of the rule, and participants are instructed to turn over only those cards that are helpful in determining whether the rule holds true. Common mistakes are to turn over cards with "Q" on the visible side or to not turn over cards with "not Q" on the visible side.

We implement two versions of this task. One version is relatively abstract, and people tend to perform poorly on it. The other version provides a more familiar context and is more intuitive. As a result, people tend to perform better.

> ***Abstract Wason selection task:*** *Suppose you have a friend who says he has a special deck of cards. His special deck of cards all have numbers (odd or even) on one side and colors (brown or green) on the other side. Suppose that the 4 cards from his deck are shown below. Your friend also claims that in his special deck of cards, even numbered cards are never brown on the other side. He says:*

"In my deck of cards, all of the cards with an even number on one side are green on the other."

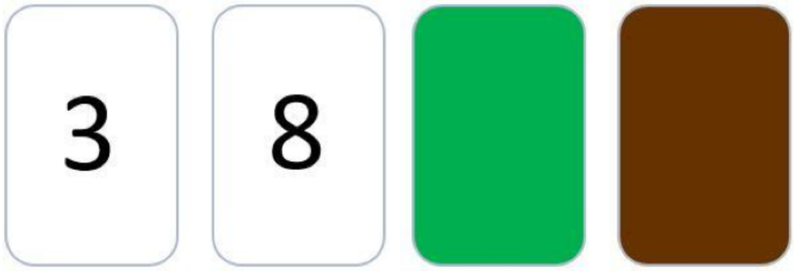


Figure 2: Abstract Wason task

Unfortunately, your friend doesn't always tell the truth, and your job is to figure out whether he is telling the truth or lying about his statement. From the cards below, turn over only those card(s) that can be helpful in determining whether your friend is telling the truth or lying. Do not turn over those cards that cannot help you in determining whether he is telling the truth or lying. Select the card(s) you want to turn over.

The correct actions are turning over the "8" and "brown" cards.

Intuitive Wason selection task: *You are in charge of enforcing alcohol laws at a bar. You will lose your job unless you enforce the following rule: If a person drinks an alcoholic drink, then they must be at least 18 years old. The cards below have information about four people sitting at a table in your bar. Each card represents one person. One side of a card tells what a person is drinking, and the other side of the card tells that person's age. In order to enforce the law, which of the card(s) below would you definitely need to turn over? Indicate only those card(s) you definitely need to turn over to see if any of these people are breaking the law.*

Select the card(s) you want to turn over.

In this "social contract" version (adapted from Cosmides, 1989), the correct actions are turning over the "Beer" and "16" cards. While this problem is logically the same as the

Figure 3: Intuitive Wason task

abstract version, this version may prompt "cheater detection" and may help participants to actually identify the correct solution more often.

In our experiments, each subject in the Wason condition completes both of these tasks in randomized order. For each task, subjects can win a fixed sum of money (that varies across treatments) if they turn over (only) the two correct cards.

2.1.3 Cognitive Reflection Test

The CRT measures people's tendency to engage in reflective thinking (Frederick, 2005). The test items have an intuitive, incorrect answer, and a correct answer that requires effortful deliberation. Research has shown that people often settle on the answer that is intuitive but wrong. We include the following two questions, both of which are widely used in the literature:

1. *A bat and a ball cost 110 KSh in total. The bat costs 100 KSh more than the ball. How much does the ball cost?* (Intuitive answer is 10, correct answer is 5).
2. *It takes 5 nurses 5 minutes to measure the blood pressure of 5 patients. How long would it take 10 nurses to measure the blood pressure of 10 patients?* (Intuitive answer is 10, correct answer is 5).

Subjects in the CRT condition complete both of these questions in randomized order. For each question, they can earn a fixed sum of money (that varies across treatments) if they provide exactly the correct response.

2.1.4 Anchoring

People have a tendency to use irrelevant information in making judgments. Substantial research has shown that arbitrary initial information can become a starting point ("anchor") for subsequent decisions, with only partial adjustment (Tversky and Kahneman, 1974). This can have consequential effects in situations such as negotiations, real estate appraisals, valuations of goods, or forecasts.[3]

To test for anchoring, we follow others in making use of a random anchor, since only an obviously random number is genuinely uninformative. To generate a random anchor, we ask participants for the last digit of their phone number. If this number is four or lower, we ask them to enter the first two digits of their year of birth into the computer, and otherwise to enter 100 minus the first two digits of their year of birth. Given that all participants were either born in the 1900s or 2000s, this procedure creates either a low anchor (19 or 20) or a high anchor (80 or 81). The experimental instructions clarify that "...you will be asked to make estimates. Each time, you will be asked to assess whether you think the quantity is greater than or less than the two digits that were just generated from your year of birth." Given these experimental procedures, the difference in anchors across subjects is transparently random.

After creating the anchor, we ask participants to solve estimation tasks as described below. Following standard procedures, in each task, we first ask subjects whether their estimate is below or above the anchor. We then ask participants to provide their exact estimate. An example sequence of questions is:

A1 *Is the time (in minutes) it takes for light to travel from the Sun to the planet Jupiter more than or less than [anchor] minutes?*

A2 *How many minutes does it take light to travel from the Sun to the planet Jupiter?*

[3] The evidence for anchoring effects in valuation tasks is mixed (Ariely et al., 2003; Fudenberg et al., 2012; Maniadis et al., 2014; Simonsohn et al., 2014; Yoon et al., 2019). Ioannidis et al. (2020) present a meta-analysis that shows that anchoring has an effect when anchors are informative or perceived to be informative. When anchors are uninformative, they find a null-effect of anchoring on valuations.

where [anchor] is replaced with the random number that is generated from a participant's phone number and year of birth. We also use three other sets of questions:

B1 *In 1911, pilot Calbraith Perry Rodgers completed the first airplane trip across the continental U.S., taking off from Long Island, New York and landing in Pasadena, California. Did the trip take more than or less than [anchor] days?*

B2 *How many days did it take Rodgers to complete the trip?*

C1 *Is the population of Uzbekistan as of 2018 greater than or less than [anchor] million?*

C2 *What is the population of Uzbekistan in millions of people as of 2018?*

D1 *Is the weight (in hundreds of tons) of the Eiffel Tower's metal structure more than or less than [anchor] hundred tons?*

D2 *What is the weight (in hundreds of tons) of the Eiffel Tower's metal structure?*

Each of these problems has a correct solution that lies between 0 and 100. Subjects are told that they can only state estimates between 0 and 100. Each participant who takes part in the anchoring condition of our experiment completes two randomly selected questions from the set above, in randomized order. For each question, participants can earn a fixed sum of money (that varies across treatments) if their guess g is within $g \in [x-2, x+2]$ for a correct response x.

2.2 Incentives and Treatment Conditions

Incentive Levels. In order to offer very high incentives and still obtain a large sample size within a feasible budget, we conduct the experiment in a low-income country: at the

Busara Lab for Behavioral Economics in Nairobi, Kenya. For each bias, there are three possible levels of incentives: a flat payment (no incentives), standard lab incentives, and high incentives. With standard lab incentives, participants can earn a bonus of 130 KSh (approx. 1.30 USD) for a correct answer. In the high incentive treatment, the size of the bonus is multiplied by a factor of 100 to equal 13,000 KSh (approx. 130 USD).

These incentives should be compared to local living standards. Kenya's GDP per capita at purchasing power parity (PPP) in 2018 was $3,468, which is 18 times lower than that of the United States. Our standard lab incentives of 130 KSh correspond to about $23.50 at PPP in the U.S. Our high incentive condition corresponds to a potential bonus of $2,350 at PPP in the U.S.

As a second point of comparison, we ask our student participants to provide information on their monthly consumption and their monthly income in a post-experiment survey. The median participant reports spending 10,000 KSh (approx. 100 USD) and earning income of 12,000 KSh (approx. 120 USD) per month. Thus, the bonus in our high incentive condition corresponds to 130% of median consumption and 108% of median income in our sample.

Treatments. In principle, our experiment requires three treatment conditions. However, because our primary interest is in the comparison between the standard incentive and the high incentive conditions, we elected to implement only two treatment conditions to increase statistical power.

The main experiment consists of two parts. Each participant is randomly assigned two of the four biases. In Part 1, all participants work on tasks for the first bias in the flat payment condition. Thus, they cannot earn a bonus in Part 1. In Part 2, they are randomly assigned to either standard lab incentives or high incentives and complete tasks for the second bias. Participants only receive instructions for Part 2 after completing Part 1, and the possibility of a bonus is never mentioned until Part 2.

With this setup, we have twice as many observations in the flat payment condition ($N = 1,236$) as in the standard lab incentive ($N = 636$) and high incentive ($N = 600$)

conditions. We keep the order of treatments constant (flat payments always followed by standard lab incentives or high incentives), so that participants working under the flat payment scheme are not influenced by the size of incentives in the first question.

Readers may be concerned that the comparison between the flat payment condition and the financially incentivized conditions is confounded by order effects. We deliberately accept this shortcoming. Formally, this means that a skeptical reader may only consider the treatment comparison between standard and high incentives valid, as this is based on randomization. Throughout the paper, we nonetheless compare the three incentive schemes side-by-side, with the implicit understanding that our main interest is in the comparison between standard and high incentives.

2.3 Procedures

Questions and randomization. In total, each participant works on two biases, where for each bias they answer two questions. Thus, each participant answers four questions in total: two in Part 1 (without any financial incentives) and two in Part 2 (with standard or high incentives).

As explained above, each bias consists of two questions. For some questions, we implement minor variations across experimental sessions to lower the risk that participants memorize the questions and spread knowledge outside the lab to other participants in the pool. For example, in the Wason tasks, we change the colors of the cards from green and brown to blue and brown. To take a different example, in the second CRT problem, we change the information from "It takes 5 nurses 5 minutes to measure the blood pressure of 5 patients" to "It takes 6 nurses 6 minutes to measure the blood pressure of 6 patients." Appendix B contains the full list of questions that we implement. We find no evidence that participants in later sessions perform better than those in earlier sessions.

Each participant completes two questions in the financially incentivized part of the experiment (Part 2). One of these two questions is randomly selected and a bonus is given for a correct answer to that question. As explained above, for the CRT and the

Wason selection task, a participant has to give exactly the correct answer to be eligible for a bonus. For base rate neglect and anchoring, the answer has to be within two of the correct answer. Appendix F contains screenshots of the full experiment, including experimental instructions and decision screens.

The stake size is randomized at the session level, mainly because the Busara Lab was worried about dissatisfaction resulting from participants comparing their payments to others in the same session. The set and order of the biases are randomized at the individual level. Within each bias, we also randomize the order of the two questions.

Salience and credibility of incentive levels. A key aspect of our design is that the stake size is both salient and credible. We take various measures in this regard. To make the stake size salient, the screen that introduces the second part of the experiment reads:

> *Part 2. We will ask you two questions on the upcoming screens. Please answer them to the best of your ability. Please remember that you will earn a guaranteed show-up fee of 450 KSh. While there was no opportunity to earn a bonus in the previous part, you will now have the opportunity to earn a bonus payment of X KSh if your answer is correct.*

where $X \in \{130; 13,000\}$. The sentence about the opportunity to earn a bonus was underlined and highlighted in red. The subsequent screen (which is the one that immediately precedes the first incentivized question) reads:

> *Remember, you will now have the opportunity to earn a bonus payment of X KSh if your answer is correct.*

To ensure credibility of the payments, we put in place three measures. First, we deliberately select the Busara lab for implementation of our experiments because the lab follows a strict no-deception rule. Second, the written instructions highlight that:

> *The study you are participating in today is being conducted by economists, and our professional standards don't allow us to deceive research subjects. Thus,*

whatever we tell you, whatever you will read in the instructions on your computer screen, and whatever you read in the paper instructions are all true. Everything will actually happen as we describe.

Third, the verbal instructions by Busara's staff likewise emphasize that all information that is provided by the experimental software is real.

Our experimental data afford two analyses to investigate whether the increase in incentives was actually salient and credible. First, the post-experimental survey included unincentivized questions that ask subjects to recall the possible bonus amounts in Parts 1 and 2 of the study. Figure 8 in Appendix D shows the distribution of responses. We see that 2/3 of participants remember *exactly* the correct bonus amount. Moreover, the distribution of responses exhibits a very clear shift across the three incentive schemes. This provides a first piece of evidence that the incentives were salient to subjects.[4] Second, as we will see below, observed response times increase significantly as the stake size increases. This indicates that the incentives were not just salient but also credible – if participants had not trusted the experimenters to actually deliver on their promises, participant effort arguably should not have increased.[5]

Timeline. Participants are told that the experiment will last approximately one hour, but have up to 100 minutes to complete it. This time limit was chosen based on pilots such that it would not provide a binding constraint to participants; indeed no participants use all of the allotted time. The timeline of the experiment is as follows: (i) electronic consent procedure; (ii) general instructions; (iii) two unincentivized questions in Part 1; (iv) screen announcing the possibility of earning a bonus in Part 2; (v) two

[4]Tables 15 and 16 in Appendix C show that our results are very similar when we restrict the sample to those tasks for which a subject recalls the incentive amount *exactly* correctly (64% of all data points).

[5]While less rigorous, it may also be helpful to provide anecdotal evidence on payment credibility. In general, Busara states that they "have deep ties to the community in terms of participants who have come many times, and in general there is a strong trust in our integrated payment systems." Likewise, the lab manager in charge of executing our particular experiments told us that "participants did not express doubt on earning or receiving the amounts." Instead, she recalls participants making statement such as: "Thank you so much! OMG! I am so excited." Finally, one of the authors (Hall) and one of the research assistants (Heniford) were present for most of the pilot part of the study. In their debrief with participants, none questioned whether the payments would be made

financially incentivized questions in Part 2; and (vi) a post-experimental questionnaire. Screenshots of each step are provided in Appendix F.

Earnings. Average earnings are 482 KSh in the standard incentive condition and 3,852 KSh in the high incentive condition. This includes a show-up fee of 450 KSh. Per the standard procedure of the Busara Lab, all payments are transferred electronically within 24 hours of participation.

2.4 Participants

The experimental sessions take place at the Busara Center for Behavioral Economics in Nairobi, Kenya. We conduct our experiments in this lab due to the lab's capabilities in administering experiments without deception as well as the lab's ability to recruit a large number of analytically capable students for whom our large incentive treatment is equal to approximately a month's worth of their consumption. Participants are recruited among students of the University of Nairobi, the largest public university in Kenya. Table 1 reports the resulting sample sizes by bias and incentive level.[6] In total, 1,236 participants completed the study between April and July 2019. The majority (93 percent) are between 18 and 24 years old (mean age 22) and 44 percent are female.

It may be helpful to compare the level of cognitive skills in our sample with that of more traditional subject pools. The two CRT questions in our study are part of the meta-study in Brañas-Garza et al. (2019). In the no incentive condition of our experiments at Busara, 34% of all CRT questions are answered correctly. In Brañas-Garza et al.'s meta-study involving 118 studies and almost 45,000 participants (91% of which were from the U.S. or Europe), the fraction of correct responses for these same two questions is

[6] Table 5 in Appendix C reports summary statistics for participant characteristics across treatments.

