COGNITIVE BIASES: MISTAKES OR MISSING STAKES?

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

Abstract—Despite decades of research on heuristics and biases, evidence on the effect of large incentives on cognitive biases is scant. We test the effect of incentives on four widely documented biases: base-rate neglect, anchoring, failure of contingent thinking, and intuitive reasoning. In laboratory experiments with 1,236 college students in Nairobi, we implement three incentive levels: no incentives, standard lab payments, and very high incentives. We find that very high stakes increase response times by 40% but improve performance only very mildly or not at all. In none of the tasks do very high stakes come close to debiasing participants.

I. 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 to motivate them to put effort into the task. Yet nonexperimental 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, nonexperimental 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 microfoundations 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

Received for publication September 23, 2020. Revision accepted for publication May 6, 2021. Editor: Brian A. Jacob.

*Enke: Harvard University and NBER; Gneezy: UC San Diego Rady School of Management; Hall: Harvard Business School; Martin: Harvard University; Nelidov, Offerman, and van de Ven: University of Amsterdam and Tinbergen Institute.

We thank three very constructive referees, Thomas Graeber, and Florian Zimmermann for helpful comments. The study received IRB approval from Harvard's IRB and 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.

A supplemental appendix is available online at https://doi.org/10.1162/rest_a_01093.

[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 Chapman & Johnson (2002); Ariely et al. (2003); Epley & Gilovich (2006). Contingent reasoning has been studied in the psychology of judgment for many decades (e.g., Bazerman & Samuelson, 1983; Johnson-Laird, 1983; Cheng & 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 & Vespa, 2014, 2016; Enke & Zimmermann, 2019; Martínez-Marquina et al., 2019; Enke, 2020).

The Review of Economics and Statistics, July 2023, 105(4): 818–832

https://doi.org/10.1162/rest_a_01093

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

Because of the discussion in the literature about the frequency of cognitive biases in abstractly versus intuitively framed problems (Cheng & Holyoak, 1985; Gigerenzer & Hoffrage, 1995; Alekseev et al., 2017), we implement two cognitive tasks (base rate neglect and the Wason selection task) 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 BRN, intuitive BRN, 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 its 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. 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, we would also like to gather meaningful information on participants' behavior without any financial incentives. To achieve this objective while maintaining 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.

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 toward debiasing people. An alternative hypothesis is that cognitive biases reflect the high difficulty of rational reasoning, 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% CI). 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

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 (Frederick, 2005; Oechssler et al., 2009; Hoppe & Kusterer, 2011; Toplak et al., 2011).

interpretation: the within-treatment correlations between cognitive effort and performance are always very small, suggesting that it is not only effort 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 nonincentivized 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. Although some variation is seen 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 experimental literature that has investigated the role of 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 & 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; although the results are a bit mixed, the bottom line is that introducing small incentives generally did not affect the presence of biases. 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 & Laury, 2002). For example, high-stakes behavior in the ultimatum game reverts back predictions based on selfishness and rationality (Slonim & Roth, 1998; Cameron, 1999; Andersen et al., 2011). Likewise, literature in experimental game theory highlights that raising the stakes often significantly increases the fraction of equilibrium play (Smith & Walker, 1993; Cooper et al., 1999; Rapoport et al., 2003; Parravano & 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 with our paper is that we focus on established tasks aimed at measuring cognitive biases. In summary, either existing experimental work on stake size variations has 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., O'Neil et al., 1995; Baumert & Demmrich, 2001; O'Neil et al., 2005; Fryer, 2011; Bettinger, 2012; Levitt et al., 2016). 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 (Bettinger, 2012; Levitt et al., 2016). Although 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.

II. Experimental Design and Procedures

A. Tasks

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 & 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 & 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. Below is the wording of the abstract and intuitive versions of the mammography problem (see appendix B for the wording of the conceptually analogous car accident problems).

> **Abstract mammography problem:** *1% of women screened at age 40 have breast cancer. If a woman has breast cancer, the probability*

[2] Nonexperimental 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 & Graddy, 2009; Pope & Schweitzer, 2011; Graddy et al., 2022; Chen et al., 2016; Jetter & 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.

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%, 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 (see appendix B for an example of the figures we used) and works only 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. 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 accident 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).

