Confidence, Self-Selection and Bias in the Aggregate*

Benjamin Enke Thomas Graeber Ryan Oprea

January 3, 2022

Abstract

Economics experiments have produced widespread evidence that people suffer from a range of cognitive biases. However, unlike experiments, real-world institutions often allow decision makers to self-select out, potentially filtering (or amplifying) the impact of biases on economic aggregates. We study the economic impacts of such self-selection and how they depend on people's meta-cognitive awareness of their own errors. In a series of online experiments that cover a wide range of classical decision biases, we document large heterogeneity in how cognitive biases are related to the intensity of bets in speculative markets, bids for property rights in auctions and contributions to collective decisions. In some tasks, rational subjects are more confident than their biased counterparts and bet, bid and vote more aggressively. As a result, self-selection tends to filter the effect of irrationalities on aggregate quantities. However, in other tasks, confidence and performance are unrelated or even negatively correlated, so that experimental institutions do not filter errors and sometimes even magnify them. As a methodological blueprint, we show that a simple measure of the relative calibration of confidence strongly predicts the degree to which institutional self-selection filters the effect of irrationalities.

Keywords: Confidence, biases, social institutions, markets

*Enke: Harvard University, Department of Economics, and NBER, enke@fas.harvard.edu; Graeber: Harvard Business School, tgraeber@hbs.edu. Oprea: UC Santa Barbara, Department of Economics, roprea@gmail.com.

1 Introduction

Decades of experimental research on judgment and decision-making have revealed that individual decision makers are subject to a wide variety of cognitive and behavioral biases. Yet much of economics is concerned not with the quality of *individual* decisions but rather with the outcomes produced by *multiple* individuals interacting in institutions such as markets and organizations. The relevance of decision errors observed in the lab to economics therefore hinges to a great degree on whether these errors influence prices, distort allocative efficiency or have redistributive effects. While many researchers – referenced below – have studied a host of classical reasons for why individual errors may not influence markets and organizations (such as wealth dynamics, arbitrage and learning from experience), we consider the possibility that behavioral economics itself may produce limits on the influence individual errors have over economic outcomes. Our inquiry focuses on the role of meta-cognition: decision makers' awareness (or lack thereof) of their own biases.

Our point of departure is the observation that, in laboratory experiments, researchers "force" subjects to make cognitively difficult decisions, while real-world institutions often give decision makers the freedom to self-select into or out of decisions. For instance, people might shy away from market competition that could exploit their fallibility; they could be reluctant to aggressively bid in auctions over the property rights of objects that they do not fully understand; or they could refrain from contributing their opinion to decision-making processes in groups and organization when they suspect that they don't understand the matter at hand. Thus, the self-selection that is endemic to many social institutions could severely filter individual-level irrationalities, relative to the unfiltered measures we observe in the lab.

To what degree does the self-selection afforded by institutions alter the severity of how cognitive biases influence economic outcomes? We propose that this hinges in large part on the way meta-cognition is distributed across decision makers. Intuitively, whether errors will actually be filtered out, persist or even be amplified in basic social institutions crucially depends on the *relative confidence calibration*: the correlation between objective performance and confidence in the population. By "confidence," we mean the strength of belief in the ex ante optimality (rationality) of one's decision, rather than potentially imperfect information. Using a simple model, we illustrate that if the error-prone are relatively less confident in the optimality of their decisions, then social institutions will tend to effectively filter errors, removing their impact on economic outcomes, even if all decision makers are overconfident. If, on the other hand, performance and confidence are unrelated, or even negatively correlated (false confidence among the confused), then social institutions will not filter errors, or might amplify their effect

on aggregate quantities. Yet, to date, we know relatively little about how aware people are of their proneness to mistakes and especially to what degree this awareness varies across the many distinct types of biases documented in behavioral economics.

We implement a series of pre-registered online experiments to study the nature of self-selection under social institutions, and how this relates to the distribution of confidence in the population. The main features of our experiment are (i) a broad set of 15 cognitive tasks and associated biases; (ii) three different social institutions in which subjects interact; and (iii) direct measurements of subjects' confidence. In total, our experimental data comprise almost 70,000 decisions obtained from 2,153 participants from a diverse online sample as well as expert forecasts from researchers in the field.

We consider 15 widely studied cognitive tasks, culled from the literatures on errors in statistical reasoning and logic, financial decision-making, and behavioral game theory. Examples include the winner's curse, base rate neglect, correlation neglect, equilibrium reasoning, portfolio choice and thinking at the margin. Each task consists of two parts. In Part 1, subjects attempt to solve the cognitive task. In Part 2, we group subjects into 10-subject cohorts to participate in one of three canonical static social institutions. In these institutions, subjects make a single decision that determines their degree of self-selection based on their Part 1 performance.

We study a set of three canonical social institutions that differ from one another in many fundamental respects, but that all afford scope for self-selection. Our motivation for studying a variety of social institutions is that self-selection decisions always occur against a backdrop of institution-specific mutual behaviors and interactions that endogenously incentivize or de-indentivize confidence-based self-selection. While for some institutions the theoretical link between individual confidence and self-selection may be direct, in others it may be weakened or even eliminated by strategic considerations such as (higher-order) beliefs about others' confidence. To study the robustness of our findings to these details, our experiments study self-selection in three different institutional environments that vary these potentially important characteristics.

In our *Betting* treatment, ten subjects participate in a parimutuel betting market, in which they bet on the optimality of their Part 1 decisions. In our *Auction* treatment, we assign subjects instead to an auction for the right to be paid a bonus based on the quality of their Part 1 decision. In our *Committee* treatment, we have all ten subjects decide how intensively to vote for their own Part 1 decision to influence a common group decision. By betting, bidding or voting less intensively, subjects can partly or fully select out of influencing institutional outcomes in a continuous way. This continuous Part 2 decision mirrors the features of many real-life institutions, in which people have a great deal of flexibility regarding how intensively to act.

The difference between how intensively cognitively biased versus unbiased subjects

act determines the aggregate outcomes institutions produce: the degree of bias in market prices in the betting market; the rate of bias among the winners in the auction; and the aggregate vote share for the optimal decision in the committee. By comparing these outcomes to average rates of bias (measured in Part 1), we study to what degree institutions "filter" biases – whether and by how much self-selection makes aggregate outcomes *appear* more rational than the raw rate of bias in the population would suggest, for the full matrix of three institutions and 15 cognitive biases.

Our experiment studies the simplest versions of speculative markets, allocative markets and committee decision making in order to focus crisply on the effect of a specific mechanism (self-selection) on outcomes. Of course, in the field, all of these social institutions may filter cognitive biases also through additional ("classical") mechanisms. We deliberately abstract away from these, not because they are unimportant, but to cleanly study the role of self-selection in a simple experimental setup.

We find that, on average across all tasks, subjects who make optimal Part 1 decisions act more intensively in the Part 2 institutions. As a result, on average, biases are filtered by self-selection in all three institutions, producing institutional outcomes that appear less biased than subjects actually are. Importantly, however, we identify strong heterogeneity across biases in the degree to which this institutional filtering occurs. Some biases (e.g., iterated reasoning and exponential growth bias) are dramatically improved under all three institutions, while others (e.g., base-rate neglect and correlation neglect) are barely affected by self-selection. Some biases such as the winner's curse are even made more severe under the average institution, due to negative selection.

Although the three social institutions we study produce different types of aggregate quantities and involve different cross-cutting strategic complications, we find that the heterogeneity in institutional filtering across cognitive tasks is very similar across institutions. It is almost always true that those errors that get filtered more effectively in one institution also get filtered more in the other institutions. The uniformity in *which* cognitive biases are most susceptible to improvement by self-selection suggests that the across-task variation is likely rooted in characteristics of the biases themselves.

Our motivating hypothesis (pre-registered prior to the experiment) is that this variation can be partly explained by gathering data on subjects' confidence. As derived in a simple conceptual framework, our key prediction is that institutional filtering by self-selection critically depends on *the correlation between performance and confidence*, which we refer to as relative confidence calibration. For example, even if all subjects are overconfident, institutions should filter errors as long as unbiased subjects tend to be more confident than biased ones.

To test this hypothesis, we measure the subjective percentage likelihood subjects assign to the proposition that they made a payoff-maximizing Part 1 decision, separately

for each task. This allows us to ask a basic question of independent interest that has received little attention in behavioral economics so far: how well-calibrated are people's beliefs about their own biases and how heterogeneous is this across biases commonly studied in economics? For reasons discussed below in Section 2, we measure subjects' confidence both in a within-subjects design (some participants both indicate their confidence and make an institutional decision) and in a between-subjects design (some participants only indicate their confidence but never make institutional decisions).

We find strong heterogeneity in the quality of relative calibration across tasks. Although subjects are almost uniformly overconfident across all cognitive tasks, the correlations between confidence and optimality range from -0.13 (for misunderstanding mean reversion) to 0.39 (for gambler's fallacy). This documents that the relative confidence calibration is much better in some tasks than in others.

As predicted by our simple framework, this relative confidence calibration is strongly predictive of institutional filtering across cognitive tasks ($r \approx 0.76 - 0.91$). In other words, in tasks in which the relative confidence calibration is strong, self-selection in social institutions effectively filters errors. This result holds regardless of whether we elicit confidence in the sample of subjects who also make an institutional decision, or in a separate sample of subjects who never make any institutional decisions. These results suggest that in order to understand and predict to what degree cognitive errors will be "filtered out" of aggregate quantities through institutional self-selection, we must understand the distribution of meta-cognition.

Given the moderate number of cognitive tasks in our study (15), it is difficult to draw definitive conclusions about which task characteristics lead to better relative confidence calibration. Nonetheless, in an exploratory analysis, we use the "peakedness" of the distribution of non-optimal answers to classify tasks according to whether cognitive errors reflect strong misleading intuitions or a high degree of complexity. We identify tentative evidence that cognitive tasks that evoke strong intuitions (such as correlation neglect) are associated with worse relative confidence calibration than tasks that do not evoke a strong gut feeling (such as backwards induction).

The importance of directly measuring the performance-confidence correlation is reinforced by the observation that it is difficult for economists to forecast the magnitude of institutional filtering or the quality of relative confidence calibrations ex ante. To underscore this point, we ran a survey asking a panel of experts in behavioral and experimental economics to guess, for a variety of cognitive tasks, (i) to which degree performance and confidence go hand-in-hand; and (ii) to which degree one of our institutions (auctions) filters errors. We find that the experts consistently overestimate both the degree of institutional filtering and the relative calibration of confidence. Moreover, the experts severely underpredict how much institutional filtering and confidence calibration vary

across cognitive tasks.

In all, we view our paper as making three contributions. (i) Our results provide direct evidence on which types of cognitive errors get filtered out through self-selection in social institutions. (ii) We document that understanding or predicting institutional filtering of a given cognitive bias demands that we take into account the relative calibration of confidence in the population (rather than just the frequency of errors themselves). This is especially valuable from a methodological perspective because it suggests a simple blueprint: researchers who study cognitive biases can readily gauge the likely strength of institutional filtering for these biases without actually implementing laboratory institutions, by appending a simple confidence question to their experiment and calculating the confidence-performance correlation. (iii) We contribute evidence on the relative calibration of people's meta-cognition about their own errors across a large set of widely studied behavioral economics biases. Documenting these meta-cognitive regularities is relevant especially in light of our finding that they predict institutional self-selection, and we believe that the documentation we've begun can and should be extended to a wider set of biases.

Our paper ties into various literatures. At a broad level, our work relates to an ongoing discussion about when behavioral anomalies affect aggregate quantities (e.g., Russell and Thaler, 1985; List, 2003, 2004; Haigh and List, 2005; Fehr and Tyran, 2005; Lacetera et al., 2012; Sonnemann et al., 2013). Various experimental contributions have studied the effect of social institutions on a range of biases and economic behaviors. This includes both market experiments on errors in statistical reasoning (e.g., Camerer, 1987; Friedman, 2010; Ganguly et al., 2000; Kluger and Wyatt, 2004; Asparouhova et al., 2015; Enke and Zimmermann, 2019) and group decision studies (Charness and Sutter, 2012; Cooper and Kagel, 2005; Charness et al., 2007). Closely related is also the literature on excess entry (Camerer and Lovallo, 1999), which studies the link between individual confidence and market entry in specific cognitive tasks. Our main contribution to this line of work is to study the effectiveness of social institutions more systematically for different institutions and a broad set of cognitive tasks, and to show that measured confidence is an effective way to conceptualize and empirically predict how and why institutional effects differ strongly across cognitive biases.

Second, we relate to active economics and psychology literatures on confidence. Much of this work focuses on average over- or underconfidence (e.g., Koriat et al., 1980; Erev et al., 1994; Klayman et al., 1999; Moore and Healy, 2008). Previous work has shown that the worst performers in various social and intellectual tasks tend to be least aware of their deficient expertise (Dunning, 2011; Kruger and Dunning, 1999). In contrast to these studies, we focus on variation in the confidence-performance relationship across tasks of particular interest to behavioral economists, and how it helps us to under-

stand institutional self-selection. Closely related to this approach of ours is psychological work that shows a link between the relative calibration of subjects' confidence and the effectiveness of group discussion in aggregating individual knowledge in wisdom-of-crowds problems (Silver et al., 2021). This work is related to ours but does not study the specific formal institutions and cognitive tasks of interest to economists.