Table 1: Number of participants by bias and incentive level

	No incentives	Standard incentives	High incentives
Base rate neglect	309	159	150
Contingent reasoning	308	160	151
CRT	311	163	146
Anchoring	308	154	153
Total	1,236	636	600

36% and therefore very similar to what we see in our sample.[7] [8]

2.5 Pre-Registration

We pre-registered the design and target sample size on www.aspredicted.org at `https://aspredicted.org/blind.php?x=5jm93d`. The pre-registration specified an overall sample size of 1,140 participants, yet our final sample consists of 1,236 participants. We contracted with the Busara lab not for a total sample size, but for a total amount of money that would have to be spent. Thus, our target sample size was based on projections of how costly the experiment would be. Because it turned out that slightly fewer subjects earned the bonus than we had anticipated, there was still "money left" when we arrived at 1,140 participants. Busara gave us a choice between forfeiting the money and sampling additional participants, and – being efficiency-oriented economists – we elected to sample additional subjects to increase statistical power. Tables 9 and 10 in

[7]For Frederick's (2005) earlier review, only averages for the entire three-question module are available. The corresponding numbers are, inter alia, 73% at MIT; 54% at Princeton; 50% at CMU; 48% at Harvard; 37% in web-based studies; 28% at University of Michigan Dearborn; 26% at Michigan State; and 19% at Toledo University. Thus, according to these metrics, our subject pool has lower average performance scores than the most selective U.S. universities, but it compares favorably with participants from more typical U.S. schools.

[8]A second, and perhaps more heuristic, comparison is to follow Sandefur (2018), who recently devised a method to construct global learning metrics by linking regional and international standardized test scores (such as TIMSS). His data suggest that Kenya has some of the highest test scores in his sample of 14 African countries. He concludes that "the top-scoring African countries produce grade 6 scores that are roughly equivalent to grade 3 or 4 scores in some OECD countries." Of course, this comparison is only heuristic because (i) it pertains to primary school rather than college students and (ii) it ignores the (likely highly positive) selection of Kenyan students into the University of Nairobi. Indeed, the University of Nairobi is the most prestigious public university in Kenya and routinely ranks as the top university in the country and among the top universities in Africa. See, for example, `https://www.usnews.com/education/best-global-universities/africa?page=2`.

Appendix C replicate our main results on a sample of the first 1,140 participants only. The results are very similar.[9]

2.6 Predictions by Experimental Economists

To complement our pre-registration and to be able to compare our results with the profession's priors, we collect predictions for our experiments (Gneezy and Rustichini, 2000; DellaVigna and Pope, 2018). In this prediction exercise, we supply forecasters with average response times and average performance for each bias in the absence of incentives, using our own experimental data. Based on these data, we ask our respondents to predict response times and performance in the standard incentive and high incentive conditions. Thus, each respondent issues 24 predictions (six tasks times two treatments times two outcome variables). Respondents are incentivized in expectation: we paid $100 to the respondent who issued the set of predictions that turned out to be closest to the actual data. The survey can be accessed at `https://hbs.qualtrics.com/jfe/form/SV_bDVhtmyvlrNKc6N`.

Our total sample of 68 researchers comprises a mix of experimental economists and Harvard students with graduate-level exposure to experimental economics. First, we contacted 231 participants of a recent conference of the Economic Science Association (the professional body of experimental economists) via email. Out of these, 45 researchers volunteered to participate in our prediction survey, 41 of which self-identified with Experimental Economics as their primary research field in our survey. In addition, we contacted all students who had completed Enke's graduate experimental economics class at Harvard in 2017–2019, which produced 23 student volunteers. The predictions of professionals and Harvard students turn out to be similar, on average.[10] We hence pool

[9] In the pre-registration, we also speculated about potential results. Specifically, we predicted that increasing incentives would result in longer decision times, reflecting higher effort. We further predicted that this would increase success in tasks such as the CRT, anchoring, BRN (when presented intuitively in the format of frequencies) and Wason (when presented intuitively in a social context). We expected that increased effort would not reduce biases in the Wason Selection task (abstract formulation) and BRN (in the classic probability format).

[10] Professional experimental economists tend to predict slightly smaller increases in response times and performance as a function of stakes, but these differences are rarely statistically significant.

them for the purpose of all analyses below.[11]

3 Results

3.1 Summary Statistics on Frequency of Cognitive Biases

A prerequisite for our study to be meaningful is the presence of cognitive biases in our sample. This is indeed the case. In the CRT, 39% of responses are correct and about 50% of all answers correspond exactly to the well-known "intuitive" response.

In the abstract base rate neglect task, 11% of all responses are approximately correct (defined as within 5 percentage points of the Bayesian posterior); the corresponding number is 26% for the intuitive version. Across all base rate neglect tasks, we see that subjects' responses tend to be too high, effectively ignoring the low base rate.

In the Wason selection task, 14% of responses are correct in the abstract frame and 57% in the intuitive frame. This level difference is consistent with prior findings. A common mistake in Wason tasks of the form $A \Rightarrow B$ is to turn over "B" rather than "not B".

In the anchoring tasks, we find statistically significant evidence of anchoring on irrelevant information. Across questions, the correlations between subjects' estimates and the anchors range between $\rho = 0.38$ and $\rho = 0.60$.

In summary, pooling across incentive conditions, we find strong evidence for the existence of cognitive biases, on average. We now turn to the main object of interest of our study, which is the effect of financial incentives.

3.2 Incentives and Effort

We start by examining whether higher stakes induce participants to increase effort, using response time as a proxy for effort. Response times are a widely used proxy for cogni-

[11]The respondents appear to exhibit a meaningful level of motivation. In our survey, we only briefly describe the study by providing the names of the experimental tasks. In addition, we provide the respondents with an option to view details on the implementation of these tasks. Across the six different tasks, 53%-84% of respondents elect to view the task details, with an overall average of 68%. Of course, some respondents may not need to look up the task details because they know the task structure.

tive effort in laboratory experiments (e.g., Luce et al., 1986; Ratcliff, 1978; Rubinstein, 2007; Krajbich et al., 2012; Spiliopoulos and Ortmann, 2018). This analysis can plausibly be understood as a “first stage” for the relationship between incentives and cognitive biases. In absolute terms, average response times range from 99 seconds per question in anchoring to 425 seconds per question in intuitive base rate neglect, which includes the time it takes participants to read the (very short) instructions on their decision screens.

Figure 4 visualizes mean response times by incentive level, separately for each experimental task. Here, to ease interpretation, response times are normalized to one in the no incentives condition. In other words, for each cognitive bias, we divide observed response times by the average response time in the no incentives condition. Thus, in Figure 4, response times can be interpreted as a percentage of response times in the no incentives condition.

We find that standard lab incentives generally do not increase response times much compared to no incentives. High incentives, however, robustly lead to greater effort, a pattern that is very similar across all tasks. Overall, response times are 39 percent higher in the high incentive condition compared to standard incentives. We observe the largest increase (52 percent) in intuitive base rate neglect, and the smallest increase (24 percent) in anchoring. Figure 10 in Appendix D shows that very similar results hold when we look at median response times.

Table 2 quantifies the effects of incentive size on response times (in seconds) using OLS regressions.[12] In these regressions, the omitted category is the standard incentive condition. Thus, the coefficients of the no incentive and the high incentive conditions identify the change in response times in seconds relative to the standard incentive condition. The last row of the table reports the p-value of a test for equality of regression coefficients between *No incentives* and *High incentives*, although again this comparison is

[12] In Table 2, we use raw response times. In mathematical psychology, researchers frequently rely on log response times, $\ln(1+RT)$, because of the oftentimes skewed nature of response time data. In our data, the residuals are indeed not normally distributed when we use raw response times. In Table 6 in Appendix C, we instead use log response times. A p-p plot of residuals shows that they follow a normal distribution in this case (Figure 11 in Appendix D). The treatment comparisons deliver the same qualitative results as with raw response times. Table 7 in Appendix C provides complementary nonparametric tests that also deliver very similar results.

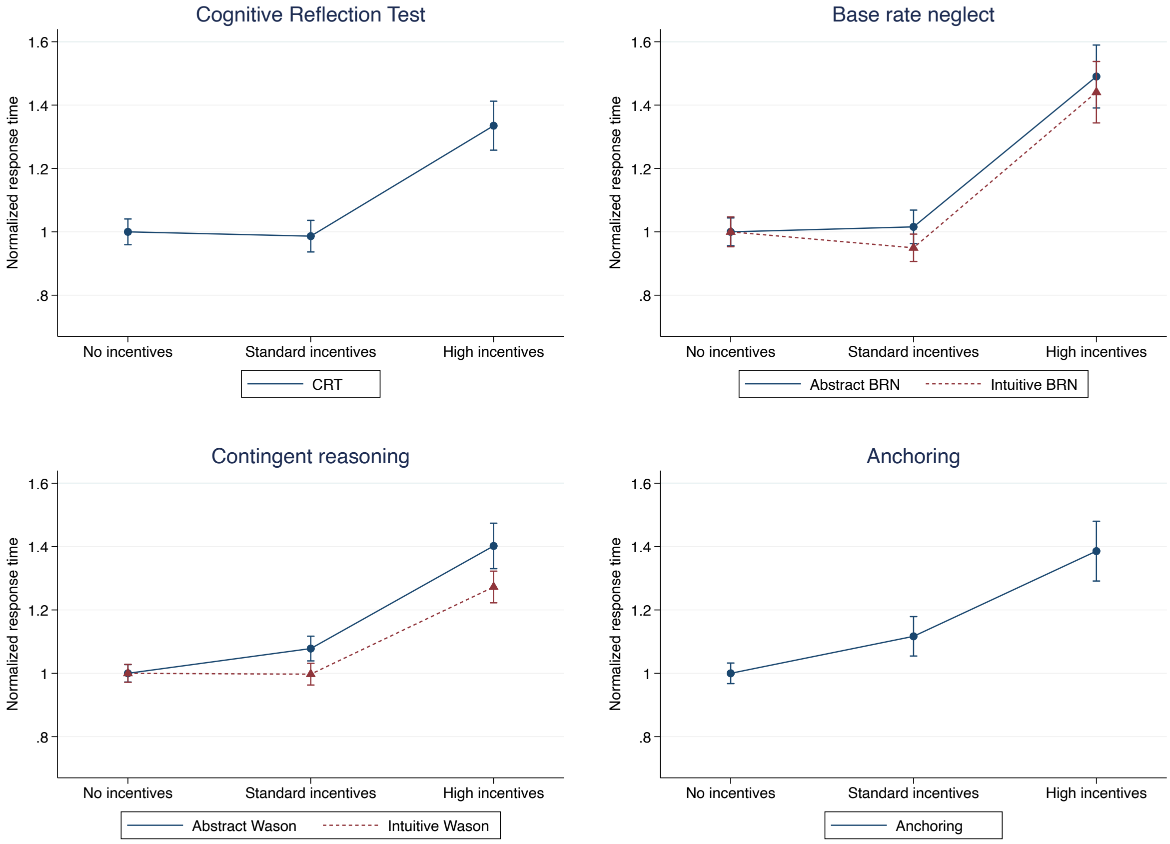


Figure 4: Average normalized response times across incentive conditions. Response times are normalized relative to the no incentive condition: for each cognitive bias, we divide observed response times by the average response time in the no incentive condition. Error bars indicate +/- 1 s.e. Average response times per question in the no incentives scheme are: 171 sec. in CRT, 335 sec. in abstract BRN, 425 sec. in intuitive BRN, 181 sec. in abstract Wason, 113 sec. in intuitive Wason, and 99 sec. in anchoring.

not based on randomization. In the regressions, an observation is the response time on a given question. However, because each subject completed two questions for the same bias, we have two observations per subject, so we always cluster the standard errors at the subject level.

We can never reject the hypothesis that cognitive effort in the flat payment and standard incentive schemes are identical. In fact, the estimated coefficient is sometimes positive and sometimes negative. While it should be kept in mind that the coefficient of the no incentive condition is potentially confounded by order effects, we still view this result as suggestive.

High stakes, on the other hand, significantly increase response times by between 24 seconds (anchoring) and 191 seconds (intuitive base rate neglect), relative to the

Table 2: Response times across incentive conditions

	Dependent variable: Response time [seconds]						
Omitted category:		Base rate neglect		Contingent reasoning			All
Standard incentives	CRT	Abstract	Intuitive	Abstract	Intuitive	Anchoring	tasks
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 if *No incentives*	2.16	-4.71	19.5	-12.7	0.28	-10.2	-1.61
	(10.25)	(20.51)	(24.79)	(7.72)	(4.72)	(6.14)	(5.50)
1 if *High incentives*	55.5**	141.6**	190.7**	52.4**	29.2**	23.6*	71.5**
	(14.63)	(33.55)	(41.17)	(13.21)	(6.43)	(9.88)	(9.42)
Constant	157.2**	303.1**	368.8**	174.5**	105.9**	97.8**	81.6**
	(7.94)	(15.77)	(16.74)	(6.31)	(3.64)	(5.44)	(5.11)
Task type FE	No	No	No	No	No	No	Yes
Observations	1240	618	618	619	619	1230	4944
R^2	0.02	0.05	0.05	0.07	0.05	0.03	0.29
p-value: *No inc.* = *High inc.*	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

Notes. OLS estimates, standard errors (clustered at subject level) in parentheses. Omitted category: standard incentive scheme. The last row reports the p-value of a test for equality of regression coefficients between *No incentives* and *High incentives*. * $p < 0.05$, ** $p < 0.01$.

standard incentive treatment. On average, response times increase by 72 seconds, see the analysis on the pooled sample in column (7). As we show in Figure 9 in Appendix D, the empirical cumulative distribution functions of response times in the high incentive conditions usually first-order stochastically dominate the CDFs in the other conditions.

Even though in relative terms high stakes induce a substantial increase in response times, the rather modest increase in absolute response times is noteworthy, given the large increase in financial incentives. Potential explanations for this – which we cannot disentangle – are the presence of substantial cognitive effort costs, overconfidence, or a belief that more cognitive effort does not improve performance on the margin.[13]

Result 1. *Very high incentives increase response times by 24–52% relative to standard lab incentives. Response times are almost identical with standard incentives and no incentives.*

[13]A related alternative explanation for the modest increase in effort could be that a high bonus signals that the task is hard, undermining a participant's confidence in their ability to solve it (see e.g., Deci (1975); Deci and Ryan (1985) for some early accounts of the informational aspects of rewards, and Benabou and Tirole (2003) for a formalization). Our data do not support this explanation as a driving factor, as we do not observe a decrease in confidence levels as the stake size increases, see Table 14 in Appendix C.

3.3 Incentives and Cognitive Biases

Figure 5 shows how variation in the stake size affects the prevalence of our cognitive biases. For the CRT, base rate neglect, and the Wason selection task, the figure shows the fraction of correct answers. For base rate neglect, following our pre-registration, we count a response as "correct" if it is within 5 percentage points of the Bayesian posterior. While subjects only received a bonus if their answer was within 2 percentage points of the Bayesian response, we work here with a slightly larger interval to allow for random computational errors. For anchoring, we plot one minus the Pearson correlation coefficient between responses and the anchor, so that higher values reflect less bias.

The main takeaway is that performance barely improves. In the CRT, performance in the high incentive condition increases by about 10 percentage points relative to the standard incentive condition. However, in all other tasks, improvements are either very small or entirely absent. Across all tasks, high incentives never come close to de-biasing participants. These results suggest that judgmental biases are not an artifact of weak incentives.

Table 3 quantifies these results through regression analysis.[14] Here, in columns (1)–(6), the dependent variable is whether a given task was solved correctly. In the first six columns, the coefficients of interest are the treatment dummies. Again, the omitted category is the standard incentive condition.

For anchoring, column (7), the object of interest is not whether a subject's answer is objectively correct, but instead how answers vary as a function of the anchor. Thus, the coefficients of interest are the interactions between the anchor and the treatment dummies.

Compared to standard incentives, the flat payment dummy usually has a negative point estimate in columns (1)–(5). While these are not statistically significant, we see in column (6) that when we pool the data across the tasks from the first five columns, performance is significantly lower without incentives, although the effect size is quite

[14]Table 8 in Appendix C provides complementary nonparametric tests that deliver very similar results.