Contingent reasoning: The wason selection task. Contingent reasoning has been studied in the psychology of judgment for many decades (e.g., Bazerman & Samuelson, 1983; Johnson-Laird, 1983; Cheng & Holyoak, 1985; Cosmides, 1989) and more recently in behavioral economics (e.g., Giglio & Shue, 2014; Esponda & Vespa, 2016; Barron et al., 2019; Enke, 2020). Although the experimental tasks in this literature differ across studies depending on the specific design objective, they all share 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."*

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 four cards showed "3," "8," "Green," and "Brown" on the visible side (or slight variations, see appendix G.3.3). 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.

The four cards showed "Drinking Beer," "Drinking Soda," "16 years old," and "25 years old" on the visible side (or slight variations). In this "social contract" version (adapted from Cosmides, 1989), the correct actions are turning over the "Drinking Beer" and "16" cards. While this problem is logically the same as the abstract version, this version may prompt "cheater detection" and may help participants to 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.

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 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?*
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?*

The intuitive answer to these questions is 10, while the 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.

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 & Kahneman, 1974). This can have consequential effects in situations such as negotiations, real estate appraisals, valuations of goods, or forecasts, although the evidence for anchoring effects in valuation tasks is mixed (see appendix A for references).

To test for anchoring, we follow others in making use of a random anchor. To generate an obviously uninformative 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 the following:

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?*

where [anchor] is replaced with the random number that is generated from a participant's phone number and year of birth. Appendix B lists the other sets of questions that we used.

Each question has a correct solution that lies between 0 and 100. Subjects are told that they can state only estimates between 0 and 100. Each participant who takes part in the anchoring condition completes two randomly selected questions from the set, 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.

B. 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, and 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 receive instructions for Part 2 only 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 consider only 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.

C. 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 G contains the 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 E2 in appendix E shows the distribution of responses. We see that two-thirds 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.[3] 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 deliver on their promises, effort arguably should not have increased.[4]

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

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

D. Participants

The experimental sessions take place at the Busara Center for Behavioral Economics in Nairobi, Kenya. We conduct our experiments in this lab because of 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. Tables D1 and D2 in appendix D report the resulting sample sizes by bias and incentive level, and summary statistics for participant characteristics across treatments. In total, 1,236 participants completed the study between April and July 2019. The majority (93%) are between 18 and 24 years old (mean age 22), and 44% 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 metastudy 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 the metastudy of Brañas-Garza et al. involving 118 studies and almost 45,000 participants (91% of which were from the United States or Europe), the fraction of correct responses for these same two questions is 36% and therefore very similar to what we see in our sample.[5,6]

E. Preregistration

We preregistered the design and target sample size on www.aspredicted.org (https://aspredicted.org/blind.php?x=5jm93d). Because we still had some money left at the end, we ended up with 1,236 instead of the pre-specified 1,140 participants. The results are very similar if we restrict the

[3] Tables D13 and D14 in appendix D 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).

[4] Although 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.

[5] 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.

[6] 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.

sample to the first 1,140 participants. See appendix C for more details about this and the speculated results.

F. Predictions by Experimental Economists

To complement our preregistration and to be able to compare our results with the profession's priors, we collect predictions for our experiments (DellaVigna & Pope, 2018; Gneezy & Rustichini, 2000). 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. We then ask them to predict response times and performance in the standard and high incentive conditions. Thus, each respondent issues 24 predictions (six tasks times two treatments times two outcome variables). We paid $100 to the respondent who issued the set of predictions that turned out to be closest to the actual data. Details about the sample (45 researchers and 23 Harvard students) are in appendix F.

III. Results

A. 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. We always report results of two-sided tests.

B. 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 cognitive effort in laboratory experiments (e.g., Luce, 1986; Ratcliff, 1978; Rubinstein, 2007; Krajbich et al., 2012; Spiliopoulos & 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 1 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 1, 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% higher in the high incentive condition compared to standard incentives. We observe the largest increase (52%) in intuitive base rate neglect, and the smallest increase (24%) in anchoring. Figure E4 in appendix E shows that very similar results hold when we look at median response times.

Table 1 quantifies the effects of incentive size on response times (in seconds) using OLS regressions.[7] 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 not based on randomization. In the regressions, an observation is the response time on a given question. Since each subject completed two questions per 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. Although 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 standard incentive treatment. On average, response times increase by 72 seconds; see the analysis on the pooled sample in column 7.