The paper proceeds as follows. Section 2 lays out our experimental design and Section 3 derives our predictions. Section 4 presents results on institutional improvements across tasks and Section 5 examines the role of confidence calibration. Section 6 reports on our expert survey. Section 7 concludes.

2 Experimental Design

2.1 Overview

Our goal is to design an experiment to answer three questions. First, to what degree do social institutions filter the effects of biases by motivating relatively biased decision makers to self-select out of participation in economic activity? Second, how strongly does this filtering vary across distinct biases and is this variation predicted by simple measures of subjects' awareness of their own proneness to mistakes (i.e., confidence)? Third, even abstracting away from institutional filtering, how is people's meta-cognition distributed across biases that are commonly studied by behavioral and experimental economists?

Our experiment consists of 15 periods, each consisting of two parts:

- **Part 1: Cognitive Task:** The subject makes a decision in one of 15 distinct cognitive tasks, randomly ordered across the 15 periods. The tasks all correspond to widely-studied cognitive biases in behavioral economics.

- **Part 2: Institutional Choice:** Each round, the subject is randomly assigned into an anonymous ten-person cohort to participate in a social institution: *Betting* markets, *Auctions* for decision rights, or *Committee* voting. She then makes an "institutional choice" linked to her Part 1 decision: a bet on the optimality of her Part 1 decision; a bid on the right to earn a bonus if her Part 1 decision was optimal; or a vote for her Part 1 decision to be adopted by her cohort. Her earnings for Part 2 depend on (i) the optimality of her Part 1 decision, (ii) her institutional choice and (iii) the institutional choices of others, in a manner that differs across institutions.

  In some treatments, subjects are not assigned to an institution in Part 2 but are instead simply asked to state their confidence (in percentage terms) that they made an optimal decision in Part 1.

The timeline is as follows: subjects (i) read computerized instructions; (ii) are required to pass a comprehension check; (iii) provide a response in the first cognitive task; (iv) indicate confidence or make a decision in a social institution related to the first task (depending on treatment); (v) repeat (iii) and (iv) for the second task etc.

2.2 Part 1: Cognitive Tasks

We selected 15 Part 1 cognitive tasks based on three principles. First, we wanted the tasks and associated biases to reflect a range of well-known and widely studied errors from behavioral and experimental economics. Second, we desired to sample tasks that relate to a variety of "Econ 101" principles of rationality and, hence, capture distinct forms of economically-relevant reasoning. Third, we wanted tasks that have very short and simple instructions, allowing us to observe every subject under all 15 tasks. In practice, this means that we selected tasks from the literature and then partly simplified the instructions or the problem setup.

We summarize the tasks in Table 1 and provide more details in Appendix A. Task instructions are provided in Appendix E. We divide the table into several sections to highlight that our tasks represent a broad swath of the different violations of "Econ 101" rationality postulates that economists and psychologists have documented. These include widely discussed errors in information-processing and statistical reasoning; well-known logic problems; errors in strategic reasoning (behavioral game theory); failure to identify constrained optima; and various errors related to financial decision-making.

Our empirical measure of performance in each task is a binary indicator that codes whether a response is (exactly) optimal, i.e., expected payoff-maximizing. This allows us to use the same performance metric across tasks. Clearly, the requirement that a response be exactly optimal is more demanding in tasks that have continuous response scales rather than two or three categorical response options. However, this is just one of several elements driving variation in the difficulties of our problems and, as we will see below, will be part of what is measured in our confidence elicitation. In addition, through pilots, we verified that none of our tasks generates a large mass of responses close-to-but-different-from the optimal response. Thus, the results are virtually identical if we instead code responses within a small window around the optimal response as optimal.

2.3 Part 2: Social Institutions

2.3.1 Overview

Our goal in Part 2 is to understand to what degree common social institutions motivate people to self-select out of participation in that institution. There are several distinct

Table 1: Overview of cognitive tasks and associated biases

Task	Bias / Description
Information Processing and Statistical Reasoning	
Base rate neglect (BRN)	Ignoring base rates when computing posteriors. ‡ Adaptation of taxi-cab problem from Tversky and Kahneman (1982).
Correlation neglect (CN)	Failing to account for non-independence of data in inference. ‡ Adaptation of tasks from Enke and Zimmermann (2019).
Balls-and-urns belief upd. (BU)	Failure to calculate Bayesian posterior. ‡ State probabilistic beliefs about which urn a colored ball is drawn from.
Gambler's fallacy (GF)	Failing to properly attribute independence to iid draws. ∘ Coin flipping task adapted from Dohmen et al. (2009).
Sample size neglect (SSN)	Failing to account for effect of sample size on precision of data. ∘ Adaptation of hospital problem from KT (1972); Bar-Hillel (1979).
Regression to mean (RM)	Failing to account for noise / failure to recognize regression to the mean. ∘ Adaptation of task from Kahneman and Tversky (1973).
Acquiring-a-company (AC)	Failing to properly condition on contingencies, à la the Winner's Curse. * Bidding task against computer as in Charness and Levin (2009).
Logic	
Wason task (WAS)	Failure to gather valuable evidence / positive hypothesis testing. ∘ Adaptation of 4-card task from Wason (1968).
Cognitive reflection test (CRT)	Following intuitive but misleading 'System 1' intuitions. ∘ Adaptation of Frederick (2005).
Strategic Reasoning	
Backw. ind. / iter. reason. (IR)	Limited depth of reasoning in recursive reasoning problems. * 1-player beauty contest game, à la Bosch-Rosa and Meissner (2020).
Equilibrium reason. (EQ)	Failure to forecast effects of incentives in dominance solvable games. ∘ Identify higher earning payoff matrix, adapted from Dal Bó et al. (2018).
Constrained Optimization	
Knapsack (KS)	Failure to identify optimal bundle in constrained optimization problem. * Knapsack problems taken from Murawski and Bossaerts (2016).
Financial Reasoning	
Thinking at the Margin (TM)	Thinking about average instead of marginal costs/benefits. * Adaptation of marginal tax task from Rees-Jones and Taubinsky (2020).
Portfolio choice (PC)	Failure to construct efficient portfolios due to 1/N heuristic. * Choose optimal portfolio vs. dominated 1/N portfolio.
Exponential growth bias (EGB)	Underestimate the exponential effects of compounding. † Interest rate forecasting problem adapted from Levy and Tasoff (2016).

Notes. Symbols indicate Part 1 payoff function in experimental currency units (ECUs). *: payoffs correspond to implied game payoffs as described in the task; ∘: 100 ECUs if optimal choice, nothing otherwise; †: 100-d and ‡: 100-3d where d = difference between response and expected payoff-maximizing / Bayesian response.

ways institutions can induce self-selection. For instance, as discussed below, institutions differ in whether – from a theoretical perspective – self-selection should primarily be governed by people's confidence or by their higher-order beliefs and/or risk aversion. As a result, we designed the experiment to study a range of institutions across which these considerations plausibly differ. Our goal is not to exhaustively cover every conceiv-

able type of institution but, rather, to provide an illuminating sampling of this variety. First, we selected two canonical types of market institutions that each rely on a different classical idea about how markets can filter out biases:

- **Betting markets:** A classical idea in economics is that well-informed bidders in speculative markets will be incentivized to bet more aggresively than less well-informed bidders, producing prices that efficiently aggregate information by putting greater weight on higher quality information. In principle, this same mechanism can apply also to traders with cognitive biases: to whatever degree traders have well-calibrated beliefs about their own decision quality, less biased traders will have incentives to bid more aggressively than more biased traders, producing prices that reflect the beliefs of the former more than the latter.

- **Allocative markets:** A second classical idea in economics is that people who more highly value products, resources and factor inputs will bid more for them in markets, causing markets to direct these resources to their most highly valued use. In standard models (absent externalities), competitive prices do just this by efficiently allocating goods to the subset of market participants who express the highest value for goods in their bids. For example, if a resource is cognitively difficult to make efficient use of (i.e., to put it to its most productive use), then if confidence is well-calibrated, relatively unbiased agents will tend to place higher value on the resource and thus outbid their competitors, acquiring the resource and thereby protecting it from inefficient use by biased competitors.

To these market mechanisms, we add a generic institutional mechanism commonly used to make decisions inside organizations:[1]

- **Committees:** Committees inside organizations aggregate opinions informally through discussion or formally through voting. Participants can often self-select out of this aggregation simply by not raising their voice in discussion, not adding their judgment to the proceedings or abstaining from voting. To the degree members' beliefs about their own biases are well-calibrated (and to the degree participants behave non-strategically), this self-selection will cause the commitee's aggregate decision to be less biased than its average member.

Notice that each of these three types of institutions are influenced by self-selection in distinct ways. In betting markets, agents are motivated to self-select out of the market (to bet less aggressively) by a desire to avoid private losses due to mistaken judgments.

[1]There is evidence that groups make more rational decisions than individuals. See, for example, Barahona et al. (2021) and Charness and Sutter (2012).

Potential institutional "filtering" occurs by improving the accuracy of the market price relative to the price that would have emerged if all agents had bet equally aggressively. In allocative markets, agents self-select out of the market by bidding less aggressively in order to avoid acquiring items that they believe they cannot effectively extract value from. Potential institutional filtering occurs by assigning resources to the least biased participants in the market rather than to bidders at random. Finally, in committees, agents are motivated to self-select out of the discussion by a fear that adding their judgments to the pool will worsen the group's aggregate decision and thereby decrease their own payoff. Potential institutional filtering occurs by producing aggregate decisions that reflect the beliefs of only the most competent participants rather than the belief of the average member of the committee.

In reality, all of these institutions potentially filter cognitive biases through many "classical" mechanisms other than self-selection, including learning from feedback, arbitrage, experimentation and wealth dynamics. We do not intend to argue that these are unimportant. However, for the sake of simplicity of the experimental design, we here abstract away from all of them and focus on the self-selection mechanism.

2.3.2 Implementation and Institutional Details

For the experiment, we aimed to find the simplest possible version of each of these institutions. In particular, we looked for versions of each institution that were static and that required only a single, simple decision from each participant.

Betting Markets: Parimutuel Betting. We implemented a *parimutuel betting market* – a particularly simple betting institution. In it, betters submit monetary bets on multiple securities, only one of which will turn out to be valuable. The total money bet is then redistributed to betters on the winning security in proportion to the amount each of those agents bet. The price of the winning asset can be summarized as the fraction of total money bet that was bet on the winning proposition. A canonical example for parimutuel betting markets is horse-race betting. However, there are also direct analogies to financial markets, where betters bet on one of multiple mutually exclusive states of the world, such as whether an asset will increase or decrease in value. Indeed, parimutuel betting markets are frequently implemented in laboratory experiments because of their simplicity and appealing resemblance of real-world markets (e.g., Plott et al., 2003).

In our implementation, participants were informed that a cohort of 9 other subjects in the study completed exactly the same Part 1 cognitive task as they did and that the ten participants would be grouped together into a betting market on their answers to these questions. In each of these Part 2 markets, each participant is endowed with 100 points (ECUs). The subject's task is to decide how many of those 100 points (if any)

to bet on the proposition that her own Part 1 response was optimal. This decision was implemented using a simple slider that ranged from 0 to 100, with no default value, see Appendix Figure 9 for an example screenshot.

The performance metric of interest in the betting market is the price of the security that is linked to the optimal Part 1 decision. Denoting the points bet by participant i as b_i and x_i as an indicator that equals 1 if the participant's Part 1 choice was optimal, the parimutuel price for this asset is given by:

$$\theta^{Betting} = \frac{\sum_{i=1}^{10} x_i b_i}{\sum_{i=1}^{10} b_i} \in [0, 1] \tag{1}$$

Notice that this price simply amounts to a re-weighting of individual Part 1 decisions, x_i, as a function of how many points each individual bets. For example, if all market participants bet the same amount (no self-selection occurs), then the market price will simply equal the raw optimality rate, $\overline{x}$ for the cohort. On the other hand, if only participants who make the optimal decision in Part 1 actually bet, the market price will equal one – the same price that would occur if all participants in the cohort were in fact unbiased. In our analyses, we can therefore easily gauge institutional filtering by comparing this price with the raw fraction of optimal Part 1 responses.

Individual payoffs are determined as follows. If a subject's Part 1 decision was not optimal, all points bet are lost and the subject only keeps the remaining endowment. If the subject's Part 2 decision *was* optimal, the subject's bonus is given by

$$\pi_i^{Betting} = \frac{b_i}{\theta^{Betting}} + (100 - b_i) \tag{2}$$

As a result, a subject is guaranteed to earn back at least what she bet (if their Part 1 decision was optimal), and the bonus is higher the more points are bet by subjects who did not take the optimal decision.