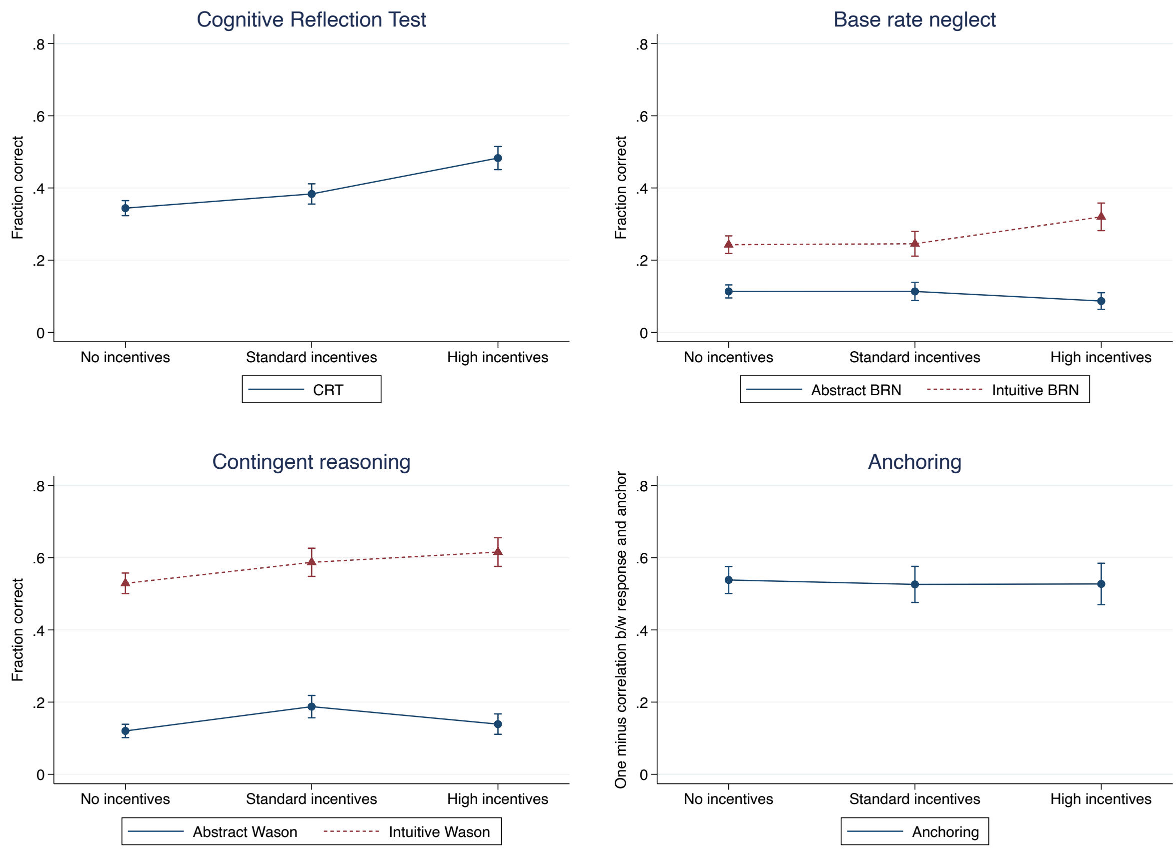


Figure 5: Average performance by incentive level. Error bars indicate +/- 1 s.e. The performance metrics are computed as follows. For the CRT, we count a response as correct if it is exactly correct. For base rate neglect, we count a response as correct if it is within 5 percentage points of the Bayesian posterior. For the Wason selection task, we count a response as correct if the participant turned over (only) the two correct cards. For anchoring, we plot one minus the correlation coefficient between responses and anchors.

small.[15] High stakes, on the other hand, lead to a statistically significant increase in performance on the CRT. For intuitive base rate neglect, the intuitive Wason task, and anchoring, the estimated coefficients of interest are positive but not statistically significant. For abstract base rate neglect and the abstract Wason task, the point estimates are even negative. In the pooled sample, column (6), performance does not increase under high incentives compared to standard incentives. Performance does improve significantly relative to no incentives, but this difference (i) is almost entirely driven by the CRT and (ii) potentially confounded by order effects.

In quantitative terms, the improvements in performance are modest. Importantly,

[15]Including the anchoring data in this pooled regression is not meaningful because for anchoring the effect of interest is not the treatment dummy but its interaction with the anchor.

Table 3: Performance by incentive level

	Dependent variable:						
	1 if answer correct						Answer
Omitted category:		Base rate neglect		Contingent reasoning		Tasks	
Standard incentives	CRT	Abstract	Intuitive	Abstract	Intuitive	1–5	Anchoring
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 if *No incentives*	-0.039	0.000061	-0.0026	-0.067	-0.058	-0.034*	5.60
	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.02)	(3.19)
1 if *High incentives*	0.099*	-0.027	0.075	-0.048	0.028	0.037	3.08
	(0.04)	(0.03)	(0.05)	(0.04)	(0.06)	(0.02)	(3.59)
Anchor							0.49**
							(0.05)
Anchor × 1 if *No incentives*							0.0037
							(0.07)
Anchor × 1 if *High incentives*							0.017
							(0.08)
Constant	0.38**	0.11**	0.25**	0.19**	0.59**	0.11**	12.7**
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.02)	(2.45)
Task type FE	No	No	No	No	No	Yes	No
Observations	1240	618	618	619	619	3714	1230
R^2	0.01	0.00	0.01	0.01	0.01	0.12	0.22
p-value: *No inc.* = *High inc.*	< 0.01	0.36	0.09	0.58	0.08	< 0.01	0.86

Notes. OLS estimates, robust standard errors (clustered at subject level) in parentheses. In columns (1)–(6), the dependent variable is a binary indicator for whether an answer is correct. In column (6), the sample includes all tasks from columns (1)–(5). In column (7), the outcome variable is the answer (between 0 and 100). Omitted category: standard incentives. In columns (1)–(6), the last row reports the p-value of a test for the equality of regression coefficients between *No incentives* and *High incentives*. In column (7), the last row reports the p-value of a test for the equality of regression coefficients between Anchor × 1 if *No incentives* and Anchor × 1 if *High incentives*. * $p < 0.05$, ** $p < 0.01$.

the weak effects of the large increase in financial incentives are not driven by a lack of statistical power. Given our large sample size, the regression coefficients are relatively tightly estimated. Looking at 95% confidence intervals derived from the regression analysis, we can rule out that, going from standard to high incentives, performance increases by more than: 18 pp in the CRT, 4 pp in abstract base rate neglect, 18 pp in intuitive base rate neglect, 3 pp in the abstract Wason task, and 14 pp in the intuitive Wason task. For anchoring, we can rule out that the OLS coefficient of the high incentives condition is smaller than 17 pp. Notably, for the more abstract tasks, we can rule out performance increass of only 3–4 pp, while in the more intuitive tasks the point estimates and confidence bands are a bit larger.[16] Indeed, for intuitive Wason and intuitive base rate neglect,

[16]It is worth pointing out that this generally small effect of incentives casts some doubt on explanations

the point estimates and confidence intervals are such that we cannot statistically reject that they are different from the point estimate for the CRT.

The last row of Table 3 reports the p-value for equality of coefficients between *No incentives* and *High incentives*. While we caution again that this comparison is not based on randomization, the results are broadly similar, except that for intuitive base rate neglect and the intuitive Wason task, the improvement in performance is marginally significant.

The results that (i) the largest and most robust performance improvements occur in the CRT and (ii) the performance increases are mildly stronger for the more intuitive versions of base rate neglect and the Wason task, are informative. The CRT was designed to capture reliance on deliberative vs. intuitive reasoning. The other tasks, however, are usually considered to be fairly difficult. It may be that the higher cognitive effort that is induced by high incentives reduces reliance on intuitive "gut feelings," but does not help with solving more complex problems. We return to this observation below.

Result 2. *Relative to standard incentives, very high incentives do not reduce cognitive biases, except for in the domain of intuitive vs. deliberative thinking. We find almost no difference in behavior between standard and no incentives.*

Types of mistakes. Up to this point, we have kept our analysis relatively simple by focusing – except for anchoring – on a binary performance classification. Here, we briefly discuss to what extent the size of incentives did or did not affect the specific types of mistakes our participants make.

First, in BRN, we only classify a response as correct if it falls within five percentage points of the Bayesian posterior. In Table 12 in Appendix C, we check whether the absolute distance between the response and the Bayesian posterior is affected by the size of incentives, as would be the case if subjects had improved but not enough to make it

of biases as arising from some version of rational inattention or optimized cognition. To take but one example, in a recent anchoring model by Lieder et al. (2018), people make rational use of finite time and limited cognitive resources, and are predicted to suffer less from anchoring effects with steeper incentives. Our study provides a direct test of this mechanism, and rejects it.

to the relatively narrow five-point band. However, in neither of the two BRN tasks does performance improve under very high incentives using this continuous measure.

Second, we investigate whether incentives affect the extent to which participants' responses directly correspond to the well-known heuristic / intuitive response patterns that the respective literatures have documented. We define these intuitive answers: the well-known impulsive answer for CRT; completely neglecting the base rate for BRN, which corresponds to simply reporting the conditional probability of a positive test given that a person is ill; and directly reporting the anchor for anchoring. For the Wason task of the form "if P then Q," we can identify two types of intuitive mistakes. A first most common mistake is to only turn over "P." A second mistake is to turn over "P" and "Q" instead of "P" "not Q."

Table 13 in Appendix C shows how the incidence of intuitive answers depends on stake size. For CRT, we find that participants are less likely to give the impulsive answer (consistent with the results reported above). For the abstract version of base-rate neglect, we find that with high incentives subjects are less likely to completely ignore the base rate compared to normal incentives but not compared to no incentives. However, there is no similar effect for the intuitive version of BRN. In the Wason task, we find no evidence that high stakes reduce the frequency of making the two specific mistakes explained above. Likewise, high incentives do not reduce the frequency of directly reporting the anchor in anchoring. In all, these results suggest that – with the exception of the CRT – very high stakes do not meaningfully reduce the frequency of particular well-known heuristic responses.

Robustness and heterogeneity analyses. In the pre-registration, we noted that we would consider heterogeneity along different sociodemographic variables such as cognitive skills. Table 11 in Appendix C shows that controlling for individual characteristics and question fixed effects leaves the results unaffected. We also conduct heterogeneity analyses with respect to college GPA, score on a Raven matrices test (a measure of intelligence), and income. We find no robust evidence of heterogeneous treatment effects

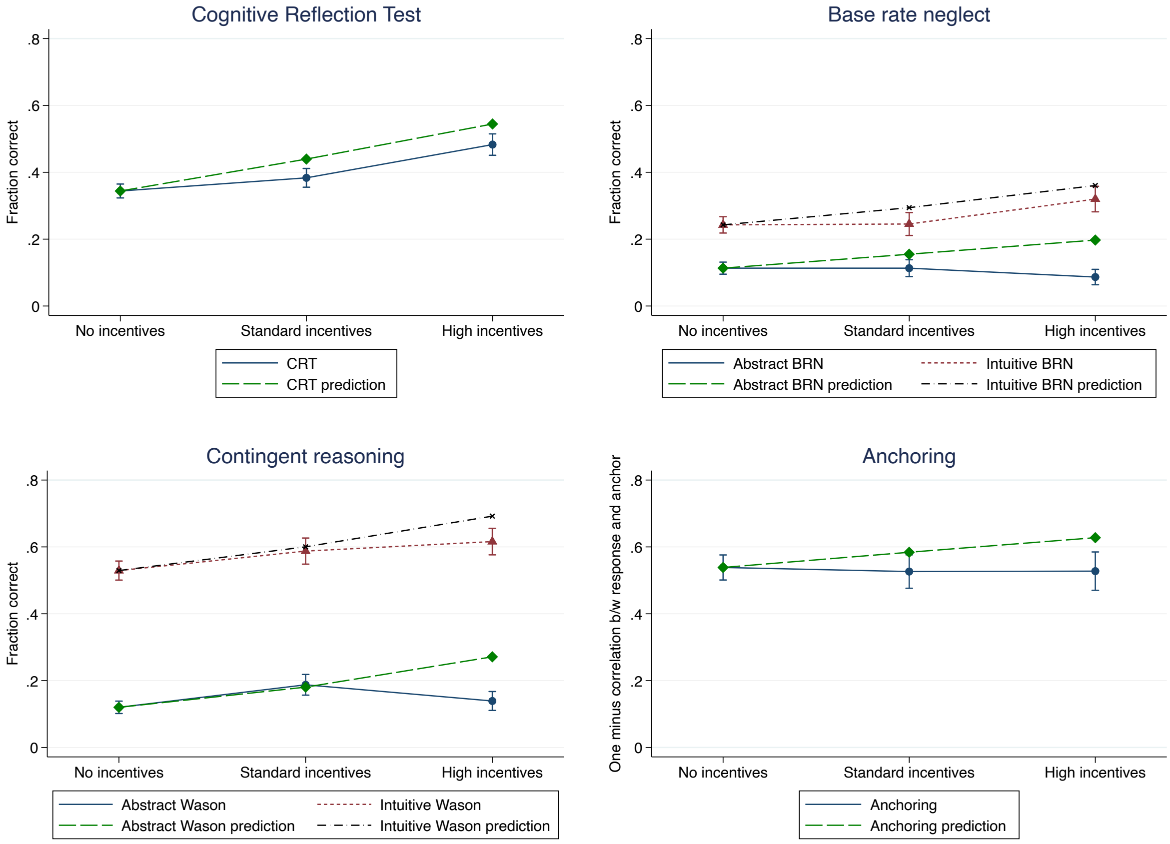


Figure 6: Average performance by incentive level and average expert predictions. Error bars on performance indicate +/- 1 s.e. The performance metrics are computed as described in the notes of Figure 5. The expert predictions for anchoring are adjusted due to a typo in our survey. We informed the experts that without any financial incentives, the correlation between answers and anchors is $\rho = 0.49$, while it is actually only 0.46. Since we are mostly interested in *changes* in respondent predictions across the different incentive levels, we adjust the data by deducting 3 pp. from each respondent prediction. Average respondent forecasts for high stakes are always higher than actual average performance, but only fall outside of a 95% confidence interval around actual performance for abstract base rate neglect and the abstract Wason task.

along these dimensions.

3.4 Comparison with Predictions

To put our results in perspective, we compare them with experimental economists' predictions, collected as described in Section 2.6. Recall that we provide these researchers with information on performance in the no incentive condition and ask them to predict performance in the standard and high incentive conditions. Figure 6 shows the results. The respondents are qualitatively correct in the sense that they predict that errors will not disappear even with very large incentives. At the same time, the respondents always predict larger performance increases than the actual data reveal. On average, researchers expect about a 25% increase in performance going from no to standard incentives, and then again a 25% increase going from standard to high incentives.[17] Mispredictions appear particularly pronounced for abstract base rate neglect, the abstract Wason task, and anchoring. Across all tasks, 56% of respondent predictions fall outside of a 95% confidence interval around average actual performance, and of these mispredictions, 90% are too high rather than too low. Prediction accuracy is highest in intuitive base rate neglect (69% inside the confidence interval) and lowest in abstract base rate neglect (24%).[18]

Result 3. *Experts correctly predict that biases do not disappear with very high incentives, yet they overestimate the responsiveness of performance to incentives, in particular for very high incentives.*

3.5 Potential Mechanisms

Effort and Performance. The results discussed up to this point suggest that the increase in response times by up to 50% that was induced by high incentives did not translate into a reduction in the frequency of biases of anything close to the same magnitude. This raises the question about the more general relationship between effort and performance in cognitive biases tasks. Indeed, while previous literature has not focused on implementing large increases in financial incentives, researchers have occasionally reported correlations between response times and observed biases. A recurring finding is that the relationship between errors and response times is statistically significant but often quantitatively small.[19]

[17]Figure 12 in Appendix D shows that the respondents also substantially overestimate the increase in response times going from no to standard and from standard to high incentives. On average, the respondents forecast increases of around 25% going from no incentives to standard incentives, and another 40–60% going from standard to high incentives.