[7] In table 1, 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 D3 in appendix D, we instead use log response times. A p-p plot of residuals shows that they follow a normal distribution in this case (figure E5 in appendix E). The treatment comparisons deliver the same qualitative results as with raw response times. Table D4 in appendix D provides complementary nonparametric tests that also deliver very similar results.

FIGURE 1.—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

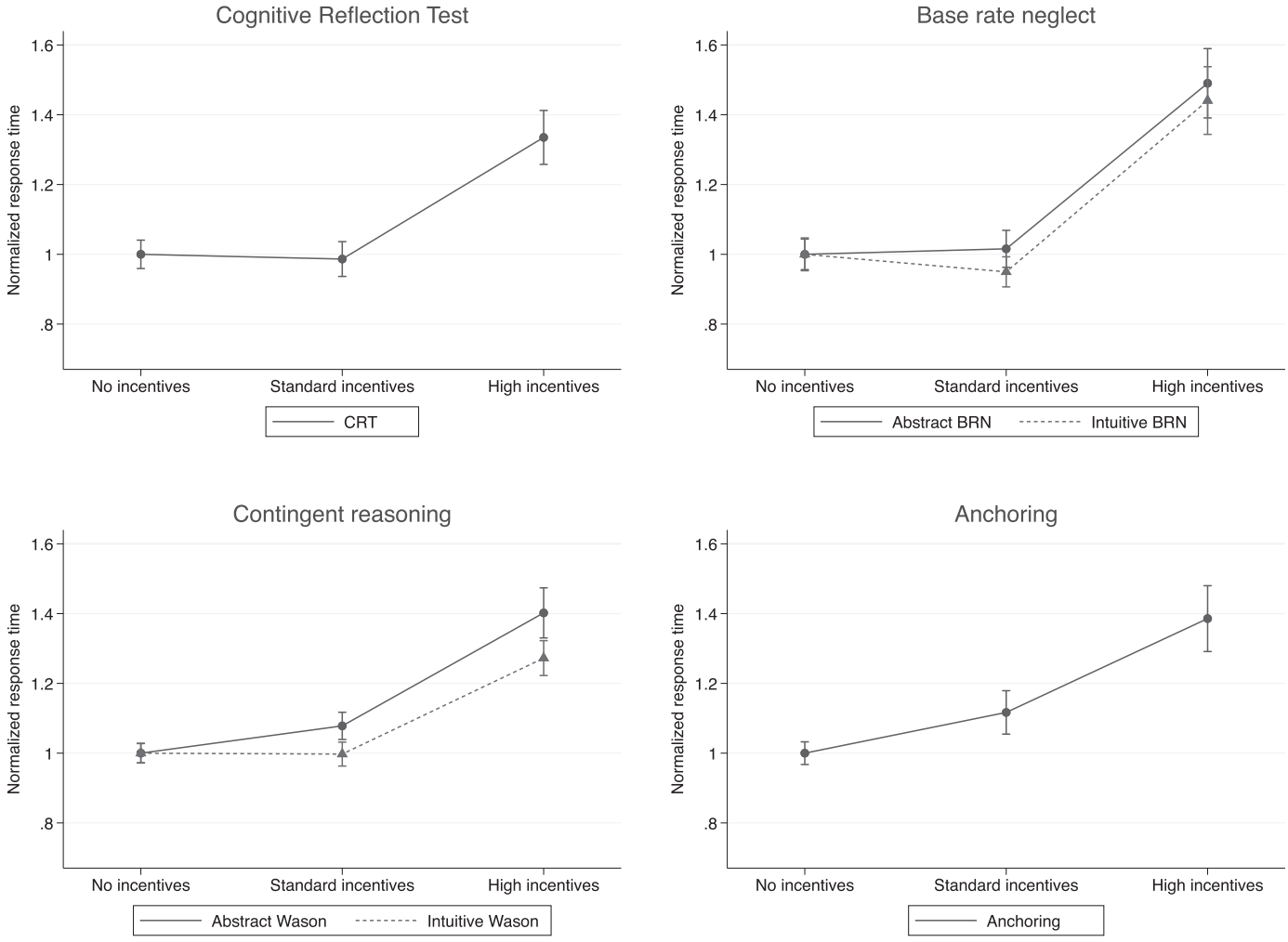
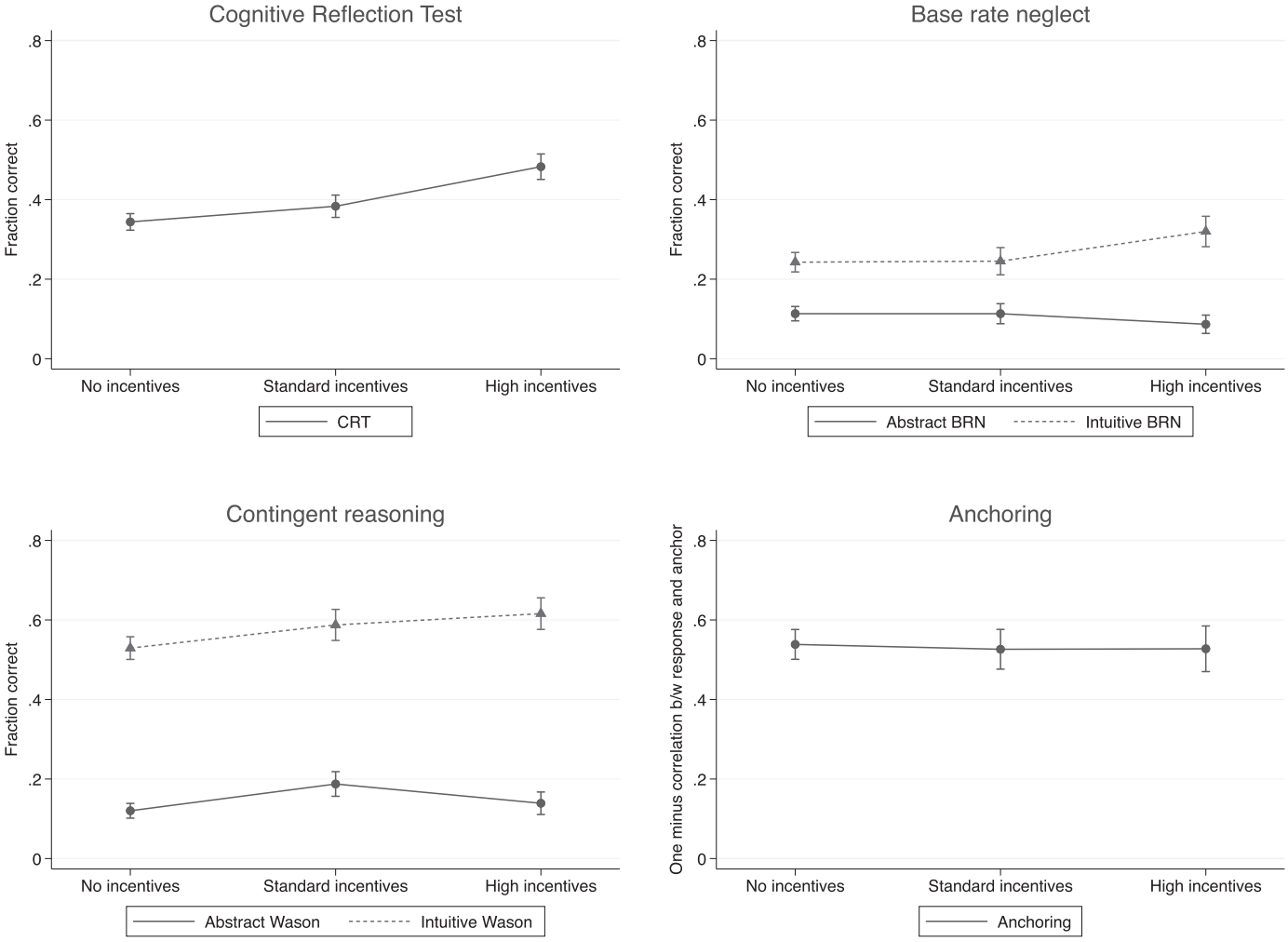