Allocative Markets: Discriminatory Auctions. For allocative markets, we implemented a sealed bid "discriminatory auction," a natural extension of a first-price auction to a setting with multiple winners. Specifically, in a group of 10, each subject receives an endowment of 100 ECU and decides how many to bid using a slider, see Appendix Figure 11. The five highest bidders win the auction and pay their own bid.[2] In exchange, the winners receive a bonus of 100 ECU if and only if their own Part 1 decision was optimal. Under standard assumptions, there is a symmetric and monotone equilibrium for

[2]If there are multiple fifth-highest bidders, the auction randomly selects from among the relevant set. The main reason we implemented a discriminatory auction with five winners rather than a single-unit auction with only one is that with five winners the performance of the institution can be more precisely estimated and doesn't rely as much on random noise in who happens to be the highest bidder.

discriminatory auctions that implements an efficient allocation to the M highest value bidders (Krishna, 2009, p.179). Intuitively, participants who believe that their Part 1 decision was incorrect have little incentive to bid.

The performance metric of interest in allocative markets is the optimality rate in the subset of participants who win the auction. Denoting the set of winners Ω:

$$\theta^{Auction} = \frac{\sum_{i \in \Omega} x_i}{5} \tag{3}$$

If no self-selection occurs (if everyone bids the same amount), resources will be assigned randomly and the expected performance will be $\overline{x}$, the raw optimality rate in the cohort. On the other hand, if five optimal participants submit the five highest bids, the performance metric will be one – the same value that would occur if *all* participants in the cohort were unbiased. In our analyses, we will then again compare this outcome of the auction with the raw Part 1 optimality rates.

Committees: Utilitarian Voting. Once again, subjects were assigned to groups of 10. Each participant was endowed with 100 votes, any number of which a subject could submit for their own Part 1 decision (the remainder are unused). These votes can be interpreted either as literal votes or instead as the intensity with which a participant argues in favor of her Part 1 solution (e.g., the number of minutes she chooses to spend arguing in a group discussion). This decision was again represented using a simple slider, see Appendix Figure 10.

The institutional performance metric of interest is the fraction of votes placed on the optimal decision. Denoting by v_i subject i's number of votes:

$$\theta^{Committee} = \frac{\sum_{i=1}^{10} x_i v_i}{\sum_{i=1}^{10} v_i} \in [0, 1] \tag{4}$$

All subjects in a group made the same profit, $\pi_i^{Committee} = 100 \times \theta^{Commitee}$. As a result, it doesn't matter for a subject's payoff whether she submitted votes herself, or that her own Part 1 decision was optimal. This captures a group decision process in which each member of a team has a common interest in the quality of the group's decision.

Note that although the incentives in committees are very different from those in parimutuel betting, the performance metric, θ, is calculated in an identical way as a function of subjects' institutional decisions. Just as in betting, if there is no self-selection (if all participants submit the same number of votes for their choices), this will just be equal to the raw optimality rate in the committee. However, if only optimal decision makers vote, the performance metric will be equal to one.

2.4 Measuring Confidence

We desire to understand (i) how aware people are of their own proneness to mistakes across a broad range of tasks and (ii) to what degree this awareness predicts institutional filtering. Throughout the paper, when we speak of "confidence," we mean the strength of belief in the ex ante optimality (rationality) of one's decision, rather than potentially imperfect information. Because confidence is at times used in ambiguous ways in the literature, we clarify that confidence is fundamentally different from the variance of one's beliefs. For instance, it is perfectly possible for a person to be fully confident that her beliefs are Bayesian, even when those beliefs have strictly positive variance.

In principle, there are two different designs in which subjects' confidence can be elicited. First, one could elicit confidence from the same set of subjects that also take institutional decisions ("within-subjects design"). Second, one could elicit confidence in a "between-subjects design," in which those subjects who report their confidence never take any institutional decisions, and vice versa. Regardless of which design is used, the objective is to assess whether we see strong (weak) institutional improvement in those tasks in which the confidence-performance correlation is strong (weak).

The two potential designs both have strengths and weaknesses. A within-subjects design has the advantage that it allows the researcher to directly observe the individual-level link between confidence and institutional behavior. This is important because a main assumption underlying this paper is that institutional decisions indeed at least partly reflect confidence. At the same time, a within-subjects design has the disadvantage that it potentially introduces consistency concerns: subjects may take institutional decisions that are in line with their previously-stated confidence not because this is what they truly desire but because they desire to appear consistent vis-a-vis the experimenter.

On the other hand, a between-subjects design introduces non-trivial measurement error. Because the researcher tries to predict the institutional improvement observed in one sample of subjects with the confidence calibration observed in another sample of subjects, the predictability is going to be attenuated in any finite sample because the researcher links the behavior of two different groups of people. Moreover, a between-subjects design does not allow the researcher to observe the individual-level link between confidence and institutional action. Given these considerations, we implement both types of experiments, see Table 2 for an overview.

Between-subjects design. This treatment, *Confidence*, follows the same outline as the institutions treatments discussed above in that it consists of two parts. After each Part 1 task, the subject is asked the exact same confidence question for all 15 tasks throughout the study, which closely follows prior work (e.g., Enke and Graeber, 2021a,b). The

Table 2: Overview of experimental treatments

Treatment	Elicitations	No. of subjects
Betting	Cognitive task; parimutuel betting	387
Auction	Cognitive task; discriminatory auction	323
Committee	Cognitive task; committee voting	337
Confidence	Cognitive task; confidence	334
Betting Within	Cognitive task; confidence; parimutuel betting	105
Auction Within	Cognitive task; confidence; discriminatory auction	105
Committee Within	Cognitive task; confidence; committee voting	104

Notes. The table lists the main treatments that are used for empirical analyses throughout the paper. Further smaller robustness treatments are reported throughout the paper as they become relevant.

instructions introduce the idea of an "optimal decision" to subjects, which we define as "the decision that maximizes your earnings, on average."[3] The confidence question then asks: "How certain are you that your decision in Part 1 was optimal?". The instructions further clarify for subjects that they are supposed to indicate the percent chance that they think their decision was optimal. Subjects used a slider to enter a value between 0% and 100%, with no initialization for the slider, see Appendix Figure 12.

It is worth pointing out that confidence should mechanically be higher in tasks that have a discrete (e.g., binary) response scale than in those that have a continuous one. This is fine for our purposes because optimality rates will tend to vary for the same reason. Moreover, our main interest is not variation in the *level* of confidence across tasks but instead in the *confidence-optimality correlation*.

Within-subjects design. Treatments *Betting Within*, *Auction Within* and *Committee Within* consisted of three parts each. In Part 1, subjects again solved a cognitive task. In Part 2, they indicated their confidence as described above (unincentivized). In Part 3, they took an incentivized institutional decision.

2.5 Incentives

Given the large variety of tasks that we deploy, the payment procedures necessarily need to differ across cognitive tasks. As summarized in Table 1, we can partition the cogni-

[3]In our main experiments, the confidence elicitation screen for each task additionally specifies the definition of "optimal." For example, in the Knapsack problem, the elicitation screen specifies that "Your decision is optimal if it maximizes your earnings.", while in the balls-and-urns belief updating task we specify that "Your decision is optimal if it corresponds to the statistically-correct option given the information you are provided." We implemented a robustness treatment in which we measure confidence without this additional explanation, with effectively identical results.

tive tasks into three sets: (i) those that have a natural implied game payoff, such as the profit from one's bid in the acquiring-a-company game; (ii) tasks that have an objectively correct (rational) solution and that feature discrete response options, such as Wason's selection task; and (iii) tasks that have a rational solution and (nearly) continuous response scales, such as a balls-and-urns belief elicitation experiment. As a result, we also deploy three types of scoring rules. Based on the insight of Danz et al. (2020) that simple scoring rules are most effective in inducing truth-telling, our overarching goal was to keep the incentive structure relatively simple and transparent. Appendix E provides the details for each task.

In tasks of type (i), payoffs follow immediately from the description of a game. In tasks of type (ii), subjects received 100 ECU if their response was correct and nothing otherwise. In tasks of type (iii), we deployed a simple linear scoring rule with maximum payoffs of 100 ECU, such as $\pi = \max\{100 - 3d; 0\}$, where d is the distance between the subject's guess and the rational response. In total, subjects in the *Confidence* treatment made 15 incentivized decisions, while subjects in the other conditions made 30 incentivized choices. For each subject, one randomly selected decision was paid out.

Treatments *Betting*, *Auction*, *Committee* and *Confidence* were implemented at the same time and subjects were randomized into these four treatments. To investigate whether our results are sensitive to financial incentives, we implemented our experiments with two slightly different stake sizes. 596 subjects took part in the experiment with an exchange rate of \$5 per 100 ECU earned, while for the remaining 785 subjects it was \$10 per 100 ECU. Given that we do not find significant differences in rates of optimality in Part 1 or in correlations between Part 1 and Part 2 decisions across these two sets of subjects, we pool the data in what follows. We did not pre-register predictions about the potential effects of the stake size variation. Treatments *Betting Within*, *Auction Within* and *Committee Within* were likewise randomized within experimental sessions with a stake size of \$5 per 100 ECU earned.

2.6 Logistics

All experiments were conducted on Prolific, an online worker platform that has been shown to deliver higher-quality data than Amazon Mechanical Turk (Gupta et al., 2021). We pre-registered that our experiments would be conducted using Prolific's "representative sample" option. However, this considerably slowed down data collection, so that we quickly switched to Prolific's general respondent pool. Average earnings in our experiments were \$11.82 for a study that took 33 minutes, on average. Depending on the treatment, this includes a \$4–6 participation fee. These average earnings are considerably higher than an hourly wage of \$9.60 that is recommended by Prolific. Data were

collected in June and November 2021.

We took two steps to ensure high data quality. First, the initial screen in the study consisted of an attention check. Second, subjects in all treatments completed a comprehension check that consisted of four questions. Any prospective participant who failed the attention check or answered one or more comprehension checks incorrectly was immediately routed out of the study and does not count towards the number of pre-registered completes. See Appendix E for the comprehension check questions in all treatments.

We pre-registered two aspects of our experiments at `https://aspredicted.org/hg4zi.pdf`. First, we pre-registered that we would sample 1,400 subjects across our four between-subjects treatments, with random assignment within each experimental session. Because slightly fewer subjects passed our comprehension checks than we anticipated, our final sample for the between-subjects treatments consists of 1,381 subjects. Second, we pre-registered that we would conduct two types of analyses: (i) the performance improvement that is caused by an institution and (ii) to which extent the correlation between performance and confidence predicts for which tasks we observe larger institutional improvements.

3 Framework and Hypotheses

This section lays out a simple empirical framework for our experimental design. The purpose of this framework is to derive and describe hypotheses for our experiment and guidance for our analysis, rather than to serve as a general micro-founded model of how confidence determines behavior across institutional environments.

Self-selection and institutional filtering. Suppose that each of N agents forms a judgment about the solution to a cognitive task. Agent i's solution is optimal (correct), $X_i = 1$, with probability p_i and incorrect, $X_i = 0$, with probability $(1 - p_i)$. Aggregate pre-institutional performance in the cognitive task is given by the raw rate of optimality in the N-agent cohort: $\Theta^{pre} = \frac{1}{N}\sum_i X_i \in [0,1]$. Θ^{pre} is a random variable with mean $\theta^{pre} \equiv \mathbb{E}[\Theta^{pre}] = \frac{1}{N}\sum_i p_i$. The agents next participate in a social institution, making an institutional decision, $k_i \in [0,1]$. These decisions represent bids in auctions, bets in betting markets and number of votes in committees. This institutional decision, k_i, is a measure of the agent's degree of self-selection into the institution: a higher k_i means that the institutionally determined outcome will be more strongly affected by the optimality of agent i's own task response.

Let $\Theta^{post} \in [0,1]$ be a performance metric produced by the institution (e.g., the vote share for the optimal option, the price of the ex post optimal security etc.), and let $\theta^{post} \equiv \mathbb{E}[\Theta^{post}]$ denote the mean of that metric. We can compare this to the same metric

calculated under the assumption that no self-selection occurs (i.e., $k_i = k_j$, $\forall i, j$). In our setting, this is just equal to Θ^{pre}, the raw rate of optimality in the cohort. We define $\mathbb{G} = \theta^{post} - \theta^{pre}$ as a measure of "expected institutional filtering" due to self-selection. It will be positive if institutions produce performance metrics *as if* the population of participants are more rational than they actually are.

This institutional filtering is directly dependent on the way the self-selection decision, k_i, is distributed in the population. The way this dependence works varies slightly across the institutions we consider. In what follows, we refer to measures of *expected* performance in a population. As discussed in Section 2.3.2, for *Betting* and *Committees* the metric of interest is

$$\theta^{post}_{bet,com} = \frac{\sum_i k_i p_i}{\sum_i k_i}. \tag{5}$$

In *Betting*, θ^{post} corresponds to the expected price produced by the parimutuel betting institution for an asset linked to the optimal decision; in *Committee*, θ^{post} is the expected vote share for the optimal decision. Institutional filtering is given by

$$\mathbb{G}_{bet,com} = \theta^{post}_{bet,com} - \theta^{pre} = \frac{\sum_i p_i \left(N k_i - \sum_j k_j\right)}{N^2 \bar{k}}. \tag{6}$$

This expression directly depends on self-selection: it is positive if and only if the better-performing agents bet more or submit more votes, i.e., if those with higher p_i bet more or submit more votes than the average subject in the cohort.