[18]Appendix E provides a more complete picture of the relationship between respondent forecasts and actual performance, with plots of the empirical distribution of respondent forecasts against the posterior distribution of actual performance on each task.

[19]See Enke and Zimmermann (2019), Enke (2020), and Graeber (2019).

In our study, similar patterns hold. As shown in Table 4, longer response times are correlated with a higher probability of solving a problem correctly, yet the magnitudes of the OLS coefficients are fairly small. Interpreted causally, the coefficients suggest that – across biases – spending one additional minute on a problem increases the probability of answering it correctly by about one percentage point. Given that the standard deviation of response times in the sample after partialling out question fixed effects is 183 seconds, this implies that response times would have to increase by 33 standard deviations (6,000 seconds) to increase the probability of answering correctly from zero to one. Our interpretation is that these "effect sizes" are much too small to plausibly explain within-treatment heterogeneity in performance purely as a result of heterogeneity in effort expended. Under this interpretation, correctly solving the types of problems that are associated with well-known cognitive biases requires not so much large amounts of effort but instead "the right way of looking at the problem." To the extent that financial incentives may only increase cognitive effort *per se* rather than substantially improving the problem solving approach, stakes might not matter all that much for performance.

In this regard, it is also informative that the largest increase in performance is visible in the CRT, where finding the correct solution arguably requires only an ability or willingness to overcome gut instincts, rather than advanced conceptual reasoning skills. That is, in the CRT – and unlike in many of the other tasks – the intuitive, wrong answers are relatively easy to disprove even without changing one's mental framework.

While these analyses are all descriptive in nature, they can be interpreted as suggesting that the difficulty in overcoming cognitive biases is often conceptual in nature, and that higher effort does not easily induce "the right way of looking at the problem." Such an interpretation is in line with other recent work that has emphasized the importance of how people look at problems and of "mental gaps," as opposed to only cost-benefit tradeoffs (see Handel and Schwartzstein, 2018, for an overview). Viewed through the lens of the popular two-systems approach to reasoning (Frederick, 2005; Stanovich and West, 2000), our results suggest that reducing reliance on "system 1" is not enough to

Table 4: Performance, response times, and cognitive skills within treatments

	Dependent variable: 1 if answer correct					Answer
		Base rate neglect		Contingent reasoning		
	CRT	Abstract	Intuitive	Abstract	Intuitive	Anchoring
	(1)	(2)	(3)	(4)	(5)	(6)
Response time [minutes]	0.018**	0.0088**	0.014**	0.0087	0.024	0.39
	(0.01)	(0.00)	(0.00)	(0.01)	(0.02)	(0.76)
Cognitive skills [z-score]	0.10**	0.0065	0.033	0.015	0.079**	-1.13
	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(1.45)
Anchor						0.52**
						(0.06)
Anchor × Response time						-0.0090
						(0.02)
Anchor × Cognitive skills						-0.037
						(0.03)
Constant	0.27**	0.028	0.22**	0.18**	0.52**	0.59
	(0.04)	(0.03)	(0.04)	(0.04)	(0.06)	(2.63)
Treatment FE	Yes	Yes	Yes	Yes	Yes	Yes
Anchor × Treatment FE	No	No	No	No	No	Yes
Question FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1240	618	618	619	619	1230
R^2	0.07	0.03	0.06	0.01	0.03	0.37

Notes. OLS estimates, robust standard errors (clustered at subject level) in parentheses. In columns (1)-(5), the dependent variable is a binary indicator for whether an answer is correct. In column (6), the outcome variable is the answer (between 0 and 100). The cognitive skills variable is the average of the z-scores of (i) GPA on the Kenya Certificate of Secondary Education exam and (ii) score on a ten-item Raven matrices test. * $p < 0.05$, ** $p < 0.01$.

overcome biases because these are often limitations of the more deliberative "system 2."

Confidence. At the end of the experiment, we elicited subjects' self-reported confidence in the correctness of their answers (on a 0–7 Likert scale). While not incentivized, the data allow us to gauge how confident subjects are in their (usually wrong) responses, and how confidence varies as a function of the stake size.

In our data, average confidence is 5.3 in base rate neglect, 6.2 in the CRT, 6.2 in the Wason tasks and 4.6 in anchoring. These data are indicative that subjects were relatively confident in their responses. As we show in Table 14 in Appendix C, reported confidence increases very little, if at all, as the stake size increases. This may suggest that while participants put in more effort when the stakes are higher, they are partially aware that this does not translate into a significantly higher probability of solving the problem correctly because they lack the skills to develop the right problem-solving approach.[20]

4 Discussion

This paper provides a systematic investigation of a long-standing question in economics: are people less likely to fall prey to cognitive biases when the stakes are very high? In experiments with a large sample of college students, we increase the financial incentives for accuracy by a factor of 100 to more than a full monthly income in the population of interest. Despite this drastic increase in incentives, performance improves either very modestly, or not at all. We view these results as having three main implications. First, our results are encouraging news for the large literature on the "heuristics and biases" program in experimental economics and psychology, as it suggests that the results in this literature need not be understood as contingent on a particular incentive level. Second, an active theoretical literature attempts to model how different cognitive biases arise, where an important question is whether systematic errors arise due to genuine cognitive limitations or as a result of inattention and low effort. Our experiments find support for the former explanation in the biases we study. Third, for economists more generally, our results highlight that the detrimental effects of the cognitive biases that are studied in the experimental economics literature plausibly play out also in decisions with large economic consequences. This result resonates with a considerable body of work on field studies with high-powered incentives – referenced in footnote 2 – that often identify systematic biases. In response to Thaler's question in our opening paragraph, this combination of field and lab data strongly suggests that people do not necessarily tend to make better decisions when the stakes are very high.

[20]An alternative interpretation centers on overconfidence. If subjects are very confident that they are getting the task right even with standard incentives, then very high stakes need not improve performance. This interpretation, however, is slightly at odds with the increase in response times.

References

Andersen, Steffen, Seda Ertaç, Uri Gneezy, Moshe Hoffman, and John A List, "Stakes matter in ultimatum games," *American Economic Review*, 2011, *101* (7), 3427–39.

Ariely, Dan, George Loewenstein, and Drazen Prelec, ""Coherent arbitrariness": Stable demand curves without stable preferences," *The Quarterly journal of economics*, 2003, *118* (1), 73–106.

_ , Uri Gneezy, George Loewenstein, and Nina Mazar, "Large stakes and big mistakes," *The Review of Economic Studies*, 2009, *76* (2), 451–469.

Arkes, Hal R, Robyn M Dawes, and Caryn Christensen, "Factors influencing the use of a decision rule in a probabilistic task," *Organizational behavior and human decision processes*, 1986, *37* (1), 93–110.

Barron, Kai, Steffen Huck, and Philippe Jehiel, "Everyday econometricians: Selection neglect and overoptimism when learning from others," Technical Report, WZB Discussion Paper 2019.

Baumert, Jürgen and Anke Demmrich, "Test motivation in the assessment of student skills: The effects of incentives on motivation and performance," *European Journal of Psychology of Education*, 2001, *16* (3), 441–462.

Bazerman, Max H and William F Samuelson, "I won the auction but don't want the prize," *Journal of conflict resolution*, 1983, *27* (4), 618–634.

Beggs, Alan and Kathryn Graddy, "Anchoring effects: Evidence from art auctions," *American Economic Review*, 2009, *99* (3), 1027–39.

Belot, Michèle, V Bhaskar, and Jeroen van de Ven, "Promises and cooperation: Evidence from a TV game show," *Journal of Economic Behavior & Organization*, 2010, *73* (3), 396–405.

Benabou, Roland and Jean Tirole, "Intrinsic and extrinsic motivation," *The review of economic studies*, 2003, *70* (3), 489–520.

Benjamin, Daniel J, "Errors in Probabilistic Reasoning and Judgmental Biases," in "Handbook of Behavioral Economics" 2019.

Berk, Jonathan B, Eric Hughson, and Kirk Vandezande, "The price is right, but are the bids? An investigation of rational decision theory," *The American Economic Review*, 1996, pp. 954–970.

Bettinger, Eric P, "Paying to learn: The effect of financial incentives on elementary school test scores," *Review of Economics and Statistics*, 2012, *94* (3), 686–698.

Binswanger, Hans P, "Attitudes toward risk: Experimental measurement in rural India," *American journal of agricultural economics*, 1980, *62* (3), 395–407.

Borghans, Lex, Huub Meijers, and Bas Ter Weel, "The role of noncognitive skills in explaining cognitive test scores," *Economic inquiry*, 2008, *46* (1), 2–12.

Brañas-Garza, Pablo, Praveen Kujal, and Balint Lenkei, "Cognitive Reflection Test: whom, how, when," *Journal of Behavioral and Experimental Economics*, 2019.

Camerer, Colin F., "Do biases in probability judgment matter in markets? Experimental evidence," *American Economic Review*, 1987, *77* (5), 981–997.

_ **and Robin M Hogarth**, "The effects of financial incentives in experiments: A review and capital-labor-production framework," *Journal of Risk and Uncertainty*, 1999, *19*, 7–41.

Cameron, Lisa A, "Raising the stakes in the ultimatum game: Experimental evidence from Indonesia," *Economic Inquiry*, 1999, *37* (1), 47–59.

Chapman, Gretchen B and Eric J Johnson, "Incorporating the irrelevant: Anchors in judgments of belief and value," *Heuristics and biases: The psychology of intuitive judgment*, 2002, pp. 120–138.

Chen, Daniel L, Tobias J Moskowitz, and Kelly Shue, “Decision making under the gambler’s fallacy: Evidence from asylum judges, loan officers, and baseball umpires,” *The Quarterly Journal of Economics*, 2016, *131* (3), 1181–1242.

Cheng, Patricia W and Keith J Holyoak, “Pragmatic reasoning schemas,” *Cognitive psychology*, 1985, *17* (4), 391–416.

Cooper, David J, John H Kagel, Wei Lo, and Qing Liang Gu, “Gaming against managers in incentive systems: Experimental results with Chinese students and Chinese managers,” *American Economic Review*, 1999, *89* (4), 781–804.

Cosmides, Leda, “The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task,” *Cognition*, 1989, *31* (3), 187–276.

Danz, David N, Lise Vesterlund, and Alistair J Wilson, “Belief elicitation: Limiting truth telling with information on incentives,” Technical Report, Mimeo 2019.

Deci, EL, “Intrinsic motivation,” 1975.

_ **and RM Ryan**, “Intrinsic motivation and self-determination in human behavior (Perspectives in Social Psychology),” 1985.

DellaVigna, Stefano and Devin Pope, “Predicting experimental results: who knows what?,” *Journal of Political Economy*, 2018, *126* (6), 2410–2456.

den Assem, Martijn J Van, Dennie Van Dolder, and Richard H Thaler, “Split or steal? Cooperative behavior when the stakes are large,” *Management Science*, 2012, *58* (1), 2–20.

Enke, Benjamin, “What you see is all there is,” *The Quarterly Journal of Economics*, 2020, *135* (3), 1363–1398.

_ **and Florian Zimmermann**, “Correlation Neglect in Belief Formation,” *Review of Economic Studies*, 2019, *86* (1), 313–332.

Epley, Nicholas and Thomas Gilovich, "When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors," *Journal of Behavioral Decision Making*, 2005, *18* (3), 199–212.

_ **and** _ , "The anchoring-and-adjustment heuristic: Why the adjustments are insufficient," *Psychological science*, 2006, *17* (4), 311–318.

_ **, Boaz Keysar, Leaf Van Boven, and Thomas Gilovich**, "Perspective taking as egocentric anchoring and adjustment.," *Journal of personality and social psychology*, 2004, *87* (3), 327.

Esponda, Ignacio and Emanuel Vespa, "Hypothetical Thinking and Information Extraction in the Laboratory," *American Economic Journal: Microeconomics*, 2014, *6* (4), 180–202.

_ **and** _ , "Hypothetical Thinking: Revisiting Classic Anomalies in the Laboratory," *Working Paper*, 2016.

Frederick, Shane, "Cognitive reflection and decision making," *Journal of Economic perspectives*, 2005, *19* (4), 25–42.

Fudenberg, Drew, David K Levine, and Zacharias Maniadis, "On the robustness of anchoring effects in WTP and WTA experiments," *American Economic Journal: Microeconomics*, 2012, *4* (2), 131–45.

Gigerenzer, Gerd and Ulrich Hoffrage, "How to improve Bayesian reasoning without instruction: frequency formats.," *Psychological review*, 1995, *102* (4), 684.

Giglio, Stefano and Kelly Shue, "No news is news: do markets underreact to nothing?," *The Review of Financial Studies*, 2014, *27* (12), 3389–3440.

Gneezy, Uri and Aldo Rustichini, "Pay enough or don't pay at all," *The Quarterly journal of economics*, 2000, *115* (3), 791–810.

— , **Stephan Meier, and Pedro Rey-Biel**, "When and why incentives (don't) work to modify behavior," *Journal of economic perspectives*, 2011, *25* (4), 191–210.

Goodie, Adam S and Edmund Fantino, "An experientially derived base-rate error in humans," *Psychological Science*, 1995, *6* (2), 101–106.

Graddy, Kathryn, Lara Loewenstein, Jianping Mei, Michael Moses, and Rachel AJ Pownall, "Anchoring or loss aversion? Empirical evidence from art auctions," 2014.

Graeber, Thomas, "Inattentive Inference," *Working Paper*, 2019.

Grether, David M., "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *Quarterly Journal of Economics*, 1980, *95*, 537–557.

Grether, David M, "Testing Bayes rule and the representativeness heuristic: Some experimental evidence," *Journal of Economic Behavior & Organization*, 1992, *17* (1), 31–57.

— **and Charles R Plott**, "Economic theory of choice and the preference reversal phenomenon," *The American Economic Review*, 1979, *69* (4), 623–638.

Handel, Benjamin and Joshua Schwartzstein, "Frictions or mental gaps: what's behind the information we (don't) use and when do we care?," *Journal of Economic Perspectives*, 2018, *32* (1), 155–78.

Holt, Charles A and Susan K Laury, "Risk aversion and incentive effects," *American economic review*, 2002, *92* (5), 1644–1655.

Hoppe, Eva I and David J Kusterer, "Behavioral biases and cognitive reflection," *Economics Letters*, 2011, *110* (2), 97–100.

Ioannidis, Konstantinos, Theo Offerman, and Randolph Sloof, "On the effect of anchoring on valuations when the anchor is transparently uninformative," *Journal of the Economic Science Association*, 2020, *6* (1), 77–94.

Jetter, Michael and Jay K Walker, "Anchoring in financial decision-making: Evidence from Jeopardy!," *Journal of Economic Behavior & Organization*, 2017, *141*, 164–176.

Johnson-Laird, Philip Nicholas, *Mental models: Towards a cognitive science of language, inference, and consciousness* number 6, Harvard University Press, 1983.

Jones, Martin and Robert Sugden, "Positive confirmation bias in the acquisition of information," *Theory and Decision*, 2001, *50* (1), 59–99.

Jr, Harold F O'Neil, Brenda Sugrue, and Eva L Baker, "Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance," *Educational Assessment*, 1995, *3* (2), 135–157.

Jr, Roland G Fryer, "Financial incentives and student achievement: Evidence from randomized trials," *The Quarterly Journal of Economics*, 2011, *126* (4), 1755–1798.

Kahneman, Daniel and Amos Tversky, "On the psychology of prediction," *Psychological Review*, 1973, *80* (4), 237–251.