TABLE 1.—RESPONSE TIMES ACROSS INCENTIVE CONDITIONS

	Response time [seconds]						
		Base rate neglect		Contingent reasoning			
Dependent variable:	CRT (1)	Abstract (2)	Intuitive (3)	Abstract (4)	Intuitive (5)	Anchoring (6)	All tasks (7)
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)
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
Test *No = High*	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

OLS estimates, standard errors (clustered at subject level) in parentheses. Omitted category: standard incentive scheme. *No incentives*: dummy variable (1 if no incentives). *High incentives*: dummy variable (1 if high incentives). The last row reports the *p*-value of a test for equality of regression coefficients between *No incentives* and *High incentives*. * $p < 0.05$ and ** $p < 0.01$.

FIGURE 2.—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



As we show in figure E3 in appendix E, 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.[8]

[8] 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 & Ryan, 1985, for early accounts of the informational aspects of rewards, and Benabou & 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 D12 in appendix D.

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.*

C. *Incentives and Cognitive Biases*

Figure 2 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 preregistration, we count a response as "correct" if it is within 5 percentage points of the Bayesian posterior. Although subjects received a bonus only 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.

TABLE 2.—PERFORMANCE BY INCENTIVE LEVEL

	1 if answer correct						Answer
		Base rate neglect		Contingent reasoning		Tasks	
Dependent variable:	CRT (1)	Abstract (2)	Intuitive (3)	Abstract (4)	Intuitive (5)	1–5 (6)	Anchoring (7)
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)
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 × *No incentives*							0.0037
							(0.07)
Anchor × *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
Test *No* = *High*	<0.01	0.36	0.09	0.58	0.08	<0.01	0.86

OLS estimates, robust s.e. (clustered at subject level) in parentheses. Dependent variable: binary indicator for correct answer (columns 1–6) or the answer (column 7, between 0 and 100). Column (6) includes all tasks from columns 1–5. Omitted category: standard incentives. *No incentives*: dummy variable (1 if no incentives). *High incentives*: dummy variable (1 if high incentives). The last row reports the *p*-value of a test for the equality of regression coefficients between *No incentives* and *High incentives* (columns 1–6) or between Anchor × 1 if *No incentives* and Anchor × 1 if *High incentives* (column 7). * $p < 0.05$, ** $p < 0.01$.

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. In all other tasks, improvements are either very small or entirely absent. Across all tasks, high incentives never come close to debiasing participants. These results suggest that judgmental biases are not an artifact of weak incentives.

Table 2 quantifies these results through regression analysis (table D5 in appendix D provides complementary nonparametric tests that deliver very similar results). 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. Although 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 small.[9] 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, 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% CIs 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 incentive condition is smaller than 17 pp. Notably, for the more abstract tasks, we can rule out performance increases of only 3–4 pp, while in the more intuitive tasks the point estimates and confidence bands are a bit larger.[10] Indeed, for intuitive Wason and intuitive base rate neglect, the point estimates and CIs are such that we cannot statistically reject that they are different from the point estimate for the CRT.

[9] Including the anchoring data in this pooled regression is not meaningful because for anchoring the effect of interest is the treatment dummy interacted with the anchor.

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

The last row of table 2 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.

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 versus 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 versus 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 classify a response as correct only if it falls within five percentage points of the Bayesian posterior. In table D9 in appendix D, 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 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 correspond to the well-known heuristic/intuitive response patterns that the literature has documented. We define these intuitive answers: the 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. One common mistake is to only turn over "P." Another common mistake is to turn over "P" and "Q" instead of "P" and "not Q."

Table D10 in appendix D 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 BRN, 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. 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 preregistration, we noted that we would consider heterogeneity along different sociodemographic variables such as cognitive skills. Table D8 in appendix D 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 along these dimensions.

D. Comparison with Predictions

To put our results in perspective, we compare them with experimental economists' predictions. 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. The respondents are qualitatively correct in the sense that they predict that errors will not disappear even with very large incentives (see also figure E6 in appendix E). At the same time, they 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.[11] Mispredictions appear particularly pronounced for abstract BRN, 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 BRN (69% inside the CI) and lowest in abstract BRN (24%).[12]

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.*

E. Potential Mechanisms

Effort and performance. 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

[11] Figure E7 in appendix E 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 to standard incentives, and another 40–60% going from standard to high incentives.

[12] Appendix F 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.

about the more general relationship between effort and performance in cognitive biases tasks. Indeed, although 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 (Enke & Zimmermann, 2019; Graeber, 2019; Enke, 2020).

In our study, similar patterns hold. Longer response times are correlated with a higher probability of solving a problem correctly, yet the magnitudes of the OLS coefficients are fairly small (see table D11 in appendix D). 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. 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." 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. Unlike in many of the other tasks, the intuitive, wrong answers are relatively easy to disprove even without changing one's mental framework.