For auctions, institutional performance is the optimality rate of the subset Ω of decision makers who won the auction. The expected institutional gain follows as:

$$\mathbb{G}_{auc} = \theta^{post}_{auc} - \theta^{pre} = \frac{1}{|W|} \sum_{j \in \Omega} p_j - \frac{1}{N} \sum_i p_i. \tag{7}$$

Thus, the auction leads to an improved aggregate outcome if the winners of the auction on average exhibit better expected performance on the task.

Self-selection and confidence. The above shows that institutional filtering ($\mathbb{G}$) is *proximally* shaped by self-selection, k_i. Our hypothesis is that this institutional filtering is *ultimately* shaped by the confidence in one's decisions, c_i. Under this assumption, two relationships are crucial:

1. The relationship, β, between confidence, c_i, and expected task performance, p_i.
2. The relationship, ω, between confidence, c_i, and institutional decisions, k_i.

Our experiment allows us to empirically measure both of these relationships, and to relate them to the efficacy of institutions at reducing bias. Suppose for simplicity that

confidence is linearly related to decision quality as follows:[4]

$$c_i = \alpha + \beta \cdot p_i \tag{8}$$

Rather than viewing eq. (8) as a behavioral micro-foundation of confidence statements, we interpret it as a linear approximation of the aggregate relationship between subjective confidence and expected performance, akin to standard calibration curves. Here, β captures the slope and hence the sensitivity of subjectively perceived performance to objective performance. Throughout the paper, we refer to the relationship between confidence and performance as *relative confidence calibration*. Average overconfidence, $d \equiv \bar{c} - \bar{p} = \alpha + (\beta - 1)\bar{p}$, is a function of both α and β.

The expression in eq. (8) highlights that confidence could be miscalibrated in two distinct ways. First, even if performance and confidence change one-for-one ($\beta = 1$), there may be average over- or underconfidence, $d \neq 0$. Second, even if there is no average over- or underconfidence ($d = 0$), variation in confidence across individuals might imperfectly reflect actual variation in underlying performance, $\beta \neq 1$. Here, a negative relationship, $\beta < 0$, implies that better-performing agents are, on average, *less* confident. Our main observation will be that β is essential for predicting whether or not a social institution filters biases.

Next, suppose that institutional self-selection has an approximately linear relationship with confidence:

$$k_i = \omega \cdot c_i \in [0, 1]. \tag{9}$$

Here, ω captures the degree to which self-selection, k_i, actually depends on confidence as opposed to other considerations. For instance, as discussed below, institutional decisions may not be *only* governed by confidence but also by, e.g., higher-order beliefs about others' confidence.

These relationships allow us to derive a pre-registered prediction about the relationship between institutional filtering ($\mathbb{G}$) and measured confidence (c_i). These predictions depend crucially on the assumption that $\omega > 0$, i.e., that more confident agents make more intensive institutional choices, which we test and strongly confirm below. In the following predictions, we also make the weak assumption that $\alpha > 0$, which says that, for an objective probability of being correct of zero, people's average subjective probability that they are correct is strictly larger than zero.

Prediction 1. *(i) If the within-task relationship between performance and confidence is positive ($\beta > 0$), institutional performance improvements are positive ($\mathbb{G} > 0$). (ii) In-*

[4]Both this formulation for c_i and the one for the institutional action k_i below are linear approximations that will fail close to the boundaries of zero and one. We choose this modeling strategy purely for the sake of simplicity. Going forward, we assume that α, β and ω are all such that $c_i \in (0, 1)$ and $k_i \in (0, 1)$.

stitutional performance improvements, $\mathbb{G}$, increase in the within-task correlation between performance and confidence, β.

The prediction holds strictly for *Betting* and *Committee*. In these institutions, the precise number of submitted bets or votes matters for the institutional outcome. The prediction holds weakly for *Auction* because only the ordering of bids matters in that case. All proofs are relegated to Appendix B.

Much of the prior literature focuses on the level of miscalibration of confidence, especially pervasive overconfidence. In contrast, our prediction highlights that institutional improvements depend on the *relative* calibration of confidence across agents. The following prediction clarifies the role of average overconfidence. In contrast to the first prediction above, this one was not pre-registered, but we test it in ancillary analyses for the sake of completeness.

Prediction 2. *The effect of mean overconfidence, d, on institutional performance improvements, $\mathbb{G}$, is ambiguous. Specifically, (i) there is no relationship in Auctions; and (ii) in Betting and Committees, the effect can be positive or negative. If $\beta > 0$, the effect of an increase in d is weakly negative.*

As we will see below, while our cognitive tasks differ widely in β (with some positive and some negative), the average correlation is positive. We therefore expect a weak negative relationship between average overconfidence and institutional filtering.

Variation across institutions. It is not obvious that Prediction 1 will hold consistently across different institutions. The framework suggests it will fail if the linkage between confidence and institutional choices is weak (if $\omega \approx 0$). There are many reasons why this might be true, some easily anticipated theoretically. For instance, in committee voting, higher-order beliefs about the cognitive performance and confidence of others are important components of the strategic environment and should theoretically compete with an agent's own confidence in shaping her institutional choices. Similarly, in parimutuel betting, the mapping between bets and confidence may vary due to heterogeneous tolerance for risk. Likewise, there may be heterogeneity in the strategies that people use to map confidence to institutional choice, perhaps due to bounded rationality. All of these mechanisms could weaken ω at the population level.

This is an important reason for why we opted to vary the type of institution itself in our design. At one extreme, in auctions, agents should, in equilibrium, submit bids that are monotonic in expected value, so we might expect ω to be quite high. At the other extreme, in committee voting, higher order beliefs about the rationality of others may theoretically weaken ω significantly. Yet, while there are theoretical reasons to hypothesize that confidence may be more predictive of institutional improvements for some

social institutions than for others, this need not be the case empirically. For instance, there is much evidence from experimental game theory that suggests that people often do not engage in the type of higher-order thinking that could attenuate ω. Indeed, it is not obvious why, in practice, people would exhibit the cognitive sophistication to solve for the equilibrium of a voting mechanism if they don't have the cognitive sophistication to solve a cognitive task in the first place. It is therefore ultimately an empirical question whether and how institutions differ in the filtering they produce, and how this depends on the relative confidence calibration.

4 Institutional Improvement Across Cognitive Tasks

Unsurprisingly, the cognitive task performance and institutional improvements observed in the between-subjects and the within-subjects treatments are very similar to each other. For the sake of brevity, we here present the results from the between-subjects treatments and always refer the reader to corresponding analyses for the within-subjects treatments in the Appendix.

4.1 Performance Across Tasks and Subjects

The average of optimal Part 1 responses across all tasks in treatments *Betting*, *Auction*, *Committee* and *Confidence* is 28%. Figure 1 shows sizable variation in the performance across the 15 cognitive tasks. While the optimality rates for 9 out of 15 tasks is clustered between 14% and 30%, the total range spans from <10% to >80%.[5] Appendix Figure 15 provides a complementary subject-level perspective by showing a CDF of the number of optimal Part 1 decisions per subject.

4.2 Which Errors do Social Institutions Filter?

Recall from Section 3 that institutions will tend to filter out errors if participants who get a Part 1 task wrong bet less, bid less, or submit fewer votes in Part 2. Figure 2 shows cumulative distribution functions (CDFs) for Part 2 choices (k_i), separately for subjects who did ("optimal") and who did not ("suboptimal") solve the corresponding Part 1 task optimally. Pooling across the 15 cognitive tasks, the CDF of optimal responses always first-order stochastically dominates that of suboptimal responses in all three institutions. The average difference in institutional decisions is slightly more pronounced in *Betting* (64.8 average bet for optimal Part 1 decisions and 47.4 for suboptimal, a difference of

[5]Appendix Figure 23 provides an analogous analysis for the within-subjects treatments.

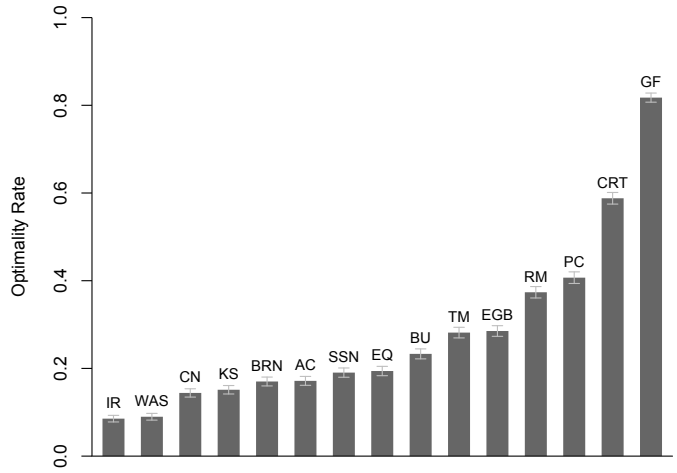


Figure 1: Fraction of optimal Part 1 responses across tasks in treatments *Betting*, *Auction*, *Committee* and *Confidence*. $N = 1,381$ participants completed each of the 15 tasks in individually randomized order. The tasks and optimal responses are described in Appendices A and E. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Thinking at the margin; WAS=Wason task. Error bars are the standard error of the binomial mean.

37%) than in *Committee* (75 average votes for optimal vs. 57.9 for suboptimal, 29%), or *Auction* (56.4 average bid for optimal vs. 43.6 for suboptimal, difference of 29%).[6]

These patterns immediately imply that, on average across tasks, all of our institutions filter errors to some extent. Our primary interest, however, is *in which tasks* institutions lead to a performance improvement, and by how much. To this effect, Figure 3 shows institutional improvements in performance, separately for each cognitive task. We calculate the percentage point improvement in, for example, market prices in the betting market, relative to the counterfactual in which no selection occurs (which, recall, is simply equal to the raw Part 1 optimality rate in each of our institutions). To take a simple example, suppose that in a given task the Part 1 optimality rate is 50%. Further suppose that, in the committee institution, those five subjects that got the task right each submit 100 votes, that one subject that got the task wrong also submits 100 votes and that all other subjects submit no votes. In this example, the institutional improvement is given by $(500/600 - 0.5) \cdot 100 = 33$ percentage points.

An immediate takeaway from Figure 3 is that there is large variation in improvement rates across tasks for all institutions. For example, in EGB (exponential growth bias) and IR (iterated reasoning / backward induction), aggregate error rates decrease substantially in all institutions, but they do not get filtered or even amplify in tasks such as EQ (equilibrium reasoning), AC (acquiring-a-company), RM (regression to the mean), BRN

[6]Appendix Figure 24 provides an analogous analysis for the within-subjects treatments.

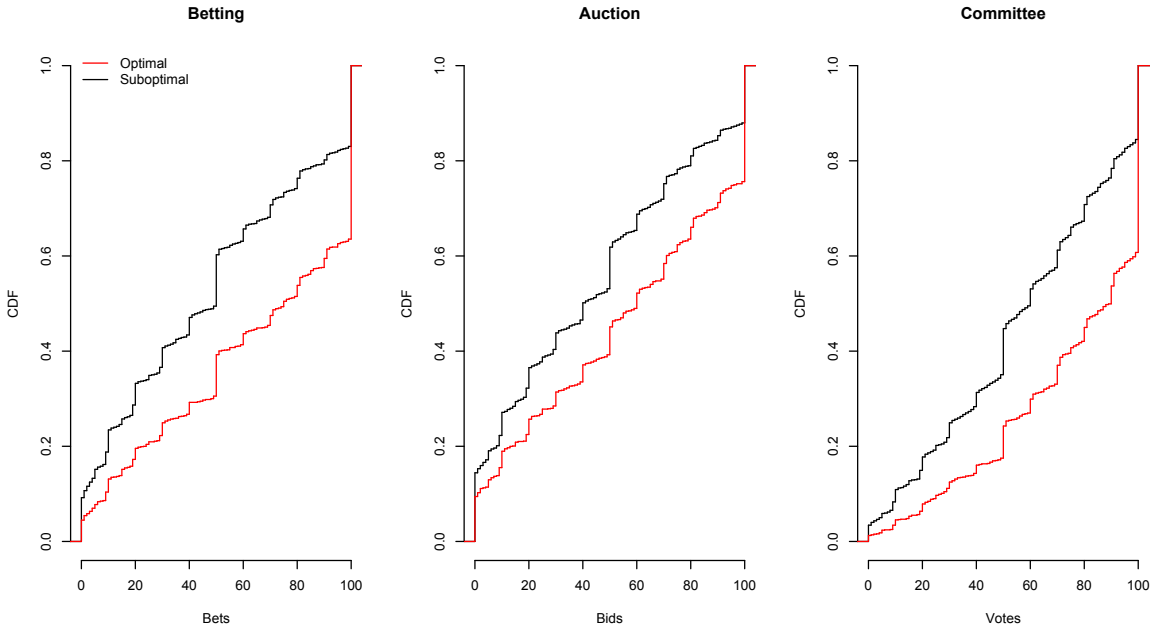


Figure 2: Part 2 institutional behavior by Part 1 optimality. Based on $N = 4,845$ Part 2 decisions in the *Auction* condition, $N = 5,805$ in *Betting* and $N = 5,055$ in *Committee*, pooled across 15 different cognitive tasks. For each institution, the sample of Part 2 decisions is split by whether the corresponding Part 1 decision was optimal and empirical distribution functions are displayed.