Klein, Barbara D, "Detecting errors in data: clarification of the impact of base rate expectations and incentives," *Omega*, 2001, *29* (5), 391–404.

Krajbich, Ian, Dingchao Lu, Colin Camerer, and Antonio Rangel, "The attentional drift-diffusion model extends to simple purchasing decisions," *Frontiers in psychology*, 2012, *3*, 193.

Levitt, Steven D, "Testing theories of discrimination: evidence from Weakest Link," *The Journal of Law and Economics*, 2004, *47* (2), 431–452.

_ , John A List, Susanne Neckermann, and Sally Sadoff, "The behavioralist goes to school: Leveraging behavioral economics to improve educational performance," *American Economic Journal: Economic Policy*, 2016, *8* (4), 183–219.

Lieder, Falk, Thomas L Griffiths, Quentin JM Huys, and Noah D Goodman, "The anchoring bias reflects rational use of cognitive resources," *Psychonomic bulletin & review*, 2018, *25* (1), 322–349.

Luce, R Duncan et al., *Response times: Their role in inferring elementary mental organization* number 8, Oxford University Press on Demand, 1986.

Maniadis, Zacharias, Fabio Tufano, and John A List, "One swallow doesn't make a summer: New evidence on anchoring effects," *American Economic Review*, 2014, *104* (1), 277–90.

Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa, "Failures in Contingent Reasoning: The Role of Uncertainty," *American Economic Review*, 2019.

Metrick, Andrew, "A natural experiment in" Jeopardy!"," *The American Economic Review*, 1995, pp. 240–253.

Nelson, Thomas E, Monica R Biernat, and Melvin Manis, "Everyday base rates (sex stereotypes): Potent and resilient.," *Journal of Personality and Social Psychology*, 1990, *59* (4), 664.

Oechssler, Jörg, Andreas Roider, and Patrick W Schmitz, "Cognitive abilities and behavioral biases," *Journal of Economic Behavior & Organization*, 2009, *72* (1), 147–152.

O'Neil, Harold F, Jamal Abedi, Judy Miyoshi, and Ann Mastergeorge, "Monetary incentives for low-stakes tests," *Educational Assessment*, 2005, *10* (3), 185–208.

Parravano, Melanie and Odile Poulsen, "Stake size and the power of focal points in coordination games: Experimental evidence," *Games and Economic Behavior*, 2015, *94*, 191–199.

Pope, Devin G and Maurice E Schweitzer, "Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes," *American Economic Review*, 2011, *101* (1), 129–57.

Rapoport, Amnon, William E Stein, James E Parco, and Thomas E Nicholas, "Equilibrium play and adaptive learning in a three-person centipede game," *Games and Economic Behavior*, 2003, *43* (2), 239–265.

Ratcliff, Roger, “A theory of memory retrieval.,” *Psychological review*, 1978, *85* (2), 59.

Rubinstein, Ariel, “Instinctive and Cognitive Reasoning: A Study of Response Times,” *Economic Journal*, 2007, *117* (523), 1243–1259.

Sandefur, Justin, “Internationally comparable mathematics scores for fourteen African countries,” *Economics of Education Review*, 2018, *62*, 267–286.

Simmons, Joseph P, Robyn A LeBoeuf, and Leif D Nelson, “The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors?,” *Journal of personality and social psychology*, 2010, *99* (6), 917.

Simonsohn, Uri, Joseph P Simmons, and Leif D Nelson, “Anchoring is not a false-positive: Maniadis, Tufano, and List’s (2014)’failure-to-replicate’is actually entirely consistent with the original,” *Working Paper*, 2014.

Slonim, Robert and Alvin E Roth, “Learning in high stakes ultimatum games: An experiment in the Slovak Republic,” *Econometrica*, 1998, pp. 569–596.

Smith, Vernon L and James M Walker, “Monetary rewards and decision cost in experimental economics,” *Economic Inquiry*, 1993, *31* (2), 245–261.

Spiliopoulos, Leonidas and Andreas Ortmann, “The BCD of response time analysis in experimental economics,” *Experimental economics*, 2018, *21* (2), 383–433.

Stanovich, Keith E and Richard F West, “Individual differences in reasoning: Implications for the rationality debate?,” *Behavioral and brain sciences*, 2000, *23* (5), 645–665.

Thaler, Richard H, “The psychology and economics conference handbook: Comments on Simon, on Einhorn and Hogarth, and on Tversky and Kahneman,” *The Journal of Business*, 1986, *59* (4), S279–S284.

Toplak, Maggie E, Richard F West, and Keith E Stanovich, "The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks," *Memory & cognition*, 2011, *39* (7), 1275.

Tversky, Amos and Daniel Kahneman, "Judgment under uncertainty: Heuristics and biases.," *Science*, 1974, *185*, 1124–1131.

Wilson, Timothy D, Christopher E Houston, Kathryn M Etling, and Nancy Brekke, "A new look at anchoring effects: basic anchoring and its antecedents.," *Journal of Experimental Psychology: General*, 1996, *125* (4), 387.

Wright, William F and Urton Anderson, "Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment," *Organizational Behavior and Human Decision Processes*, 1989, *44* (1), 68–82.

Yoon, Sangsuk, Nathan M Fong, and Angelika Dimoka, "The robustness of anchoring effects on preferential judgments.," *Judgment & Decision Making*, 2019, *14* (4).

ONLINE APPENDIX

A Additional Summary of Related Literature

Base-rate neglect. The results are mixed. Arkes et al. (1986) gave their participants either cash for each correct judgment ($.10/correct), a cash award ($5) for being the best judge in their group, or no monetary incentive. The incentives groups had fewer correct judgments than did the no-incentive group. Nelson et al. (1990) incentivized participants ($50 reward to the best answer), but they did not perform better than those in the control group. Similarly, Goodie and Fantino (1995) paid participants 10 cents for each point earned, with the top two earners receiving a bonus of $25 at the end. Incentives did not alleviate the base rate neglect error. By contrast, Klein (2001) told participants the top 30% would be entered into a lottery for $100, and found the incentivized participants performed better than the non-incentivized participants.

Wason task. Only a few experiments study the effect of incentives on performance in the task. Jones and Sugden (2001) paid their participants but did not vary the level of incentives. Behavior was closer to Bayesian rationality than in many no-incentives selection task experiments.

Anchoring. Tversky and Kahneman (1974) found that payment for accuracy did not reduce anchoring, as did Wilson et al. (1996). By contrast, Wright and Anderson (1989) and Epley and Gilovich (2005) and Epley et al. (2004) found that incentives for accuracy reduced anchoring. Simmons et al. (2010) showed that accuracy motivation through the use of incentives failed to increase the gap between anchors and final estimates when people were uncertain about the direction of adjustment, but increased anchor-estimate gaps when people were certain about the direction.

CRT. Borghans et al. (2008) tested the effect of incentives on performance on 10 problems, one of which was the CRT "bat and ball" (Frederick, 2005). He crossed time constraint (no time constraint, 60 seconds, or 30 seconds) with incentive (no pay, €0.10, €0.40, or €1.00 for each correct answer). Higher incentives increased time investment in answering the questions. In turn, CRT scores were higher for any level of incentive pay except in the case of the short time limit of 30 seconds. Yet, in a meta-analysis of CRT studies, Brañas-Garza et al. (2019) found that "paying subject for correct answers on the CRT does not increase performance levels."

Standardized tests and academic performance. Fryer Jr (2011) finds no effect of incentives of $30 to fourth graders and $60 to seventh graders on either math or reading test scores. Bettinger (2012) finds that incentives of $20 have a significant impact on third through sixth graders' performance in math but no impact on reading, social science, or science. Levitt et al. (2016) find that while an incentive of $20 delivered immediately improves test performance, both rewards delivered with a one month delay fail and a lower incentive of $10 fail to have an impact. Other studies that announced incentives immediately before the test and provided rewards with a delay also find mixed evidence. O'Neil et al. (1995, 2005) report that delayed incentives can improve eighth grade test scores but fail to have an impact on twelfth grade test scores, even at high incentives of $100. Similarly, Baumert and Demmrich (2001) find no effects of incentives on ninth grade test scores. So overall, the effects of incentives on test performance are quite mixed (see Gneezy et al., 2011, for a review). In cases where a positive effect is reported, scores improve about 0.10 (Levitt et al., 2016) to 0.15 Bettinger (2012) standard deviations. For comparison, in our CRT task, the only task in which we observe a significant improvement in performance, the score improves about 0.20 standard deviations when going from standard stakes to very high stakes.

B Experimental Questions

B.1 Base rate neglect

BRN 1. Suppose the Kenyan police set up a road block to test drivers for drunk driving. They stop every bus and taxi driver that passes with an Alcoblow test.

The Alcoblow shows a red light when it detects that the person is drunk, and a green light when it detects the person is not drunk. However, the test is not completely reliable and can give a wrong indication.

Now suppose that 100 out of every 10,000 drivers who are stopped at a routine police control are actually drunk.

When ACTUAL drunk drivers are tested, the Alcoblow shows a red light for 55 of those 100 drunk drivers.

But, of the remaining 9,900 drivers who are NOT drunk, the Alcoblow test also shows a red light for 500 of these 9,900 non-drunk drivers.

To make this very clear, a diagram presenting this information is shown below:

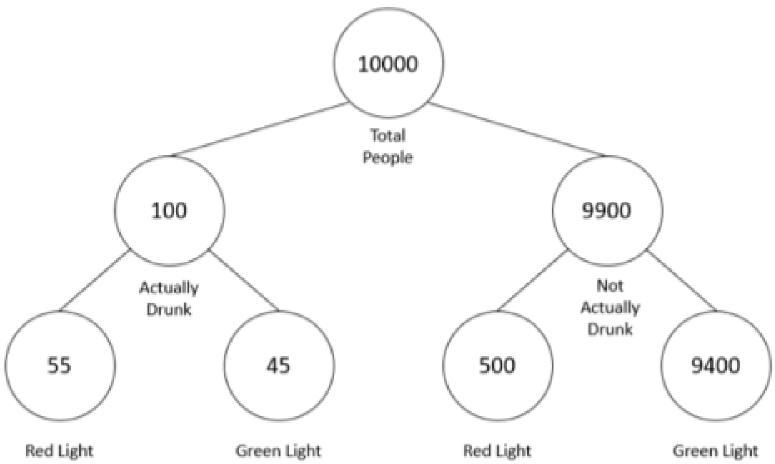


Figure 7: Diagram used to illustrate the “intuitive” car accidents base rate neglect task

Now suppose that in a new sample of 2000 drivers, the Alcoblow test showed a red light for 100 drivers. Of these 100 drivers, how many drivers do you expect to have actually been drunk? (options 0 to 100 in steps of 1)

BRN 2. 1% of women screened at age 40 have breast cancer.

If a woman has breast cancer, the probability is 80% that she will get a positive mammography.

If a woman does not have breast cancer, the probability is 9.6% that she will get a positive mammography.

A 40-year-old woman had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? (answer in %, options 0 to 100% in steps of 1)

BRN 3. 1% of drivers who are stopped by routine police control are drunk.

If a driver is drunk, the probability is 55% that the driver will test positive on an alcohol test.

If a driver is not drunk, the probability is 5.1% that the driver will test positive on an alcohol test.

Suppose a driver has tested positive on an alcohol test in a routine police control. What is the probability that the driver was actually drunk? (answer in %, options 0 to 100% in steps of 1)

BRN 4. Suppose a Kenyan medical center routinely tests women at age 40 to determine if they have breast cancer. They use an x-ray machine for this. The machine produces images, and a medical expert examines the images.

If the medical expert detects breast cancer on the images, the expert brings bad news to the woman. If the medical expert does not detect breast cancer on the images, the expert brings good news to the woman. However, the images are not always clear and the expert can reach the wrong conclusion.

Now suppose that 10 out of every 1,000 women at age 40 who get routinely tested actually have breast cancer.

When the women with actual breast cancer get tested, the expert brings bad news to 8 out of these 10 women.

But, of the remaining 990 women who do NOT have breast cancer, the expert also gives bad news to 95 out of these 990 women.

To make this very clear, a diagram presenting this information is shown below: [see Figure in Section 2.1.1]

Now suppose that in a new sample of 1000 women, 100 women at age 40 received bad news in the routine test. Of these 100 women, how many women do you expect to actually have breast cancer? (options 0 to 100 in steps of 1)

B.2 Wason selection task

Wason 1. Suppose you have a friend who says he has a special deck of cards. His special deck of cards all have numbers (odd or even) on one side and colors (brown or green) on the other side. Suppose that the 4 cards from his deck are shown below.

Your friend also claims that in his special deck of cards, even numbered cards are never brown on the other side. He says:

“In my deck of cards, all of the cards with an even number on one side are green on the other.”

Unfortunately, your friend doesn’t always tell the truth, and your job is to figure out whether he is telling the truth or lying about his statement.

From the cards below, turn over only those card(s) that can be helpful in determining whether your friend is telling the truth or lying. Do not turn over those cards that cannot help you in determining whether he is telling the truth or lying.

Select the card(s) you want to turn over: [see Figure in Section 2.1.2]

Wason 2. Suppose you have a friend who says he has a special deck of cards. His special deck of cards all have numbers (odd or even) on one side and colors (brown or blue) on the other side. Suppose that the 4 cards from his deck are shown below.

Your friend also claims that in his special deck of cards, even numbered cards are never brown on the other side. He says:

“In my deck of cards, all of the cards with an even number on one side are blue on the other.”

Unfortunately, your friend doesn’t always tell the truth, and your job is to figure out whether he is telling the truth or lying about his statement.

From the cards below, turn over only those card(s) that can be helpful in determining whether your friend is telling the truth or lying. Do not turn over those cards that cannot help you in determining whether he is telling the truth or lying.

Select the card(s) you want to turn over: [cards shown are 4 Card, 9 Card, Blue Card, Brown Card]

Wason 3. You are in charge of enforcing alcohol laws at a bar. You will lose your job unless you enforce the following rule:

If a person drinks an alcoholic drink, then they must be at least 18 years old.

The cards below have information about four people sitting at a table in your bar. Each card represents one person. One side of a card tells what a person is drinking, and the other side of the card tells that person’s age.

In order to enforce the law, which of the card(s) below would you definitely need to turn over? Indicate only those card(s) you definitely need to turn over to see if any of these people are breaking the law.

Select the card(s) you want to turn over: [see Figure in Section 2.1.2]

Wason 4. You are in charge of enforcing alcohol laws at a bar. You will lose your job unless you enforce the following rule:

If a person drinks an alcoholic drink, then they must be at least 18 years old.

The cards below have information about four people sitting at a table in your bar. Each card represents one person. One side of a card tells what a person is drinking, and the other side of the card tells that person’s age.

In order to enforce the law, which of the card(s) below would you definitely need to turn over? Indicate only those card(s) you definitely need to turn over to see if any of these people are breaking the law.

Select the card(s) you want to turn over [cards shown are Drinking Wine, Drinking Juice, 17 Years Old, 22 Years Old]:

B.3 Anchoring

For this task, please first do the following. Take the last digit of your phone number.

- If it is 4 or less, please enter below the first two digits of your year of birth.
- If it is 5 or above, please enter below 100 minus the first two digits of your year of birth.

Enter the two digits:

In this task, you will be asked to make two estimates. Each time, you will be asked to

1. assess whether you think the quantity is greater than or less than the two digits that were just generated from your year of birth
2. give an estimate of the quantity (a number between 0 and 100). Your answer will be counted as correct if it is no more than 2 away from the actual number.

Anchoring 1. In 1911, pilot Calbraith Perry Rodgers completed the first airplane trip across the continental U.S., taking off from Long Island, New York and landing in Pasadena, California.

Did the trip take more than or less than ANCHOR days?