Although 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 & Schwartzstein, 2018, for an overview).

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). Although not incentivized, the data allow us to gauge how confident subjects are in their (usually wrong) responses (Enke & Graeber, 2019), 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 D12 in appendix D, reported confidence increases very little, if at all, as the stake size increases. This may suggest that although 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. 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.

IV. Discussion

This paper provides a systematic investigation of a longstanding 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 they suggest 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 because of 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, our results strongly suggest that people do not necessarily tend to make better decisions when the stakes are very high.

REFERENCES

Alekseev, Aleksandr, Gary Charness, and Uri Gneezy, "Experimental Methods: When and Why Contextual Instructions Are Important," *Journal of Economic Behavior and Organization* 134 (2017), 48–59.

Andersen, Steffen, Seda Ertaç, Uri Gneezy, Moshe Hoffman, and John A. List, "Stakes Matter in Ultimatum Games," *American Economic Review* 101:7 (2011), 3427–3439. 10.1257/aer.101.7.3427

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar, "Large Stakes and Big Mistakes," *Review of Economic Studies* 76:2 (2009), 451–469. 10.1111/j.1467-937X.2009.00534.x

Ariely, Dan, George Loewenstein, and Drazen Prelec, "'Coherent Arbitrariness': Stable Demand Curves without Stable Preferences," *Quarterly Journal of Economics* 118:1 (2003), 73–106. 10.1162/00335530360535153

Barron, Kai, Steffen Huck, and Philippe Jehiel, "Everyday Econometricians: Selection Neglect and Overoptimism When Learning from Others," 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* 16:3 (2001), 441–462. 10.1007/BF03173192

Bazerman, Max H., and William F. Samuelson, "I Won the Auction but Don't Want the Prize," *Journal of Conflict Resolution* 27:4 (1983), 618–634. 10.1177/0022002783027004003

Beggs, Alan, and Kathryn Graddy, "Anchoring Effects: Evidence from Art Auctions," *American Economic Review* 99:3 (2009), 1027–1039. 10.1257/aer.99.3.1027

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

Benabou, Roland, and Jean Tirole, "Intrinsic and Extrinsic Motivation," *Review of Economic Studies* 70:3 (2003), 489–520. 10.1111/1467-937X.00253

Benjamin, Daniel J., "Errors in Probabilistic Reasoning and Judgmental Biases," in *Handbook of Behavioral Economics: Applications and Foundations* (pp. 69–186) (Amsterdam: Elsevier, 2019).

Berk, Jonathan B., Eric Hughson, and Kirk Vandezande, "The Price Is Right, but Are the Bids? An Investigation of Rational Decision Theory," *American Economic Review* 86 (1996), 954–970.

Bettinger, Eric P., "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores," this REVIEW 94:3 (2012), 686–698.

Binswanger, Hans P., "Attitudes toward Risk: Experimental Measurement in Rural India," *American Journal of Agricultural Economics* 62:3 (1980), 395–407. 10.2307/1240194

Brañas-Garza, Pablo, Praveen Kujal, and Balint Lenkei, "Cognitive Reflection Test: Whom, How, When," *Journal of Behavioral and Experimental Economics* 82 (2019), 101455.

Camerer, Colin F., "Do Biases in Probability Judgment Matter in Markets? Experimental Evidence," *American Economic Review* 77:5 (1987), 981–997.

Camerer, Colin F., and Robin M. Hogarth, "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty* 19 (1999), 7–41. 10.1023/A:1007850605129

Cameron, Lisa A., "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia," *Economic Inquiry* 37:1 (1999), 47–59. 10.1111/j.1465-7295.1999.tb01415.x

Chapman, Gretchen B., and Eric J. Johnson, "Incorporating the Irrelevant: Anchors in Judgments of Belief and Value," in *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge: Cambridge University Press, 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," *Quarterly Journal of Economics* 131:3 (2016), 1181–1242. 10.1093/qje/qjw017

Cheng, Patricia W., and Keith J. Holyoak, "Pragmatic Reasoning Schemas," *Cognitive Psychology* 17:4 (1985), 391–416. 10.1016/0010-0285(85)90014-3

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* 89:4 (1999), 781–804. 10.1257/aer.89.4.781

Cosmides, Leda, "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task," *Cognition* 31:3 (1989), 187–276. 10.1016/0010-0277(89)90023-1

Danz, David N., Lise Vesterlund, and Alistair J. Wilson, "Belief Elicitation: Limiting Truth Telling with Information on Incentives," mimeo (2019).