(base rate neglect) or CN (correlation neglect).[7]

These patterns suggest that the relationship between Part 1 responses and Part 2 behavior – *who* self-selects in institutional decisions – varies substantially across tasks. In some tasks, it is mostly people who make suboptimal decisions that select out. In other tasks, optimal and suboptimal decision makers make roughly the same Part 2 decisions. Indeed, in our data, the within-task correlation between bids/bets/votes and optimality ranges from $r = -0.12$ in Acquiring-a-Company (AC) bets to $r = 0.49$ in Exponential Growth Bias (EGB) bets, see Appendix Figure 18.

Although there is some variation in which tasks are most and least improved across institutions, there is for the most a strong agreement. If a given cognitive bias does or does not get filtered to a great degree by one institution, then it also does or does not get filtered to a great degree in the other institutions. The pairwise correlations in improvements between institutions range from 0.85 to 0.91. This striking commonality across different institutions suggests that the differential patterns of institutional filtering across cognitive biases is not driven by random noise or institutional peculiarities. Rather, the uniformity of results suggests that the across-task variation in institutional filtering is rooted in characteristics of the biases themselves. We next investigate our pre-registered hypothesis that this heterogeneity is determined by differences in meta-

[7]Appendix Figure 25 provides an analogous analysis for the within-subjects treatments.

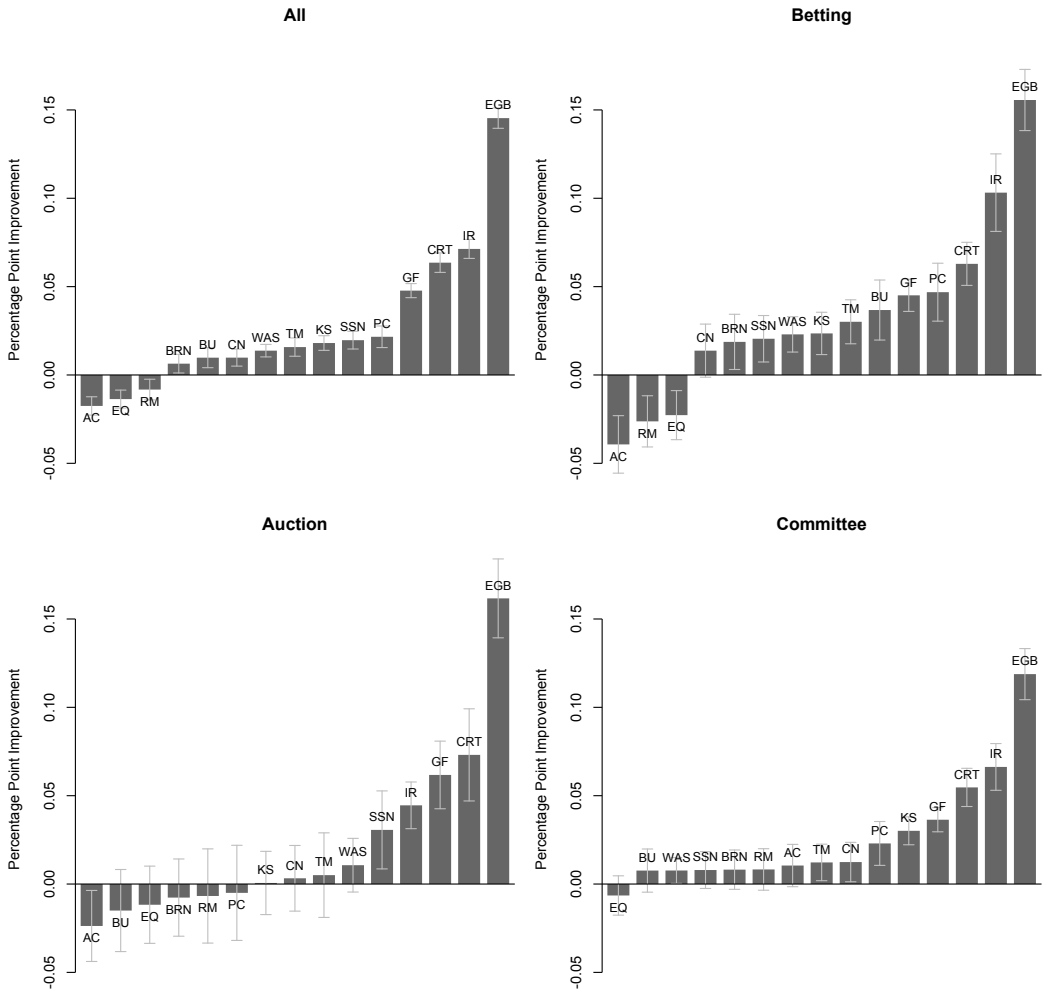


Figure 3: Performance improvement through institutions across tasks. Percentage point improvement is computed as the aggregate performance rate after institutional filtering minus the fraction of optimal Part 1 responses. The aggregate performance rate is based on 10,000 randomly constructed ten-subject cohorts for each institution, taking the mean over all samples. Based on $N = 323$ participants in the *Auction* condition, $N = 387$ in *Betting* and $N = 337$ in *Committee*. One-standard error bars are conservatively calculated as the ratio of the standard deviations of improvements over these random cohorts divided by the square root of the number of cohorts available in the dataset (e.g., 387/10=38.7 in Betting). Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

cognitive awareness of errors across tasks.

Robustness: Efficiency measure. Our main measure of institutional improvement is a measure of *absolute* improvement. An alternative perspective is to consider an efficiency measure: *what fraction of the theoretically possible improvement* is realized, given the actual distribution of performance. This measure is of interest because in our institutional groups of ten subjects each, it will sometimes happen that even if a social planner se-

lected the most competent players, the post-institution performance would not be 100% because not enough subjects actually get the task right. In *Betting* and *Committee*, at least one out of ten subjects needs to get a task right in order for the theoretically possible post-institutional performance to be 100%, yet this is not always the case. In *Auctions*, we awarded the right to a bonus to five out of ten participants, so that the institutional performance metric can only equal one if at least five participants get a task right. To account for this, we compute an additional efficiency metric of improvement, which is given by the fraction of the theoretically possible improvement (given the distribution of performance among subjects) that is achieved by an institution. Appendix Figure 16 shows a similar degree of task heterogeneity for this efficiency measure.

5 The Role of Relative Confidence Calibration

5.1 Confidence Across Subjects and Tasks

Pooling across all 15 cognitive tasks in treatment *Confidence*, we find that optimal decisions are associated with higher confidence. Average confidence in the pool of optimal decisions is 76%, while it is 64% in the pool of suboptimal decisions, see Appendix Figure 17. As in previous work, we find that individual-level heterogeneity in confidence is correlated with demographics, see Appendix Table 3: (i) people are overconfident on average; (ii) men are more overconfident than women; and (iii) subjects with lower performance are more overconfident than those with high performance (the "Dunning-Kruger effect", Kruger and Dunning (1999)). These familiar correlations suggest that we are effectively measuring confidence using our unincentivized question.

Our main interest, however, is in the variation of the performance-confidence correlation *across tasks*. Figure 4 shows the within-task correlation between Part 1 optimality and Part 2 confidence across our 15 tasks. We see large variation in the cross-sectional calibration of confidence. In no task is the Pearson correlation coefficient north of $r = 0.5$, and in six tasks the correlation is actually *negative*, meaning that, if anything, sub-optimal respondents tend to be *more* certain that they solved the task correctly. This is true in particular for RM (misattribution of regression to the mean) and TM (thinking at the margin rather than the average in a tax minimization problem), for which we can statistically reject the hypothesis of no correlation between confidence and optimality. Appendix Figure 26 presents an analogous analysis for the within-subjects treatments.

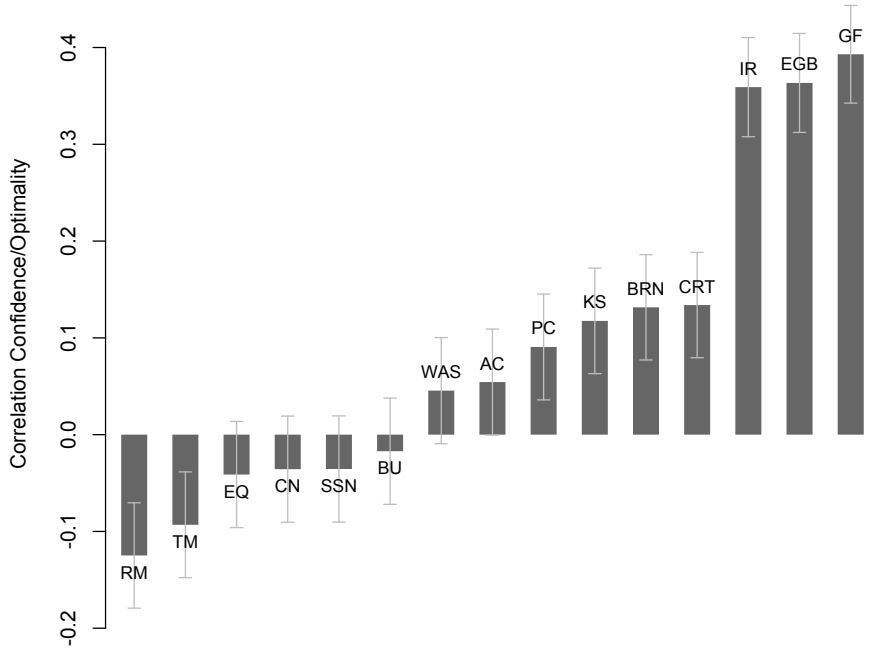


Figure 4: Within-task correlation between confidence and Part 1 optimality across tasks in treatment *Confidence*. Displayed are the Pearson correlation coefficients, based on $N = 334$ participants. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

5.2 Relative Confidence Calibration and Institutional Improvement

Our hypothesis is that the sign and magnitude of the optimality-confidence correlation illustrated in Figure 4 is predictive of institutional improvement. As discussed in Section 2.4, this question can be analyzed in both a within-subjects and a between-subjects design. Figure 5 shows the results for both approaches.

In the left panel (between-subjects data), the vertical axis shows the magnitude of institutional improvement in percentage points, averaged across treatments *Betting*, *Auction* and *Committee*. The horizontal axis shows the within-task correlation between optimality and confidence in treatment *Confidence*. Thus, in this figure, we predict the institutional improvement observed in one sample of subjects with the relative confidence calibration observed in another sample of subjects.

In the right panel (within-subjects data), the vertical axis and horizontal axes show the same quantities as discussed above, except that they are all derived from treatments *Betting Within*, *Auction Within* and *Committee Within*. Thus, we here predict the institutional improvement observed in one sample of subjects with the relative confidence

calibration of those same subjects.

We make two main observations. First, the figures visually confirm our hypothesis. In tasks with a strong relative calibration of subjects' confidence (in a rank-order sense), the institutional improvement is large. This is the case for tasks such as exponential growth bias, iterated reasoning and gambler's fallacy. Opposite patterns hold for attribution (understanding regression to the mean), thinking at the margin, correlation neglect and equilibrium reasoning. Second, these patterns are slightly more pronounced in the within-subjects data. In the between-subjects data the Pearson correlation between institutional improvement and the confidence-optimality correlation is $r = 0.76$, while it is $r = 0.93$ in the within-subjects data. As discussed in Section 2.4, this is unsurprising because in any finite sample the between-subjects approach necessarily introduces measurement error because institutional improvement and confidence calibration are observed in different samples of people. Indeed, Figure 5 visually suggests that the between-subjects results are a noisier version of the within-subjects results. In any case, as we hypothesized, the relationship between institutional improvement and relative confidence calibration is always strong.

An immediate question is whether the predictability of institutional improvement through confidence calibration is similar across the different institutions that we study. For both the between- and the within-subjects data, we find that this is indeed the case. The correlations between institutional improvement and relative confidence calibration are $r^{\text{auction}} = 0.69$, $r^{\text{betting}} = 0.73$, $r^{\text{committee}} = 0.77$, $r^{\text{auction within}} = 0.9$, $r^{\text{betting within}} = 0.9$ and $r^{\text{committee within}} = 0.91$, see Appendix Figures 20 and 28. This is arguably a non-obvious result: even though optimal behavior in some of the institutions relies on considerations other than confidence (e.g., committee voting in principle may be influenced by higher-order beliefs), we find that the predictive power of confidence is roughly the same. These results suggest that when researchers would like to understand (or predict) the degree to which social institutions may filter biases through self-selection, gathering data on the correlation between confidence and performance will yield valuable insights.