How many days did it take Rodgers to complete the trip? (options 0 to 100 in steps of 1)

Anchoring 2. Is the time (in minutes) it takes for light to travel from the Sun to the planet Jupiter more than or less than ANCHOR minutes?

How many minutes does it take light to travel from the Sun to the planet Jupiter? (options 0 to 100 in steps of 1)

Anchoring 3. Is the population of Uzbekistan as of 2018 greater than or less than ANCHOR million?

What is the population of Uzbekistan in millions of people as of 2018? (options 0 to 100 in steps of 1)

Anchoring 4. Is the weight (in hundreds of tons) of the Eiffel Tower's metal structure more than or less than ANCHOR hundred tons?

What is the weight (in hundreds of tons) of the Eiffel Tower's metal structure? (options 0 to 100 in steps of 1)

B.4 Cognitive Reflection Test

CRT 1. A bat and a ball cost 110 KSh in total. The bat costs 100 KSh more than the ball. How much does the ball cost? (Please provide your answer in KSh)

CRT 2. A pencil and an eraser cost 110 KSh in total. The pencil costs 100 KSh more than the eraser. How much does the eraser cost? (Please provide your answer in KSh)

CRT 3. It takes 5 nurses 5 minutes to measure the blood pressure of 5 patients. How long would it take 10 nurses to measure the blood pressure of 10 patients? (Please provide your answer in minutes)

CRT 4. It takes 5 workers 5 minutes to pack 5 boxes. How long would it take 10 workers to pack 10 boxes? (Please provide your answer in minutes)

CRT 5. It takes 6 nurses 6 minutes to measure the blood pressure of 6 patients. How long would it take 12 nurses to measure the blood pressure of 12 patients? (Please provide your answer in minutes)

CRT 6. It takes 6 workers 6 minutes to pack 6 boxes. How long would it take 12 workers to pack 12 boxes? (Please provide your answer in minutes)

C Additional Tables

Table 5: Participants' background characteristics across the two treatment conditions

	Flat-standard lab	Flat-high	p-value test (1)=(2)
Age (years)	22.1	22.1	0.740
Female	0.42	0.46	0.191
GPA	4.6	4.7	0.227
Raven score (0-10)	3.2	3.4	0.078
Monthly consumption level (1,000 KSh)	23.5	13.7	0.335
Monthly income (1,000 KSh)	16.1	17.1	0.363
N	636	600	

Notes. The large difference in average consumption across treatments is driven by one extreme outlier (perhaps a typo by the participant). Median consumption is identical across treatments (10,000 KSh).

Table 6: Response times across incentive conditions: Log-transformed response times

	Dependent variable: Log (1+Response time [seconds])						
Omitted category:		Base rate neglect		Contingent reasoning			All
Standard incentives	CRT	Abstract	Intuitive	Abstract	Intuitive	Anchoring	tasks
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 if *No incentives*	-0.051	-0.040	0.00042	-0.068	-0.015	-0.052	-0.041
	(0.05)	(0.07)	(0.06)	(0.05)	(0.05)	(0.05)	(0.02)
1 if *High incentives*	0.26^{**}	0.32^{**}	0.35^{**}	0.21^{**}	0.23^{**}	0.17^{*}	0.24^{**}
	(0.07)	(0.09)	(0.07)	(0.07)	(0.05)	(0.07)	(0.03)
Constant	4.80^{**}	5.49^{**}	5.75^{**}	5.05^{**}	4.58^{**}	4.36^{**}	4.33^{**}
	(0.04)	(0.06)	(0.05)	(0.04)	(0.04)	(0.04)	(0.03)
Task type FE	No	No	No	No	No	No	Yes
Observations	1240	618	618	619	619	1230	4944
R^2	0.03	0.04	0.05	0.04	0.04	0.02	0.37

Notes. OLS estimates, standard errors (clustered at subject level) in parentheses. Omitted category: standard incentives. The dependent variable is ln(1+RT). * $p < 0.05$, ** $p < 0.01$.

Table 7: Non-parametric tests for response time differences across incentive levels

Task	High vs. no	High vs. standard	Standard vs. no
CRT	< 0.01	< 0.01	0.38
Abstract BRN	< 0.01	< 0.01	0.40
Intuitive BRN	< 0.01	< 0.01	0.53
Abstract Wason	< 0.01	< 0.01	0.04
Intuitive Wason	< 0.01	< 0.01	0.61
Anchoring	< 0.01	< 0.01	0.39

Notes. P-values for Wilcoxon ranksum tests of response times between incentive levels. * $p < 0.05$, ** $p < 0.01$.

Table 8: Non-parametric tests for performance differences across incentive levels

Task	High vs. no	High vs. standard	Standard vs. no
CRT	< 0.01	0.01	0.23
Abstract BRN	0.38	0.44	0.99
Intuitive BRN	0.08	0.15	0.95
Abstract Wason	0.57	0.25	0.05
Intuitive Wason	0.08	0.61	0.23

Notes. P-values for Wilcoxon ranksum tests of performance [0–1] between incentive levels. * $p < 0.05$, ** $p < 0.01$.

Table 9: Response times across incentive conditions: Pre-registered sample

	Dependent variable: Response time [seconds]						
Omitted category:		Base rate neglect		Contingent reasoning			All
Standard incentives	CRT	Abstract	Intuitive	Abstract	Intuitive	Anchoring	tasks
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 if *No incentives*	-0.49	-8.99	12.1	-14.9	0.98	-7.42	-3.26
	(10.98)	(21.96)	(25.12)	(8.09)	(4.96)	(6.21)	(5.76)
1 if *High incentives*	48.1**	133.6**	195.9**	52.9**	30.2**	25.0*	69.7**
	(14.89)	(35.04)	(43.55)	(13.78)	(6.61)	(10.15)	(9.84)
Constant	161.2**	308.6**	374.1**	177.9**	105.9**	95.1**	81.6**
	(8.59)	(16.87)	(18.16)	(6.58)	(3.78)	(5.48)	(5.36)
Task type FE	No	No	No	No	No	No	Yes
Observations	1144	569	569	572	572	1134	4560
R^2	0.02	0.05	0.06	0.07	0.05	0.02	0.29

Notes. OLS estimates, standard errors (clustered at subject level) in parentheses. Omitted category: standard incentives. The sample is restricted to the first 1,140 subjects who completed the experiment. * $p < 0.05$, ** $p < 0.01$.

Table 10: Performance by incentive level: Pre-registered sample

	Dependent variable: 1 if answer correct						Answer
Omitted category:		Base rate neglect		Contingent reasoning		Tasks	
Standard incentives	CRT	Abstract	Intuitive	Abstract	Intuitive	1–5	Anchoring
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 if *No incentives*	-0.030	0.0012	0.013	-0.074	-0.067	-0.031	6.60*
	(0.04)	(0.03)	(0.04)	(0.04)	(0.05)	(0.02)	(3.31)
1 if *High incentives*	0.10*	-0.018	0.099	-0.063	0.025	0.041	3.96
	(0.04)	(0.04)	(0.05)	(0.04)	(0.06)	(0.02)	(3.67)
Anchor							0.48**
							(0.05)
Anchor × 1 if *No incentives*							0.0029
							(0.07)
Anchor × 1 if *High incentives*							0.0051
							(0.08)
Constant	0.38**	0.11**	0.23**	0.20**	0.59**	0.11**	12.4**
	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.02)	(2.51)
Task type FE	No	No	No	No	No	Yes	No
Observations	1144	569	569	572	572	3426	1134
R^2	0.01	0.00	0.01	0.01	0.01	0.12	0.21

Notes. OLS estimates, robust standard errors (clustered at subject level) in parentheses. In columns (1)–(5), the dependent variable is a binary indicator for whether an answer is correct. In column (6), the outcome variable is the answer (between 0 and 100). Omitted category: standard incentives. The sample is restricted to the first 1,140 subjects who completed the experiment. * $p < 0.05$, ** $p < 0.01$.

Table 11: Performance by incentive level: Adding covariates

	Dependent variable:						
	1 if answer correct						Answer
Omitted category:		Base rate neglect		Contingent reasoning		Tasks	
Standard incentives	CRT	Abstract	Intuitive	Abstract	Intuitive	1–5	Anchoring
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 if *No incentives*	-0.052	0.0041	0.0060	-0.063	-0.063	-0.035*	5.99
	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.02)	(3.17)
1 if *High incentives*	0.10*	-0.026	0.082	-0.045	0.016	0.037	4.83
	(0.04)	(0.03)	(0.05)	(0.04)	(0.06)	(0.02)	(3.60)
Age	-0.026**	-0.0080	-0.020**	-0.016**	-0.013	-0.019**	0.055
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.45)
1 if male	0.13**	0.014	0.11**	0.012	-0.098*	0.051**	-5.37**
	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.01)	(1.79)
1 if above median income	-0.027	0.020	0.030	0.0036	-0.11**	-0.017	-0.23
	(0.03)	(0.02)	(0.03)	(0.03)	(0.04)	(0.01)	(1.76)
Cognitive skills [z-score]	0.079**	0.0074	0.026	0.0048	0.079**	0.044**	-2.51*
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.99)
1 if STEM major	0.013	-0.028	0.027	0.069*	0.013	0.021	-0.29
	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.02)	(2.00)
Anchor							0.50**
							(0.05)
Anchor × 1 if *No incentives*							-0.0036
							(0.06)
Anchor × 1 if *High incentives*							-0.0044
							(0.08)
Constant	0.83**	0.24	0.66**	0.53**	0.98**	0.94**	13.9
	(0.21)	(0.14)	(0.17)	(0.13)	(0.22)	(0.09)	(10.28)
Question FE	Yes	Yes	Yes	Yes	Yes	Yes	No
Observations	1240	618	618	619	619	3714	1230
R^2	0.09	0.02	0.05	0.03	0.05	0.15	0.24

Notes. OLS estimates, robust standard errors (clustered at subject level) in parentheses. In columns (1)-(5), the dependent variable is a binary indicator for whether an answer is correct. In column (6), the outcome variable is the answer (between 0 and 100). Omitted category: standard incentives. * $p < 0.05$, ** $p < 0.01$.

Table 12: Performance by incentive level: Additional analyses for base rate neglect

	Dependent variable:							
	|Answer - Bayesian posterior|				1 if |Answer - Bayesian posterior| ≤ 2			
Omitted category:	Base rate neglect							
Standard incentives	Abstract		Intuitive		Abstract		Intuitive	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1 if *No incentives*	2.85	2.40	-0.048	-0.60	0.027	0.028	0.015	0.027
	(2.57)	(2.60)	(3.07)	(3.09)	(0.02)	(0.02)	(0.04)	(0.04)
1 if *High incentives*	-0.12	-0.55	-4.09	-4.34	0.022	0.023	0.060	0.068
	(2.99)	(3.02)	(3.56)	(3.57)	(0.03)	(0.03)	(0.05)	(0.05)
Age		-0.11		1.24*		0.00080		-0.020**
		(0.48)		(0.60)		(0.00)		(0.01)
1 if male		-2.36		-5.28*		0.0049		0.11**
		(2.22)		(2.61)		(0.02)		(0.03)
1 if above median income		-1.79		-0.73		0.0019		0.056
		(2.19)		(2.55)		(0.02)		(0.03)
Cognitive skills [z-score]		-0.35		-1.75		0.0096		0.033
		(1.24)		(1.46)		(0.01)		(0.02)
1 if STEM major		1.95		2.62		-0.0098		0.019
		(2.46)		(2.97)		(0.02)		(0.04)
Constant	36.2**	40.7**	42.8**	18.4	0.0036	-0.017	0.25**	0.59**
	(2.06)	(10.67)	(2.96)	(13.53)	(0.02)	(0.10)	(0.04)	(0.17)
Question FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	618	618	618	618	618	618	618	618
R^2	0.04	0.04	0.03	0.04	0.03	0.03	0.01	0.05

Notes. OLS estimates, robust standard errors (clustered at subject level) in parentheses. The dependent variable is the absolute difference between a subject's response and the Bayesian posterior. Omitted category: standard incentives. * $p < 0.05$, ** $p < 0.01$.

Table 13: Specific mistakes

	Dependent variable: Gives Intuitive Answer							
Omitted category:		Base rate neglect		Contingent reasoning				
Standard incentives	CRT	Abstract	Intuitive	Abstract		Intuitive		Anchoring
	Intuitive response	Report diagnosticity		Turns only P	Turns P and Q	Turns only P	Turns P and Q	Guess = anchor
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1 if *No incentives*	0.06	-0.05	0.02	-0.03	0.02	-0.04	0.03	0.03
	(0.04)	(0.05)	(0.05)	(0.03)	(0.05)	(0.04)	(0.02)	(0.02)
1 if *High incentives*	-0.09*	-0.11*	0.03	-0.04	0.04	-0.02	0.02	0.01
	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)	(0.04)	(0.02)	(0.02)
Constant	0.48**	0.38**	0.31**	0.15**	0.30**	0.17**	0.04*	0.04**
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.02)	(0.01)
Observations	1240	618	618	619	619	619	619	1230
R^2	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00

Notes. OLS estimates, robust standard errors (clustered at subject level) in parentheses. The dependent variable is a binary indicator that equals one if the subject's response corresponds to the "intuitive response." In the CRT, this is given by the well-known intuitive responses. In BRN, it is given by the signal diagnosticity (the conditional probability P(event|state)). For Wason, we consider two types of intuitive responses, for a given rule of the type "if P then Q." The first is that subjects turn over only the "P" card (the "8" in the example in Section 2). The second is that subjects turn over "P" and "Q" (the "8" and "green" in the example in Section 2). In anchoring, the intuitive response is given by reporting the anchor. * $p < 0.05$, ** $p < 0.01$.

Table 14: Confidence by incentive level

	Dependent variable: Confidence [0–7]					
Omitted category:		Base rate neglect		Contingent reasoning		
Standard incentives	CRT	Abstract	Intuitive	Abstract	Intuitive	Anchoring
	(1)	(2)	(3)	(4)	(5)	(6)
1 if *No incentives*	-0.11	0.085	-0.035	0.051	-0.058	0.15
	(0.09)	(0.13)	(0.13)	(0.11)	(0.09)	(0.15)
1 if *High incentives*	0.12	-0.032	-0.16	0.23	-0.0074	0.21
	(0.10)	(0.17)	(0.16)	(0.13)	(0.10)	(0.18)
Constant	6.22**	5.25**	5.45**	5.89**	6.43**	4.49**
	(0.07)	(0.11)	(0.11)	(0.09)	(0.07)	(0.12)
Observations	1240	618	618	619	619	1230
R^2	0.01	0.00	0.00	0.01	0.00	0.00

Notes. OLS estimates, robust standard errors (clustered at subject level) in parentheses. The dependent variable is self-reported confidence. Omitted category: standard incentives. * $p < 0.05$, ** $p < 0.01$.

Table 15: Response times across incentive conditions: Restricting sample to correct recall of incentive level

	Dependent variable: Response time [seconds]						
Omitted category:		Base rate neglect		Contingent reasoning			All
Standard incentives	CRT	Abstract	Intuitive	Abstract	Intuitive	Anchoring	tasks
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 if *No incentives*	-2.50	-5.87	10.6	-13.0	-0.94	-13.7	-5.32
	(11.75)	(26.83)	(29.32)	(9.72)	(5.56)	(7.74)	(6.49)
1 if *High incentives*	43.3**	107.4**	189.1**	50.2**	30.3**	28.3*	62.8**
	(16.61)	(39.42)	(50.01)	(15.17)	(8.06)	(13.68)	(10.53)
Constant	155.3**	313.7**	375.7**	177.9**	107.0**	98.5**	86.2**
	(9.14)	(20.42)	(20.28)	(8.03)	(4.36)	(7.06)	(6.09)
Task type FE	No	No	No	No	No	No	Yes
Observations	832	386	386	421	421	814	3260
R^2	0.02	0.03	0.06	0.07	0.06	0.04	0.31

Notes. OLS estimates, standard errors (clustered at subject level) in parentheses. Omitted category: standard incentives. The sample is restricted to observations for which a subject recalled exactly the correct incentive amount in the post-experimental questionnaire. * $p < 0.05$, ** $p < 0.01$.