Deci, E. L., *Intrinsic Motivation* (Berlin: Springer, 1975).

Deci, E. L., and R. M. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior (Perspectives in Social Psychology)* (Berlin: Springer, 1985).

DellaVigna, Stefano, and Devin Pope, "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy* 126:6 (2018), 2410–2456. 10.1086/699976

Enke, Benjamin, "What You See Is All There Is," *Quarterly Journal of Economics* 135:3 (2020), 1363–1398. 10.1093/qje/qjaa012

Enke, Benjamin, and Thomas Graeber, "Cognitive Uncertainty," National Bureau of Economic Research working paper w26518 (2019).

Enke, Benjamin, and Florian Zimmermann, "Correlation Neglect in Belief Formation," *Review of Economic Studies* 86:1 (2019), 313–332.

Epley, Nicholas, and Thomas Gilovich, "The Anchoring-and-Adjustment Heuristic: Why the Adjustments Are Insufficient," *Psychological Science* 17:4 (2006), 311–318. 10.1111/j.1467-9280.2006.01704.x, PubMed: 16623688

Esponda, Ignacio, and Emanuel Vespa, "Hypothetical Thinking and Information Extraction in the Laboratory," *American Economic Journal: Microeconomics* 6:4 (2014), 180–202. 10.1257/mic.6.4.180

——— "Hypothetical Thinking: Revisiting Classic Anomalies in the Laboratory," working paper (2016).

Frederick, Shane, "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives* 19:4 (2005), 25–42. 10.1257/089533005775196732

Fryer, Roland G. Jr., "Financial Incentives and Student Achievement: Evidence from Randomized Trials," *Quarterly Journal of Economics* 126:4 (2011), 1755–1798. 10.1093/qje/qjr045

Gigerenzer, Gerd, and Ulrich Hoffrage, "How to Improve Bayesian Reasoning without Instruction: Frequency Formats.," *Psychological Review* 102:4 (1995), 684–704. 10.1037/0033-295X.102.4.684

Giglio, Stefano, and Kelly Shue, "No News Is News: Do Markets Underreact to Nothing?" *Review of Financial Studies* 27:12 (2014), 3389–3440. 10.1093/rfs/hhu052

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel, "When and Why Incentives (Don't) Work to Modify Behavior," *Journal of Economic Perspectives* 25:4 (2011), 191–210. 10.1257/jep.25.4.191

Gneezy, Uri, and Aldo Rustichini, "Pay Enough or Don't Pay at All," *Quarterly Journal of Economics* 115:3 (2000), 791–810. 10.1162/003355300554917

Graddy, Kathryn, Lara Loewenstein, Jianping Mei, Michael Moses, and Rachel A. J. Pownall, "Anchoring or Loss Aversion? Empirical Evidence from Art Auctions," *Journal of Cultural Economics* (2022, forthcoming), 1–23.

Graeber, Thomas, "Inattentive Inference," working paper (2019).

Grether, David M., "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *Quarterly Journal of Economics* 95 (1980), 537–557. 10.2307/1885092

——— "Testing Bayes Rule and the Representativeness Heuristic: Some Experimental Evidence," *Journal of Economic Behavior & Organization* 17:1 (1992), 31–57.

Grether, David M., and Charles R. Plott, "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review* 69:4 (1979), 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* 32:1 (2018), 155–78. 10.1257/jep.32.1.155, PubMed: 29693346

Holt, Charles A., and Susan K. Laury, "Risk Aversion and Incentive Effects," *American Economic Review* 92:5 (2002), 1644–1655. 10.1257/000282802762024700

Hoppe, Eva I., and David J. Kusterer, "Behavioral Biases and Cognitive Reflection," *Economics Letters* 110:2 (2011), 97–100. 10.1016/j.econlet.2010.11.015

Jetter, Michael, and Jay K. Walker, "Anchoring in Financial Decision-Making: Evidence from Jeopardy!" *Journal of Economic Behavior & Organization* 141 (2017), 164–176.

Johnson-Laird, Philip Nicholas, *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* Vol. 6 (Cambridge, MA: Harvard University Press, 1983).