Mechanism: Confidence and institutional self-selection. Our hypothesis for why the confidence-optimality correlation is so strongly predictive of the magnitude of institutional improvement is that less confident subjects are more like to self-select out in the institution: that they bet less, bid lower amounts, and submit fewer votes. Through the lens of our conceptual framework in Section 3, this amounts to saying that $\omega > 0$. In our within-subjects treatments, we can directly test this assumption. Appendix Figure 27 shows binned scatterplots of institutional actions against stated confidence separately for each institution. We find that the correlations between stated confidence and bids,

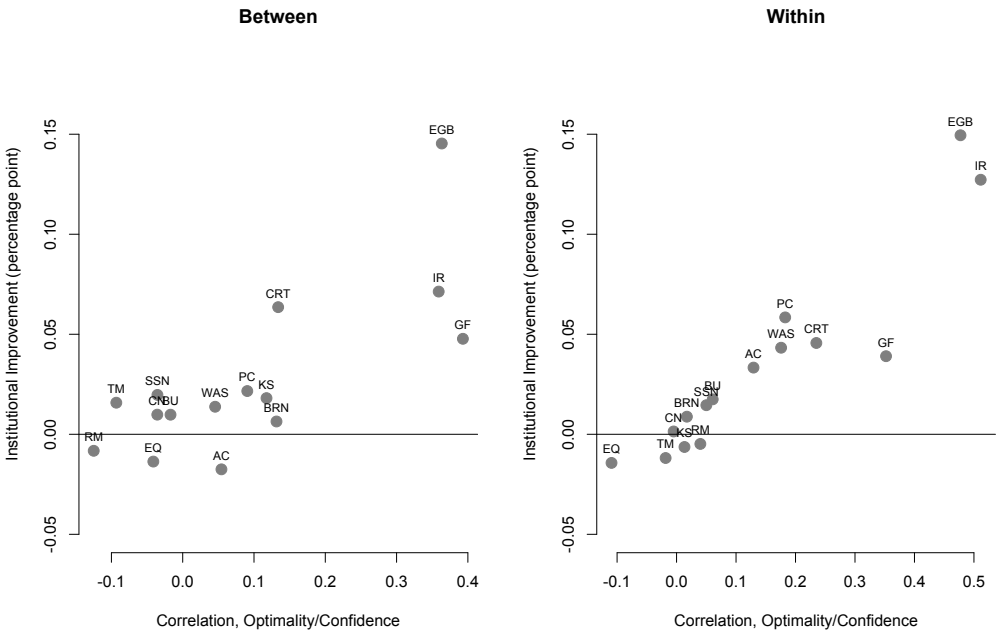


Figure 5: Confidence-optimality correlation and institutional improvement. The left panel shows the results for the between-subjects treatments and the right panel those for the within-subjects treatments. In the left panel, the horizontal axis shows the within-task correlation between confidence and optimality in treatment *Confidence*. The vertical axis shows the average institutional improvement across treatments *Betting*, *Auction* and *Committee*. In the right panel, we show analogous quantities, except that they are all derived from treatments *Betting Within*, *Auction Within* and *Committee Within*. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

bets and votes are $r = 0.79$, $r = 0.85$ and $r = 0.89$, respectively.[8]

Robustness I: Efficiency measure. Our main analysis considers absolute institutional improvement. Appendix Figure 21 shows corresponding results using the efficiency measure introduced in Section 4 that normalizes actual improvement by the theoretically possible improvement. The calibration of confidence turns out to be an even stronger predictor of institutional performance under this alternative specification, with r rising to 0.87 in the between-subjects data and to $r = 0.94$ in the within-subjects data.

Robustness II: Sensitivity to outliers. To examine whether our finding about the relationship between the confidence calibration and institutional improvement is driven by specific tasks, we perform a leave-two-out analysis: we compute 10,000 correlation coefficients, in each run excluding two randomly selected tasks. The resulting distribution

[8]Appendix Figure 19 shows that confidence and institutional action are strongly correlated not just across individuals but also across cognitive tasks: in those tasks in which subjects are on average more confident, they also bet / bid / vote more intensively, on average.

of correlations confirms that the result is not driven by individual tasks. In the between-subjects treatments, the Pearson correlation coefficients vary between 0.61 and 0.83 when pooling all institutions, with a mean of 0.76. In the within-subjects treatments, they vary between 0.86 and 0.98, with a mean of 0.93.

5.3 Which Types of Errors Have Strong Confidence Calibration?

Our results immediately raise the question of *what characteristics* of tasks make decision makers more or less likely to be well-calibrated? Given that we are looking at a moderately-sized sample of tasks, an analysis of this question is naturally tentative in nature and ought to be interpreted with care because, with relatively few data points, one faces the risk that any "theory" will overfit the data.

A natural starting point is the role of misleading intuitions. Many "classical" task paradigms in the decision-making literature are associated with a compelling, yet flawed intuition, such as in the CRT (e.g., Kahneman, 2011). Other tasks, such as backwards induction, constrained optimization in the Knapsack problem or the acquiring-a-company task arguably do not elicit similarly strong intuitions. Instead, we can arguably loosely think of these errors as "complexity-driven." There are reasons to hypothesize that the confidence calibration will be less accurate for intuition-based biases. Indeed, a long literature in psychology on processing fluency and the "feeling of rightness" (e.g Thompson, 2009; Thompson et al., 2011) posits that flawed intuitions are particularly misleading if they are associated with the experience of high confidence.

In the absence of an established definition of the strength of misleading intuitions in a problem, we construct a proxy by looking at the mass of responses on the modal suboptimal answer. According to this classification, a task is more likely to generate a false strong intuition the larger the number of people who choose the exact same wrong answer (conditional on being wrong). For instance, in the CRT or correlation neglect, large fractions of people produce exactly the same wrong answer, while in exponential growth calculations that is not the case. We construct this measure only for those nine tasks for which there are more than ten possible responses. This is because if there are only two or three response options, it is impossible to disentangle whether people jump to a specific wrong solution because of a misleading intuition or because of, e.g., random judgment noise.

Appendix Figure 22 provides some tentative evidence that tasks in which wrong responses are strongly peaked ("intuition problems") indeed see smaller institutional improvements. For example, in correlation neglect and balls-and-urns belief updating, about 50% of all responses are concentrated on a single answer, and the optimality-confidence correlation in these tasks is roughly zero. In iterated reasoning, exponential

growth bias and the Knapsack problem, on the other hand, the fraction of concentrated responses is between 10% and 30%, and the within-task optimality-confidence correlation in these tasks is always strictly positive.

We acknowledge that this analysis is tentative in nature, for at least two reasons. First, it is based on only nine tasks. Second, it ignores that the response scales across these nine tasks differ widely. Future research is needed to shed more light on the determinants of the quality of the relative confidence calibration.

5.4 The Role of Average Overconfidence

Our conceptual framework in Section 3 accounted for two forms of potential miscalibrations in the distribution of confidence: (i) for a given average level of confidence, the correlation between confidence and optimality (β) could be less than one; and (ii) for a given β, average confidence could be too high or low, $d \neq 0$. In this section, we empirically explore the potential implications of average over- or underconfidence for institutional filtering.

Figure 6 plots average confidence in treatment *Confidence* against the optimization rate in each task.[9] Under perfectly calibrated average confidence, all dots should be located close to the diagonal line. We observe two main patterns: first, there is average overconfidence in all of the 15 tasks. Second, confidence and the optimization rate are strongly positively correlated at the task level, $r = 0.75$. As a consequence, absolute overconfidence is much more pronounced in some tasks than others. This insensitivity of confidence statements to the optimization rate mirrors previous research on metacognition (e.g., Erev et al., 1994; Moore and Healy, 2008).

How does such average overconfidence relate to institutional behavior and resulting performance improvements? As discussed in Section 3, average overconfidence and the resulting more aggressive average behavior could translate into (weakly) lower institutional improvement when confidence and performance are positively correlated, $\beta > 0$. In line with this prediction, our data indeed show relatively weak negative relationships between average overconfidence (computed as average confidence minus average performance) and institutional improvement. Figure 7 illustrates the results by again averaging the institutional improvement across all institutions. The correlations between overconfidence and institutional improvement are given by $r = -0.34$ for the between-subjects experiments and by $r = -0.32$ for the within-subjects experiments (neither significantly different from zero at conventional levels).[10] These results are in line with

[9] Appendix Figure 29 shows the results for the within-subjects treatments.

[10] Regarding the specific institutions, the correlations are $r = -0.28$ for *Betting*, -0.36 for *Committee*, -0.36 for *Auction*, $r = -0.24$ for *Betting Within*, $r = -0.40$ for *Auction Within* and $r = -0.23$ for *Committee Within*.

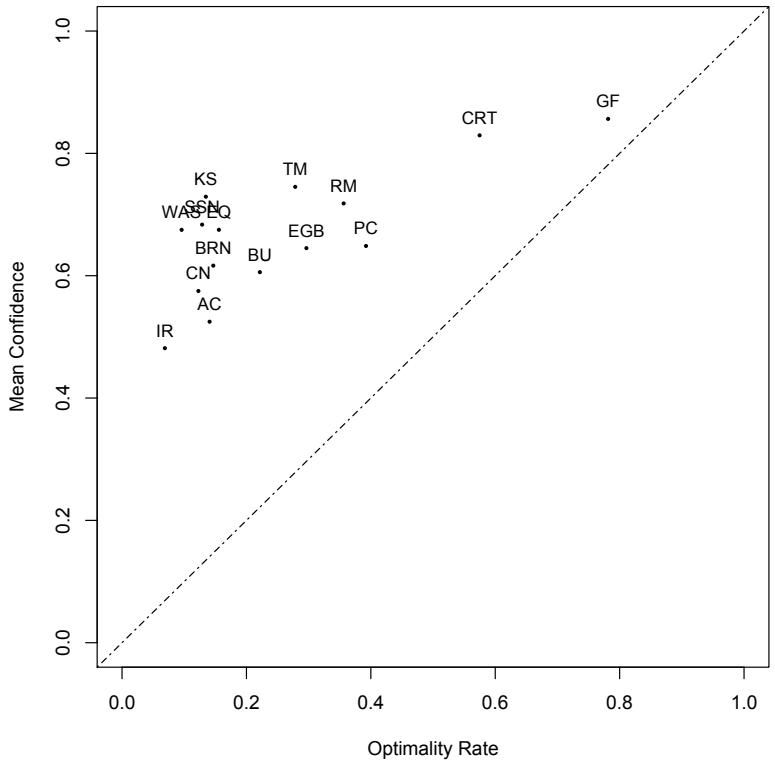


Figure 6: Fraction of optimal Part 1 responses and average confidence in treatment *Confidence*, separately for each task. Based on $N = 334$ respondents. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Minimize taxation; WAS=Wason task.

Prediction 2 from our framework in Section 3, and highlight that what matters for institutional filtering is indeed mostly the confidence-optimality correlation, rather than average overconfidence.

6 Expert Predictions

We compare our experimental results with the predictions of a sample of experts. The expert survey was conducted using the Social Science Prediction Platform.[11] We distributed the survey among participants of the *CESifo Area Conference on Behavioral Economics 2021* and attendees of the online speaker series *VIBES – The Virtual Behavioral Economics Seminar*. We obtained a total of $N = 38$ complete responses. Among those who indicated their professional level, 57% are faculty at all levels, 10% are post-doctoral re-

[11]Public study ID *sspp-2021-0028-v1*, see `https://socialscienceprediction.org/s/b04a0x`. We thank Stefano DellaVigna and Nicholas Otis for excellent comments and support.

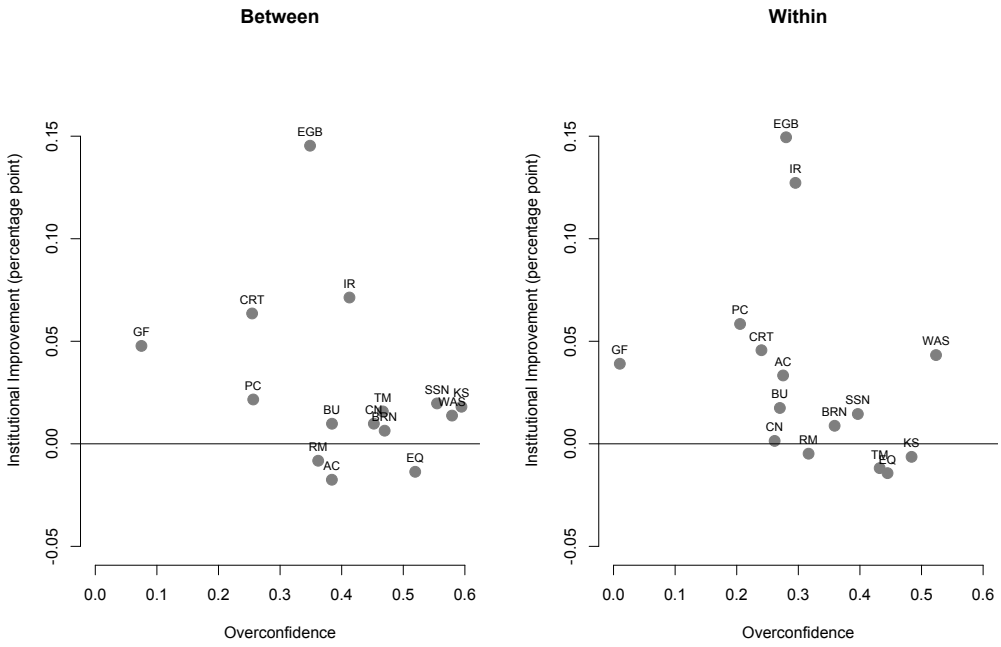


Figure 7: Overconfidence and institutional improvement. The left panel shows the results for the between-subjects treatments *Betting*, *Auction*, *Committee* and *Confidence*, while the right panel shows the results for *Betting Within*, *Auction Within* and *Committee Within*. Institutional improvement is averaged across institutions. Overconfidence is computed as average confidence minus the optimality rate in a task. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

searchers and 33% are graduate students. Over 85% of the sample indicated behavioral or experimental economics as their main field of expertise.