Table 16: Performance by incentive level: Restricting sample to correct recall of incentive level

	Dependent variable:						
	1 if answer correct						Answer
Omitted category:		Base rate neglect		Contingent reasoning		Tasks	
Standard incentives	CRT	Abstract	Intuitive	Abstract	Intuitive	1–5	Anchoring
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1 if *No incentives*	-0.068	0.022	-0.019	-0.039	0.0082	-0.028	6.10
	(0.04)	(0.04)	(0.06)	(0.04)	(0.06)	(0.02)	(4.00)
1 if *High incentives*	0.10^{*}	-0.025	0.053	-0.043	0.12	0.053^{*}	4.16
	(0.05)	(0.04)	(0.07)	(0.05)	(0.07)	(0.03)	(4.50)
Anchor							0.52^{**}
							(0.06)
Anchor × 1 if *No incentives*							-0.048
							(0.08)
Anchor × 1 if *High incentives*							-0.013
							(0.10)
Constant	0.41^{**}	0.11^{**}	0.30^{**}	0.18^{**}	0.57^{**}	0.12^{**}	12.4^{**}
	(0.03)	(0.03)	(0.05)	(0.04)	(0.05)	(0.02)	(3.19)
Task type FE	No	No	No	No	No	Yes	No
Observations	832	386	386	421	421	2446	814
R^2	0.02	0.00	0.00	0.00	0.01	0.13	0.21

Notes. OLS estimates, robust standard errors (clustered at subject level) in parentheses. In columns (1)–(5), the dependent variable is a binary indicator for whether an answer is correct. In column (6), the outcome variable is the answer (between 0 and 100). Omitted category: standard incentives. The sample is restricted to observations for which a subject recalled exactly the correct incentive amount in the post-experimental questionnaire. * $p < 0.05$, ** $p < 0.01$.

D Additional Figures

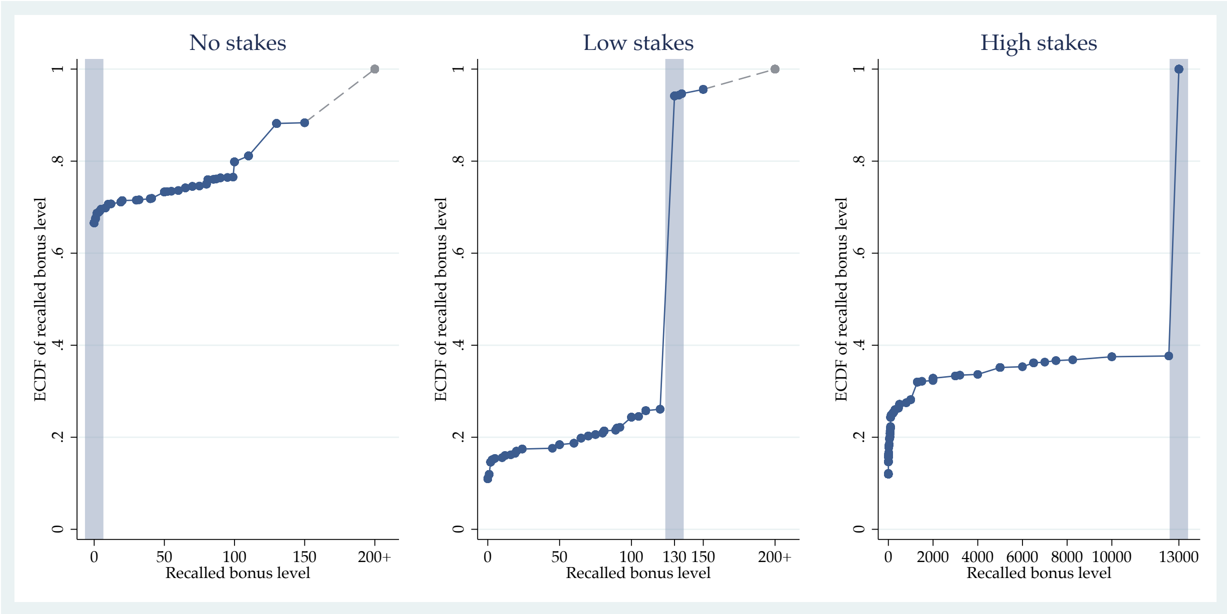


Figure 8: Empirical CDF of recalled bonus amount by incentive level.

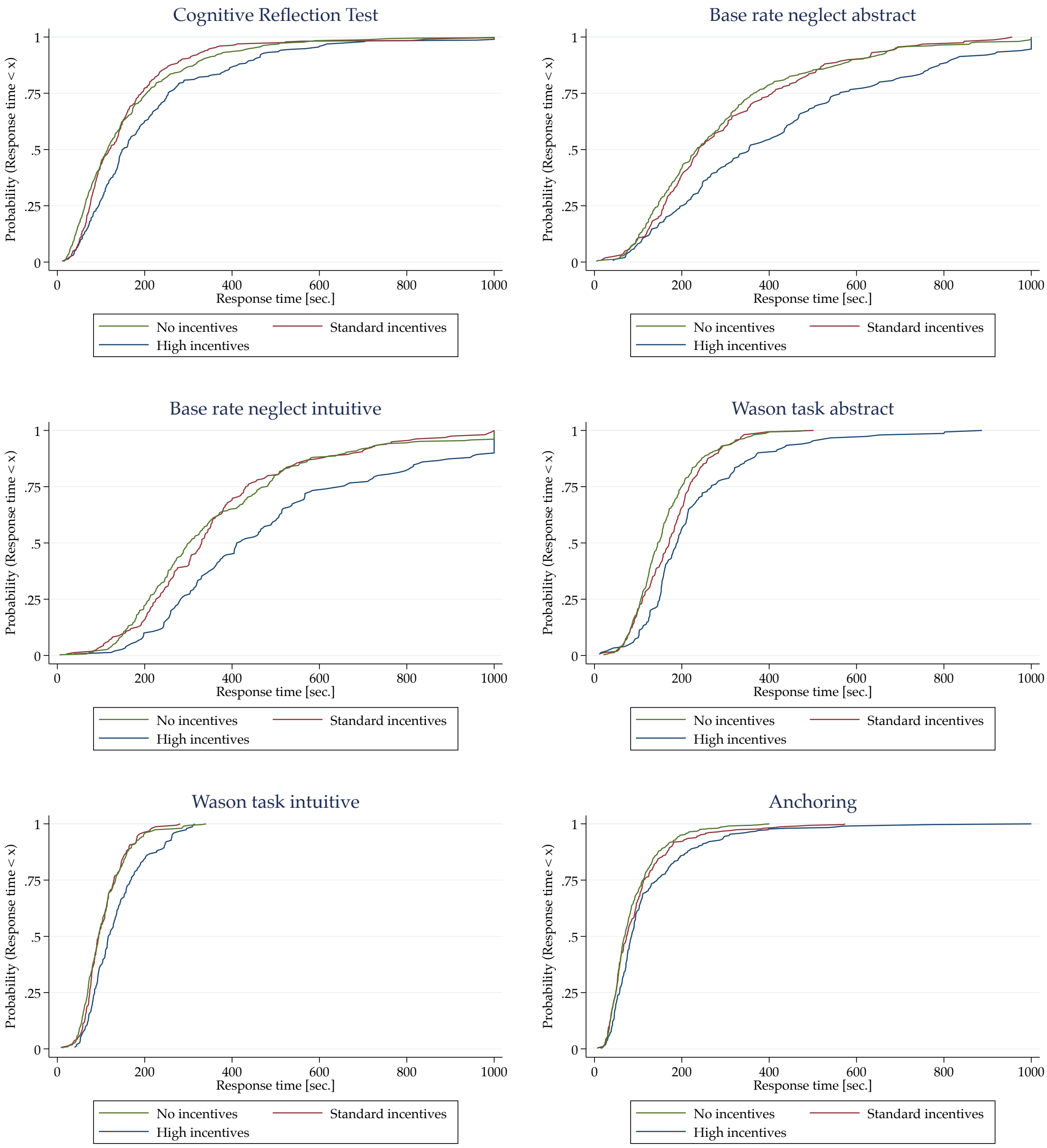


Figure 9: Empirical CDFs of response times. For the purposes of this figure, response times are winsorized at 1,000 seconds, which corresponds approximately to the 99th percentile across all tasks and subjects.

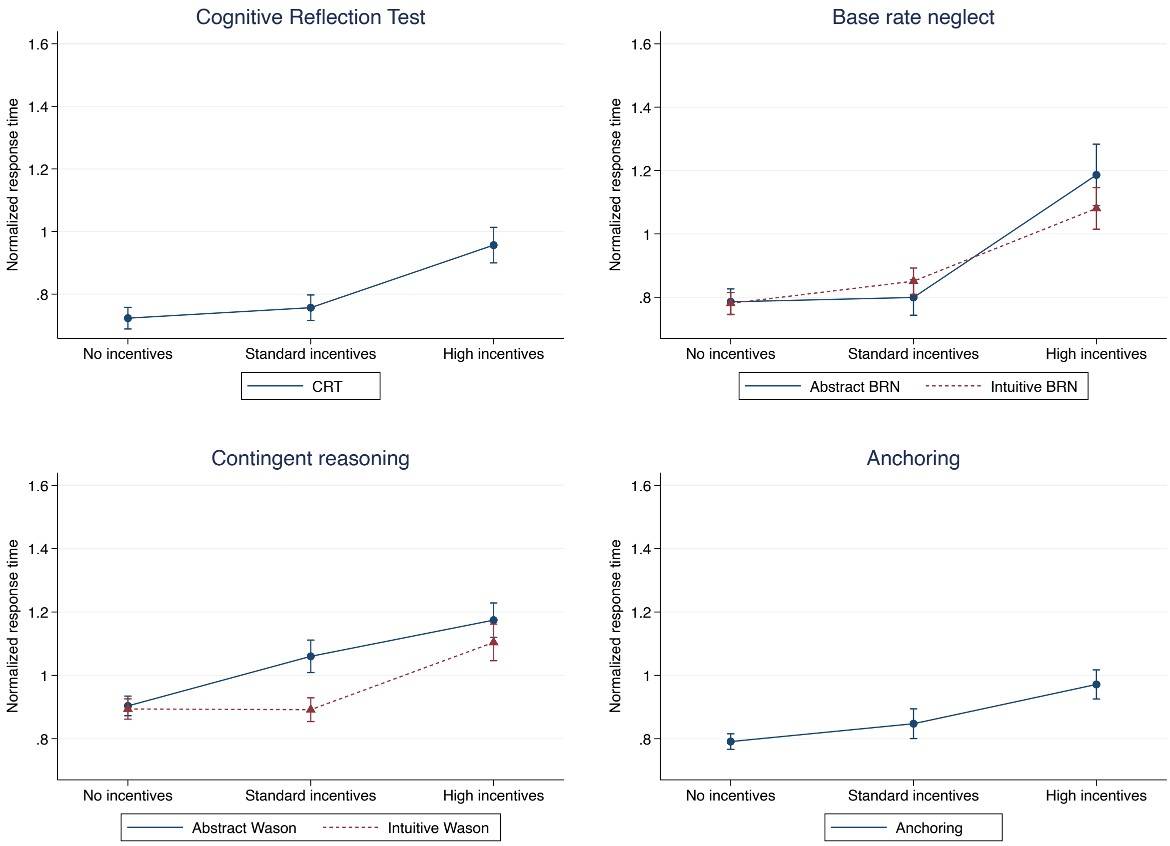


Figure 10: Median normalized response times across incentive conditions. Response times are normalized relative to the no incentive condition: for each cognitive bias, we divide observed response times by the average response time in the no incentive condition. Error bars indicate +/- 1 s.e.

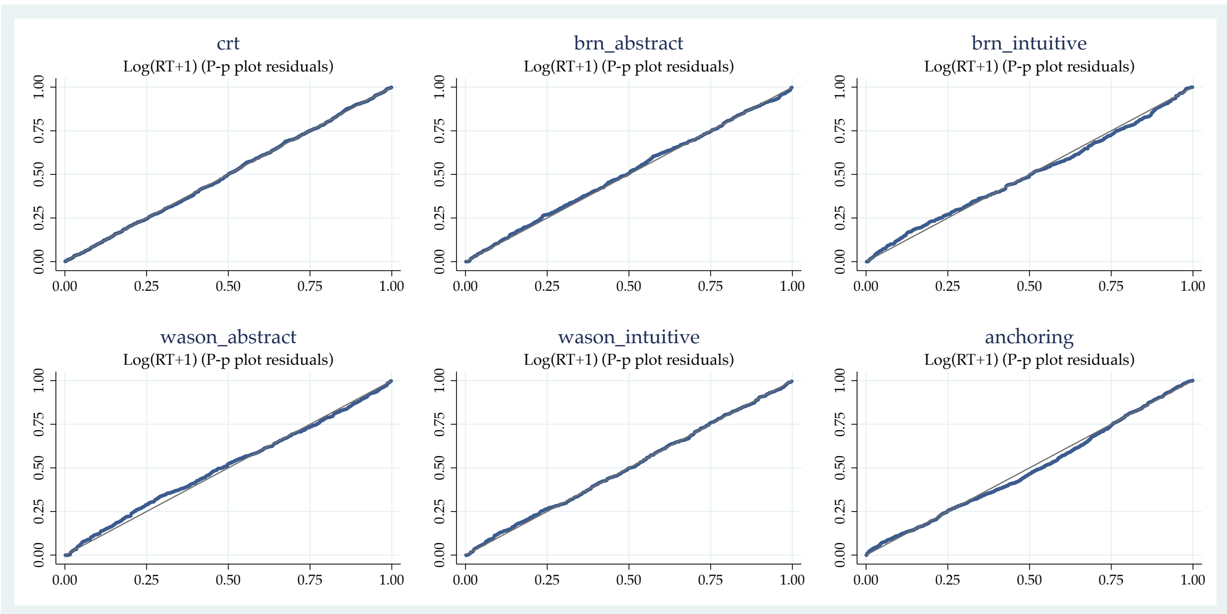


Figure 11: p-p plot of residuals from a regression of ln(1+RT) on incentive condition indicators.