Kahneman, Daniel, and Amos Tversky, "On the Psychology of Prediction," *Psychological Review* 80:4 (1973), 237–251. 10.1037/h0034747

Krajbich, Ian, Dingchao Lu, Colin Camerer, and Antonio Rangel, "The Attentional Drift-Diffusion Model Extends to Simple Purchasing Decisions," *Frontiers in Psychology* 3 (2012), 193. 10.3389/fpsyg.2012.00193

Levitt, Steven D., "Testing Theories of Discrimination: Evidence from Weakest Link," *Journal of Law and Economics* 47:2 (2004), 431–452. 10.1086/425591

Levitt, Steven D., 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* 8:4 (2016), 183–219. 10.1257/pol.20130358

Lieder, Falk, Thomas L. Griffiths, Quentin J. M. Huys, and Noah D. Goodman, "The Anchoring Bias Reflects Rational Use of Cognitive Resources," *Psychonomic Bulletin & Review* 25:1 (2018), 322–349.

Luce, R. Duncan, *Response Times: Their Role in Inferring Elementary Mental Organization* Vol. 8 (Oxford: Oxford University Press on Demand, 1986).

Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa, "Failures in Contingent Reasoning: The Role of Uncertainty," *American Economic Review* 109 (2019), 3437–3474.
Metrick, Andrew, "A Natural Experiment in 'Jeopardy!'" *American Economic Review* 85 (1995), 240–253.
Oechssler, Jörg, Andreas Roider, and Patrick W. Schmitz, "Cognitive Abilities and Behavioral Biases," *Journal of Economic Behavior & Organization* 72:1 (2009), 147–152.
O'Neil, Harold F., Jamal Abedi, Judy Miyoshi, and Ann Mastergeorge, "Monetary Incentives for Low-Stakes Tests," *Educational Assessment* 10:3 (2005), 185–208.
O'Neil, Harold F. Jr., Brenda Sugrue, and Eva L. Baker, "Effects of Motivational Interventions on the National Assessment of Educational Progress Mathematics Performance," *Educational Assessment* 3:2 (1995), 135–157.
Parravano, Melanie, and Odile Poulsen, "Stake Size and the Power of Focal Points in Coordination Games: Experimental Evidence," *Games and Economic Behavior* 94 (2015), 191–199. 10.1016/j.geb.2015.05.001
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* 101:1 (2011), 129–57. 10.1257/aer.101.1.129
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* 43:2 (2003), 239–265. 10.1016/S0899-8256(03)00009-5
Ratcliff, Roger, "A Theory of Memory Retrieval.," *Psychological Review* 85:2 (1978), 59. 10.1037/0033-295X.85.2.59
Rubinstein, Ariel, "Instinctive and Cognitive Reasoning: A Study of Response Times," *Economic Journal* 117:523 (2007), 1243–1259. 10.1111/j.1468-0297.2007.02081.x
Sandefur, Justin, "Internationally Comparable Mathematics Scores for Fourteen African Countries," *Economics of Education Review* 62 (2018), 267–286. 10.1016/j.econedurev.2017.12.003
Slonim, Robert, and Alvin E. Roth, "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic," *Econometrica* 66 (1998), 569–596. 10.2307/2998575
Smith, Vernon L., and James M. Walker, "Monetary Rewards and Decision Cost in Experimental Economics," *Economic Inquiry* 31:2 (1993), 245–261. 10.1111/j.1465-7295.1993.tb00881.x
Spiliopoulos, Leonidas, and Andreas Ortmann, "The BCD of Response Time Analysis in Experimental Economics," *Experimental Economics* 21:2 (2018), 383–433. 10.1007/s10683-017-9528-1, PubMed: 29720889
Thaler, Richard H., "The Psychology and Economics Conference Handbook: Comments on Simon, on Einhorn and Hogarth, and on Tversky and Kahneman," *Journal of Business* 59:4 (1986), S279–S284. 10.1086/296366
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* 39:7 (2011), 1275.
Tversky, Amos, and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases.," *Science* 185 (1974), 1124–1131. 10.1126/science.185.4157.1124
van den Assem, Martijn J., Dennie Van Dolder, and Richard H. Thaler, "Split or Steal? Cooperative Behavior When the Stakes Are Large," *Management Science* 58:1 (2012), 2–20. 10.1287/mnsc.1110.1413