To keep the number of total predictions for each forecaster manageable, we picked one specific institution (*Auction*) and a subset of seven tasks.[12] Each expert made two separate sets of predictions for each task. First, we provided the raw optimality rate of answers to a given cognitive task and asked experts to predict the average optimality rate among the five winners of the auction. This allows us to compute predicted institutional improvement. Second, we asked experts to predict average confidence among subjects that took optimal / suboptimal decisions. This allows us to compare actual with predicted confidence calibration. Screenshots of the elicitation screens are reproduced in Appendix Figures 13 and 14.

Panel A of Figure 8 plots the median forecast of institutional improvement through the auction against actual improvement. Panel B plots the predicted difference in confidence between optimal and suboptimal decision makers against the corresponding empirical counterpart. The 45-degree lines represent the hypothetical case of perfect

[12]These tasks are RM, CN, TM, BRN, AC, CRT and EGB. Our expert elicitation also included WAS and BU, but due to a coding error the corresponding forecasts are not usable.

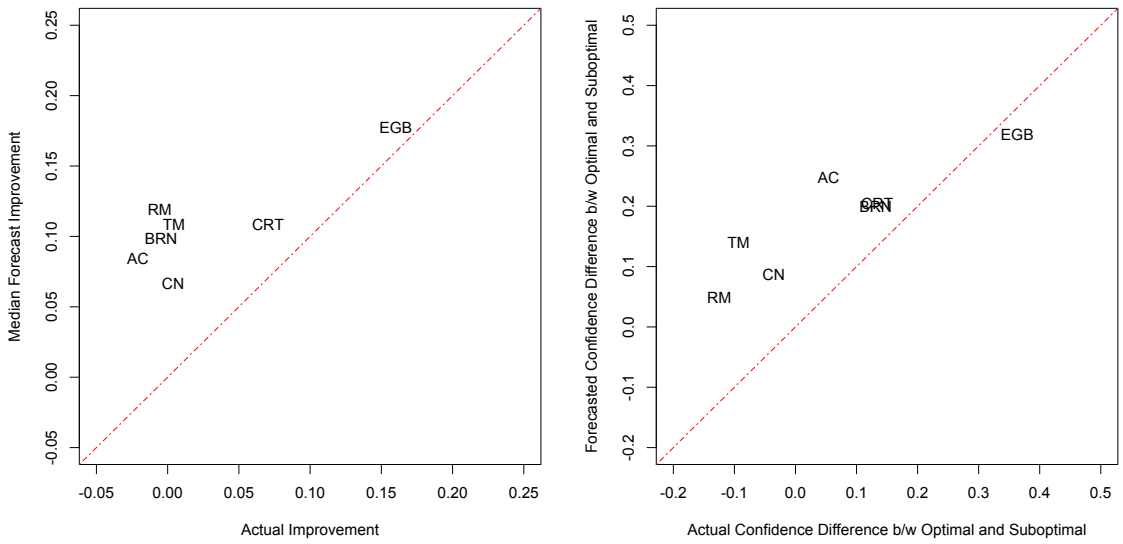


Figure 8: Expert forecasts and empirical analogues. The left panel plots predicted institutional improvements against actual ones. The right panel plots the predicted difference in confidence between optimal and suboptimal decisions against the true difference. The diagonal line indicates perfect calibration. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive Reflection Test; EGB=Exponential growth calc.; RM=Attribution; TM=Minimize taxation; WAS=Wason task.

calibration of the experts. The main takeaway is that the experts generally overpredict both the magnitude of institutional improvement and the degree of confidence calibration. These results are internally consistent with each other in our framework: if one believes that confidence is better calibrated than it actually is, then one should also believe that the institutional improvement will be larger than it actually is. Moreover, Panels A and B of Figure 8 show that the expert forecasts are excessively compressed relative to the truth: experts predict that the degree of confidence calibration and institutional improvement is more similar across tasks than is actually the case. At the same time, while the sample of nine tasks is very small, it does appear that while the experts are systematically overoptimistic about the degree of confidence calibration, they have a reasonably good qualitative sense of on which tasks experimental subjects are more or less well-calibrated.

7 Discussion

When we as experimental economists use average behavior in experiments to measure the severity of a bias, we are measuring a special case, that of no self-selection. Many of our most important and ubiquitous economic and social institutions create significant scope for self-selection out of decision-making, and this self-selection can produce rates

of bias in aggregate outcomes that differ from the raw rate of bias in the population. As a result, sample means from experiments may over- or under-state the influence biases are likely to have have for the aggregate outcomes produced by real-world institutions.

In this paper, we take some steps towards understanding the influence of self-selection over institutional outcomes for a wide range of biases, using maximally-simple variants of canonical institutions like speculative and allocative markets and organizations. We take a broad approach, studying 15 of the most famous and economically relevant biases from behavioral economics and three of the most important types of institutions in which self-selection can occur. We find that self-selection can have large effects on bias, but, more importantly, that the degree to which this is true varies wildly across distinct biases. We show that this heterogeneity is strongly related to heterogeneity in meta-cognition: the correlation between performance and confidence in the population.

Though our experiment takes a wide-ranging approach, we view it as a piece of a broader agenda. We see at least three ways in which future research could make use of and expand upon the preliminary findings reported here.[13]

First, our paper suggests a simple *methodological takeaway*. Future researchers can account for the likely impact of self-selection on the biases they measure, without undertaking the logistical challenges of implementing full-fledged social institutions. Our research suggests that simple and unincentivized measures of meta-cognition (confidence) can be used to produce an index of the susceptibility of biases to filtering by self-selection. Simply by (i) asking subjects at the end of an experiment how likely they think it is they made an optimal decision and (ii) reporting the correlation of this confidence measure with actual performance, researchers can provide evidence on how strongly self-selection should attenuate the impact of biases. We believe that this simple methodological blueprint can allow researchers to provide valuable context on the likely impact of lab-measured biases on real aggregate outcomes at very low cost. Given that our experts results suggest that researchers (ourselves included) do not have well-honed intuitions on how biases are likely to be influenced by self-selection, this is a practice we encourage for future research in experimental economics.

Second, it would be valuable for future research to apply our methods to *a broader range of social institutions and social mechanisms*. Our selection of institutions include some of the most salient in economics, including speculative and allocative markets and committee mechanisms of the sort that often govern organizations. But many other social mechanisms also produce potential scope for self-selection. For instance, cultural

[13]Beyond self-selection, there are a number of other ways institutions may work to reduce or magnify the impact of biases on economic outcomes. For instance, methods like ours could be profitably deployed to examine the role behavioral parameters and meta-cognition play in mechanisms that don't involve self-selection like wealth dynamics (the hypothesis that markets tend to systematically drive biased agents out of markets by impoverishing them) or learning.

transmission and social learning operate by mutual imitation. Whether imitation produces less biased aggregate outcomes (whether social learning "works") depends on whether the more biased tend to imitate the less biased or vice versa, a regularity that may well be driven by meta-cognition just as with the institutions we study here. To give another example, the effectiveness of persuasion (in, e.g., political debate, social media etc.) at reducing bias depends on whether the unbiased are more persuasive than the biased, and the biased more persuadable than the unbiased. This, too, may be directly related to meta-cognition. Equally important, our implementations of the institutions we study are the simplest and plausibly the least impactful versions of these institutions. For instance, we examine static contexts that feature no interaction, feedback or opportunities to mutually learn. Future research may examine dynamic betting markets, allocative markets and committees in which subjects can mutually adjust behavior, mutually evaluate relative confidence and possibly revise their biased beliefs. We conjecture that self-selection will have even stronger impacts on institutional outcomes in richer variations of the institutions we study here.

Finally, our work suggests that further study of *meta-cognition* is important for behavioral economics. Far from being a second-order psychological curiosity, meta-cognition may be of first-order importance for understanding the way findings of behavioral economics influence aggregate social science outcomes. Given that behavioral economists have traditionally invested substantially more effort documenting the existence of biases than on decision makers' awareness of them, we conjecture that more research energies might profitably be spent measuring and understanding meta-cognition and the sensitivity of meta-cognition to features of the choice environment. There are least three avenues that seem especially promising. (i) We have studied 15 salient biases from behavioral economics but there are dozens of others that could be similarly and retrospectively studied through the lens of meta-cognition. (ii) Alhough behavioral economists have put great energies into studying how nudges, frames, feedback, familiarity and learning influence biases themselves, we know next to nothing about how these same drivers of choice influence meta-cognition. For example, the effect of policy interventions on meta-cognition may be every bit as important for social science outcomes as their effect on biases themselves. (iii) For future theorizing and practical predictions, it would be very useful to understand *why* it is that in some tasks people's confidence in their decisions is reasonably well-calibrated, but not in others. Why do the "right" people sometimes believe that they are getting things wrong, and sometimes the "wrong" people?

References

Asparouhova, Elena, Peter Bossaerts, Jon Eguia, and William Zame, "Asset pricing and asymmetric reasoning," *Journal of Political Economy*, 2015, *123* (1), 66–122.

Bar-Hillel, M., "The base-rate fallacy in probability judgments," *Acta Psychologica*, 1980, *44*, 211–233.

Bar-Hillel, Maya, "The role of sample size in sample evaluation," *Organizational Behavior and Human Performance*, 1979, *24* (2), 245–257.

Barahona, Ricardo, Stefano Cassella, and Kristy AE Jansen, "Do Teams Alleviate or Exacerbate Cognitive Biases? Evidence from Extrapolation in Mutual Funds," *Working Paper*, 2021.

Benartzi, Shlomo and Richard H Thaler, "Naive diversification strategies in defined contribution saving plans," *American economic review*, 2001, *91* (1), 79–98.

Bó, Ernesto Dal, Pedro Dal Bó, and Erik Eyster, "The demand for bad policy when voters underappreciate equilibrium effects," *The Review of Economic Studies*, 2018, *85* (2), 964–998.

Bosch-Rosa, Ciril and Thomas Meissner, "The one player guessing game: a diagnosis on the relationship between equilibrium play, beliefs, and best responses," *Experimental Economics*, 2020, pp. 1–19.

Camerer, Colin and Dan Lovallo, "Overconfidence and Excess Entry: An Experimental Approach," *American Economic Review*, 1999, *89*, 306–318.

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

Charness, Gary and Dan Levin, "The Origin of the Winner's Curse: A Laboratory Study," *American Economic Journal: Microeconomics*, 2009, pp. 207–236.

_ **and Matthias Sutter**, "Groups make better self-interested decisions," *Journal of Economic Perspectives*, 2012, *26* (3), 157–76.

_ **, Edi Karni, and Dan Levin**, "Individual and group decision making under risk: An experimental study of Bayesian updating and violations of first-order stochastic dominance," *Journal of Risk and uncertainty*, 2007, *35* (2), 129–148.

Cooper, David J and John H Kagel, "Are two heads better than one? Team versus individual play in signaling games," *American Economic Review*, 2005, *95* (3), 477–509.

Danz, David, Lise Vesterlund, and Alistair J Wilson, "Belief elicitation: Limiting truth telling with information on incentives," Technical Report, National Bureau of Economic Research 2020.

Dohmen, Thomas, Armin Falk, David Huffman, Felix Marklein, and Uwe Sunde, "Biased probability judgment: Evidence of incidence and relationship to economic outcomes from a representative sample," *Journal of Economic Behavior & Organization*, 2009, *72* (3), 903–915.

Dunning, David, "The Dunning–Kruger effect: On being ignorant of one's own ignorance," in "Advances in experimental social psychology," Vol. 44, Elsevier, 2011, pp. 247–296.

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

_ **and Thomas Graeber**, "Cognitive uncertainty," Technical Report, National Bureau of Economic Research 2021.

_ **and** _ , "Cognitive uncertainty in intertemporal choice," *Working Paper*, 2021.

Erev, Ido, Thomas S Wallsten, and David V Budescu, "Simultaneous over-and underconfidence: The role of error in judgment processes.," *Psychological review*, 1994, *101* (3), 519.

Fehr, Ernst and Jean-Robert Tyran, "Individual irrationality and aggregate outcomes," *Journal of Economic Perspectives*, 2005, *19* (4), 43–66.