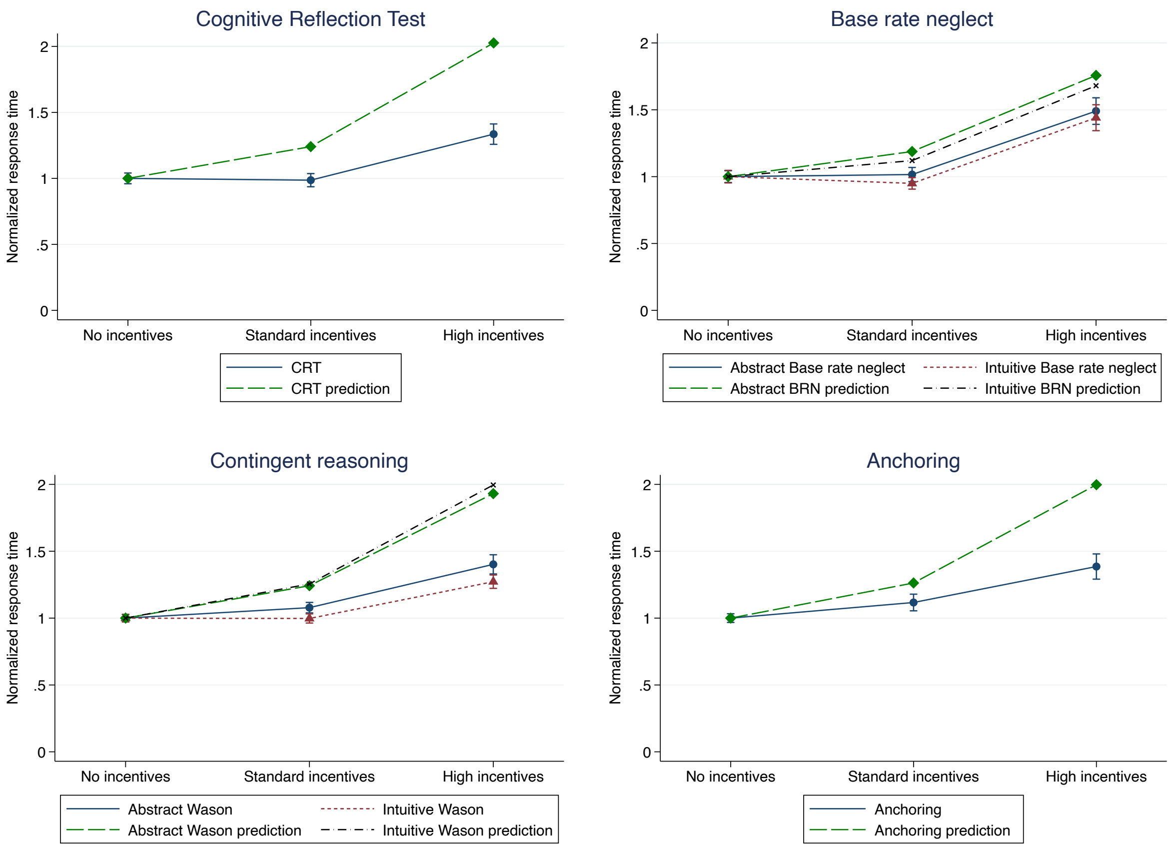


Figure 12: Average response time by incentive level relative to average expert predictions. Error bars on response times indicate +/- 1 s.e.

E Comparison with Expert Forecasts

This appendix characterizes the differences between the distributions of participant responses and expert forecasts. In Figure 13, we plot the empirical distribution of our expert predictions alongside the posterior distribution of actual performance for each task. Each posterior distribution is centered around actual performance and has a standard deviation equal to the corresponding standard error in Figure 5. For all tasks, there is excess mass on the right tail of the posterior distribution of performance.

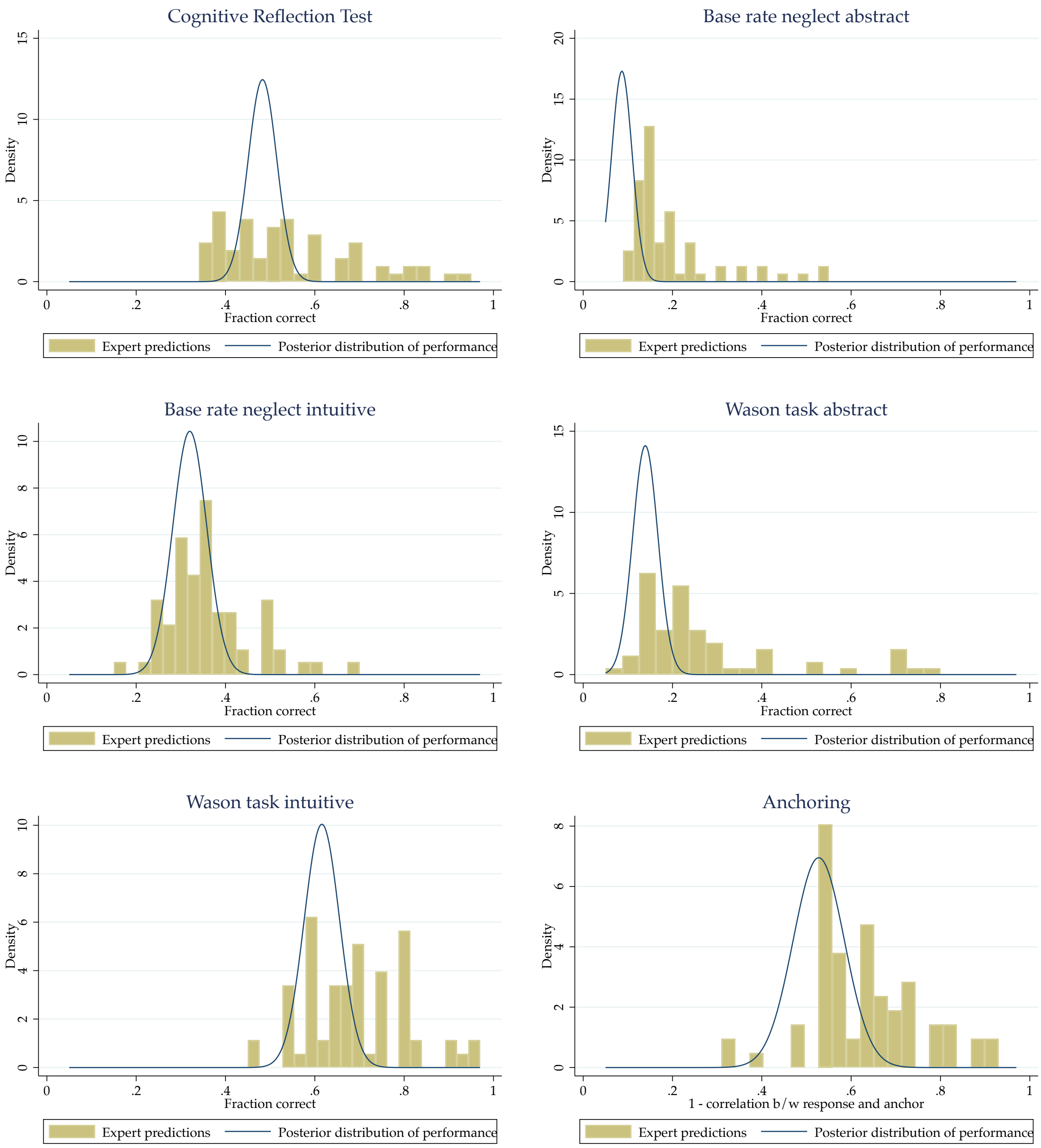


Figure 13: Empirical distributions of expert predictions and posterior distributions of actual performance.

F Experimental Instructions and Decision Screens

F.1 General Instructions for Part 1

Instructions

Thank you for participating in this experiment. Please read the instructions below carefully.

The experiment consists of several parts, each of which is independent.

At completion of the study, you will receive 450 KSh via Mpesa for your participation. 50 KSh will be given for arriving on time. This will be paid within 24 hours.

These payments will be made regardless of whether your answers are correct or not.

If you have a question at any point during the experiment, please raise your hand and one of us will come to your desk.

What follows are the instructions for the first part. Once you are ready to start, please click the “Start” button to proceed.

Part 1

We will ask you two questions on the upcoming screens. Please answer them to the best of your ability.

F.2 General Instructions for Part 2

You have completed **Part 1** of the study.

Part 2 will now begin.

Part 2

We will ask you two questions on the upcoming screens. Please answer them to the best of your ability.

Please remember that you will earn a guaranteed show-up fee of 450 KSh.

While there was no opportunity to earn a bonus in the previous part, you will now have the opportunity to earn a <u>bonus payment of 13000 KSh</u> (thirteen thousand KSh) if your answer is correct.

One of the questions will be randomly selected for payment. If your answer to that question is correct, you will receive the bonus payment of 13000 KSh (thirteen thousand KSh). If your answer to that question is incorrect, you will not earn the bonus payment.

This payment will be sent via Mpesa and will arrive within 72 hours.

Remember, you will now have the opportunity to earn a <u>bonus payment of 13000 KSh</u> (thirteen thousand KSh) if your answer is correct.

→

F.3 Sample Task Instructions

Recall that different participants solved different sets of tasks for each bias. The full set of tasks is outlined in Appendix B, and our process for randomization is described in Section 2.

F.3.1 Cognitive Reflection Test

A pencil and an eraser cost 110 KSh in total. The pencil costs 100 KSh more than the eraser. How much does the eraser cost? (Please provide your answer in KSh)

It takes 5 workers 5 minutes to pack 5 boxes. How long would it take 10 workers to pack 10 boxes? (Please provide your answer in minutes)

→

F.3.2 Base rate neglect

In this task, we will ask you to make two estimates. Your answers can range from 0 to 100. We will compare each answer to the estimate of an expert. Your answer will be counted as correct if it is no more than 2 away from the expert's estimate.

Suppose the Kenyan police set up a road block to test drivers for drunk driving. They stop every bus and taxi driver that passes with an Alcoblow test.

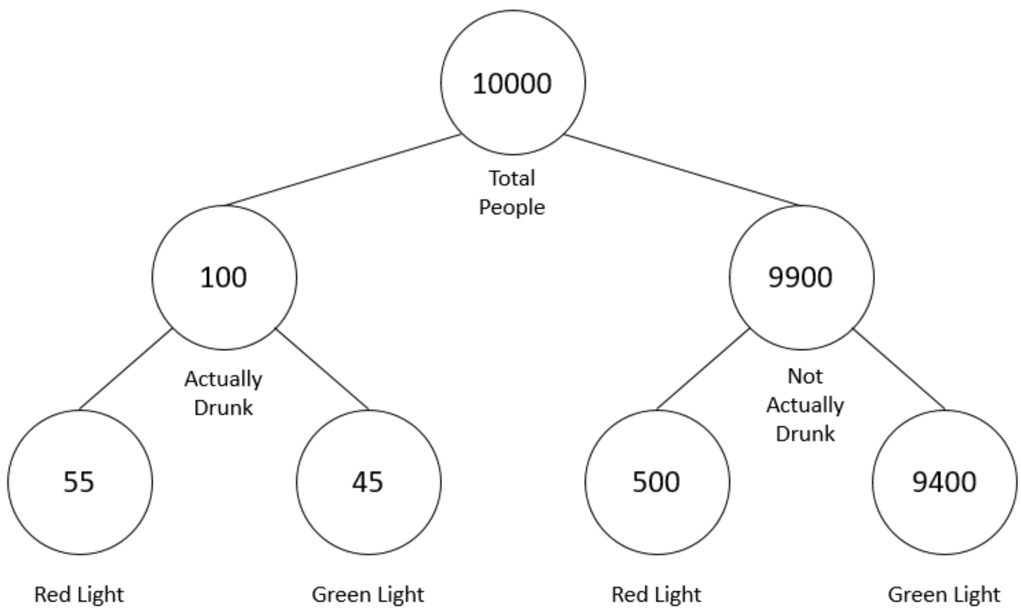
The Alcoblow shows a red light when it detects that the person is drunk, and a green light when it detects the person is not drunk. However, the test is not completely reliable and can give a wrong indication.

Now suppose that 100 out of every 10,000 drivers who are stopped at a routine police control are actually drunk.

When ACTUAL drunk drivers are tested, the Alcoblow shows a red light for 55 of those 100 drunk drivers.

But, of the remaining 9,900 drivers who are NOT drunk, the Alcoblow test also shows a red light for 500 of these 9,900 non-drunk drivers.

To make this very clear, a diagram presenting this information is shown below:



Now suppose that in a new sample of 2000 drivers, the Alcoblow test showed a red light for 100 drivers. Of these 100 drivers, how many drivers do you expect to have actually been drunk? **(options 0 to 100 in steps of 1)**

1% of women screened at age 40 have breast cancer.

If a woman has breast cancer, the probability is 80% that she will get a positive mammography.

If a woman does not have breast cancer, the probability is 9.6% that she will get a positive mammography.

A 40-year-old woman had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? **(answer in %, options 0 to 100% in steps of 1)**

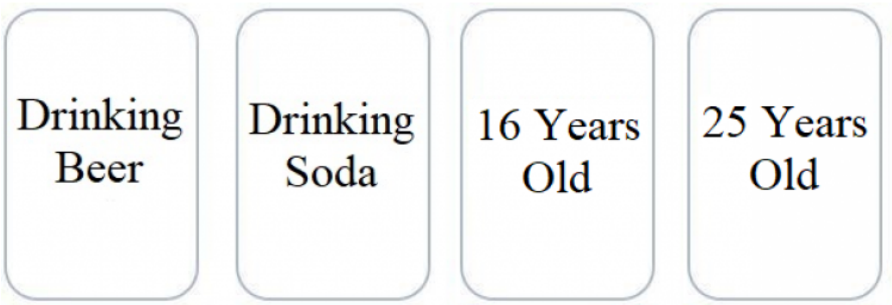
F.3.3 Wason selection task

You are in charge of enforcing alcohol laws at a bar. You will lose your job unless you enforce the following rule:

If a person drinks an alcoholic drink, then they must be at least 18 years old.

The cards below have information about four people sitting at a table in your bar. Each card represents one person. One side of a card tells what a person is drinking, and the other side of the card tells that person's age.

In order to enforce the law, which of the card(s) below would you definitely need to turn over? Indicate **only** those card(s) you definitely need to turn over to see if any of these people are breaking the law**.**



Select the card(s) you want to turn over:

Drinking Beer   Drinking Soda   16 Years Old   25 Years Old

Suppose you have a friend who says he has a special deck of cards. His special deck of cards all have numbers (odd or even) on one side and colors (brown or blue) on the other side. Suppose that the 4 cards from his deck are shown below.

Your friend also claims that in his special deck of cards, even numbered cards are **never brown** on the other side. He says:

"In my deck of cards, all of the cards with an even number on one side are blue on the other."

Unfortunately, your friend doesn't always tell the truth, and your job is to figure out whether he is telling the truth or lying about his statement.

From the cards below, turn over **only** those card(s) that can be helpful in determining whether your friend is telling the truth or lying. Do not turn over those cards that cannot help you in determining whether he is telling the truth or lying.

Select the card(s) you want to turn over:

4 Card   9 Card   Blue Card   Brown Card

F.3.4 Anchoring

For this task, please first do the following. Take the last digit of your phone number.

- If it is 4 or less, please enter below the first two digits of your year of birth.
- If it is 5 or above, please enter below 100 minus the first two digits of your year of birth.

Enter the two digits:

In this task, you will be asked to make two estimates. Each time, you will be asked to

1. assess whether you think the quantity is greater than or less than the two digits that were just generated from your year of birth
2. give an estimate of the quantity (a number between 0 and 100).

Your answer will be counted as correct if it is no more than 2 away from the actual

---

Is the weight (in hundreds of tons) of the Eiffel Tower's metal structure more than or less than 19 hundred tons?

More than 19 hundred tons

Less than 19 hundred tons

What is the weight (in hundreds of tons) of the Eiffel Tower's metal structure? (options 0 to 100 in steps of 1)

F.4 Excerpt from Post-Experimental Questionnaire

You will now be asked some questions that test your understanding of the study so far.

Was it possible to earn a bonus in **part 1** of the study? (In this part, you answered questions on selecting cards.)

No

Yes, there was a possible bonus of

Was it possible to earn a bonus in **part 2** of the study? (In this part, you made estimates with respect to a number you calculated from your phone number and year of birth.)

No

Yes, there was a possible bonus of

In Part 2 of the survey, you answered the following question:

In 1911, pilot Calbraith Perry Rodgers completed the first airplane trip across the continental U.S., taking off from Long Island, New York and landing in Pasadena, California.

How many days did it take Rodgers to complete the trip? (options 0 to 100 in steps of 1)

Your answer to this question was: 33

How confident are you that this answer is within 2 of the correct solution?

	Not Confident at All	Not Confident	Somewhat Not Confident	Neutral	Somewhat Confident	Confident	Very Confident
How confident are you?	○	○	○	○	○	○	○

→