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

Friedman, Daniel, "laboratory financial markets," in "Behavioural and Experimental Economics," Springer, 2010, pp. 178–185.

Ganguly, Ananda R., John H. Kagel, and Donald V. Moser, "Do asset market prices reflect traders' judgment biases?," *Journal of Risk and Uncertainty*, 2000, *20* (3), 219–245.

Gupta, Neeraja, Luca Rigott, and Alistair Wilson, "The Experimenters' Dilemma: Inferential Preferences over Populations," *arXiv preprint arXiv:2107.05064*, 2021.

Haigh, Michael S and John A List, "Do professional traders exhibit myopic loss aversion? An experimental analysis," *The Journal of Finance*, 2005, *60* (1), 523–534.

Kahneman, Daniel, *Thinking, Fast and Slow*, Macmillan, 2011.

_ **and Amos Tversky**, "Subjective probability: A judgment of representativeness," *Cognitive psychology*, 1972, *3* (3), 430–454.

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

Klayman, Joshua, Jack B Soll, Claudia Gonzalez-Vallejo, and Sema Barlas, "Overconfidence: It depends on how, what, and whom you ask," *Organizational behavior and human decision processes*, 1999, *79* (3), 216–247.

Kluger, Brian D. and Steve B. Wyatt, "Are judgment errors reflected in market prices and allocations? Experimental evidence based on the Monty Hall problem," *Journal of Finance*, 2004, *59* (3), 969–998.

Koriat, Asher, Sarah Lichtenstein, and Baruch Fischhoff, "Reasons for confidence.," *Journal of Experimental Psychology: Human learning and memory*, 1980, *6* (2), 107.

Krishna, Vijay, *Auction theory*, Academic press, 2009.

Kruger, Justin and David Dunning, "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments.," *Journal of personality and social psychology*, 1999, *77* (6), 1121.

Lacetera, Nicola, Devin G Pope, and Justin R Sydnor, "Heuristic thinking and limited attention in the car market," *American Economic Review*, 2012, *102* (5), 2206–36.

Levy, Matthew and Joshua Tasoff, "Exponential-growth bias and lifecycle consumption," *Journal of the European Economic Association*, 2016, *14* (3), 545–583.

List, John A, "Does market experience eliminate market anomalies?," *The Quarterly Journal of Economics*, 2003, *118* (1), 41–71.

_ , "Neoclassical theory versus prospect theory: Evidence from the marketplace," *Econometrica*, 2004, *72* (2), 615–625.

Mathews, George B, "On the partition of numbers," *Proceedings of the London Mathematical Society*, 1896, *1* (1), 486–490.

Moore, Don A and Paul J Healy, "The trouble with overconfidence.," *Psychological review*, 2008, *115* (2), 502.

Murawski, Carsten and Peter Bossaerts, "How humans solve complex problems: The case of the Knapsack problem," *Scientific reports*, 2016, *6* (1), 1–10.

Plott, C., J. Wit, and W. Yang, "Parimutuel betting markets as information aggregation devices: experimental results," *Economic Theory*, 2003, *22*, 311–351.

Rees-Jones, Alex and Dmitry Taubinsky, "Measuring "schmeduling"," *The Review of Economic Studies*, 2020, *87* (5), 2399–2438.

Russell, Thomas and Richard Thaler, "The relevance of quasi rationality in competitive markets," *The American Economic Review*, 1985, *75* (5), 1071–1082.

Silver, Ike, Barbara A Mellers, and Philip E Tetlock, "Wise teamwork: Collective confidence calibration predicts the effectiveness of group discussion," *Journal of Experimental Social Psychology*, 2021, *96*, 104157.

Sonnemann, Ulrich, Colin F Camerer, Craig R Fox, and Thomas Langer, "How psychological framing affects economic market prices in the lab and field," *Proceedings of the National academy of Sciences*, 2013, *110* (29), 11779–11784.

Thompson, Valerie A, *Dual-process theories: A metacognitive perspective.*, Oxford University Press, 2009.

_ , Jamie A Prowse Turner, and Gordon Pennycook, "Intuition, reason, and metacognition," *Cognitive psychology*, 2011, *63* (3), 107–140.

Tversky, Amos and Daniel Kahneman, "Evidential Impact of Base Rates. Judgment under Uncertainty: Heuristics and Biases. D. Kahneman, P. Slovic and A. Tversky," 1982.

Wason, Peter C, "Reasoning about a rule," *Quarterly journal of experimental psychology*, 1968, *20* (3), 273–281.

ONLINE APPENDIX

A More on the Cognitive Tasks

We describe the 15 cognitive tasks in some detail here. See Appendix E for the full task instructions.

Base rate neglect (BRN). An important principle of rational information-processing is to take into account base rates, but a voluminous line of work documents that people tend to neglect base rates (e.g., Bar-Hillel, 1980; Kahneman and Tversky, 1973). We devised a simple variant of Tversky and Kahneman's (1982) well-known taxi-cab problem, which is known to generate responses that neglect the base rate. In our problem, a quality control machine of a bike manufacturer classifies bikes as good or defective but misclassifies any given bike 25% of the time. Subjects are asked to state the percentage chance that a bike is actually defective, given that the base rate for a defective bike is 10% and that the quality control machine classifies the bike as defective. A common incorrect answer is 75%, whereas the statistically correct answer is 25%.

Correlation neglect (CN). Taking into account potential non-independence of data is a core principle of both rational belief updating and econometrics courses. We devised a simplified version of the correlation neglect task developed in Enke and Zimmermann (2019). Subjects are asked to estimate the weight of a bucket. The hypothetical characters Ann and Bob have each examined the bucket and produced an estimate. They share their estimate with Charlie, who computes the average of their two guesses. The subject has access to Ann's estimate of 70 and Charlie's estimate of 40. A common incorrect answer is to compute the average of 40 and 70, i.e., 55. The correct answer is 40.

Balls-and-urns belief updating (BU). A widely used paradigm to study belief updating are so-called balls-and-urns experiments. In our setup, there are two bags. One contains 70 red and 30 blue chips, and one contains 30 red and 70 blue chips. One of them gets selected at random with 50-50 chance, and a balls gets drawn from the selected bag. Subjects are asked to indicate the percentage chance that the selected bag is the one that contains more red chips, given that the drawn chip is red. Subjects commonly exhibit a conservatism bias in this setup and state posteriors strictly between 50% and 70%. The Bayesian answer is 70%.

Gambler's fallacy (GF). Following Dohmen et al. (2009), subjects were asked to predict the next outcome in the following sequence of tosses of a fair coin, where the last three tosses came up Heads: T - T - T - H - T - H - H - H . In this type of task, subjects occasionally believe that Tails is "due." The correct answer is to select that "Both are equally likely."

Sample size neglect (SSN). Adapting the classic "hospital problem" (Kahneman and Tversky, 1972; Bar-Hillel, 1979), subjects were asked whether a factory that produces 45 chairs each day or a factory that produces 15 chairs each day has more days on which more than 20% of chairs are defective. A common incorrect response is that this happens equally often in both factories. The correct answer, however, is that this outcome is more likely in the smaller factory.

Regression to the mean / misattribution (RM). People exhibit a well-known tendency to attribute outcomes to internal factors rather than random noise, which leads them to neglect mean reversion. In a variant of classical work on the failure to appreciate regression (Kahneman and Tversky, 1973), subjects are asked to assess whether the true IQ of a test-taker is more likely to be above or below 140, given that their IQ test score is 140, the average score in the sample is 100, and test scores reflect a combination of true ability and random chance. Subjects commonly believe that the person's true IQ is equally likely to be above or below 140. The correct answer is that the person's true IQ is more likely to be below than above 140. Such failure to account for mean reversion is a special case of a more general class of biases that reflect misattribution.

Acquiring-a-company (AC). This game is one of the most widely studied tasks in experimental economics, both because it reflects a general class of errors in contingent reasoning and because its adverse selection logic has many applications in economics, such as in auctions. Following Charness and Levin (2009), we implement a version of the AC game in which subjects play against a computerized opponent. A seller has a company that is worth either 20 or 120 points to him. The company's value to the buyer is 1.5 times as much as the value to the seller. The subject makes a take-it-or-leave-it-offer, which the seller accepts if and only if the offer is at least as high as the value of the company to him. Subjects frequently bid strictly more than the theoretically optimal bid of 20, which neglects the adverse selection logic of the problem.

Wason selection task (WAS). The Wason task (Wason, 1968) is a widely used task in the social sciences because it captures failures in contingent reasoning and a tendency towards positive hypothesis testing in a simple way. In the problem, there is a deck of

four cards that have numbers (odd or even) on one side and a color (brown or green) on the other. Subjects are tasked with turning over (only) those cards that can be useful in assessing whether the statement "All of the cards with an even number on one side are green on the other" is true. In this task, subjects frequently engage in positive testing by turning over the card with the even number and the one that is green. The logically correct choice is to pick the brown card in addition to the card with the even number.

Cognitive Reflection Test (CRT). The CRT is likewise a widely studied task in economics because it effectively captures the intuitive "System 1" responses that the early heuristics and biases program emphasized. Moreover, responses in the CRT have been shown to be correlated with various economic behaviors (Frederick, 2005). We implement the question "It takes 6 machines 6 days to produce 6 cars. How long would it take 12 machines to produce 12 cars?" A tempting, incorrect answer is 12 days. The correct answer is 6 days.

Iterated reasoning / Backward induction (IR). To capture the well-known and widely-studied tendency to iterate the best-response function only a small number of times ("level k reasoning"), we follow Bosch-Rosa and Meissner (2020) in implementing a one-player guessing game that only requires iteration of the best response function but is independent of beliefs about others. Subjects are asked to pick two numbers between 0 and 100, inclusive. Their task is to select numbers whose average is as close as possible to 2/3 of either number. While zero is the correct solution, most subjects state strictly positive values.

Equilibrium Reasoning / Predicting response to incentives (EQ). Many empirical regularities in behavioral economics can be understood as people failing to accurately predict others' behavior from their incentives. One elegant demonstration of this is the experiment in Dal Bó et al. (2018), which we simplify here. Subjects are presented with two similar-looking 2×2 payoff matrices. In Game A, all payoffs are higher than in Game B, but Game B has a cooperative equilibrium, while Game A has a prisoner's dilemma structure. Subjects are asked to predict in which game past participants made more money, on average. A majority of subjects incorrectly believes that people make more in Game A because they fail to anticipate differences in equilibrium play.

Knapsack / identifying constrained optima (KS). Knapsack problems are a simple to explain but canonical instance of constrained optimization, which lies at the heart of a large class of economic consumer and firm maximization problems (Mathews, 1896; Murawski and Bossaerts, 2016). In our implementation, subjects pick from a set of 12

items, each of which has a known value and weight. The objective is to maximize the sum of values chosen, while satisfying a budget constraint on the weights. Experiments typically show that subjects fail to identify the value-maximizing bundle (which, in our instance consists of 4 of the 12 items).

Portfolio choice (PC). Various studies have documented failures to construct efficient financial portfolios. One well-known example of this is failure is the use of the so-called 1/N heuristic (Benartzi and Thaler, 2001), according to which people split their investment uniformly across the available assets. To get at this, we ask subjects to choose between two portfolios that are constructed from four assets each. The portfolios are identical except that one allocates 1/4 of the budget to each asset in a way that makes this "1/N portfolio" strictly dominated by the other available portfolio.

Thinking at the margin (TM). One of the core lessons of economics is to think at the margin rather than the average, yet people have consistently been shown to think in terms of averages. We developed a simplified version of the taxation problem in Rees-Jones and Taubinsky (2020). Subjects are tasked with deciding which of two bank accounts to allocate 20 points to. Both bank accounts already contain 40 points. The trick is that the marginal tax rate for additional 20 points is lower in the bank account that has a higher average tax rate.

Exponential growth bias (EGB). An ability to compute compound interest is essential in numerous economics models and decision contexts, including exponential time discounting, savings and investment. Following previous studies (Levy and Tasoff, 2016), we ask subjects to guess how much a stock that is worth $100 today is worth after 20 years if its value increases 5% each year. People tend to give a response of $200 in this problem, which entirely misses the compounding effect. The correct answer is $265.

B Model Derivations

In the following, we derive our main predictions separately for the model of committee voting or parimutuel betting (Appendix B.1) and for the model of auctions (Appendix B.2).

B.1 Committees and Parimutuel Betting Markets

Note that expanding eq. (6) from Section 3, institutional gain $\mathbb{G}_{bet,com}$ may be expressed as:

$$\mathbb{G}_{bet,com} = \frac{\sum_i p_i \left(N k_i - \sum_j k_j \right)}{N^2 \bar{k}} \tag{10}$$

$$= \omega\beta \frac{\sum_i p_i (p_i - \bar{p})}{N\bar{k}} \tag{11}$$

$$= \frac{(N-1)\beta}{N\bar{c}} s_p^2 \tag{12}$$

where s_p^2 is the sample variance of p. Informally, we may consider $\beta/\bar{k} = \partial \log \bar{k} / \partial \bar{p}$, as the percent increase in the average bid given an increase in the average confidence among the players.

Proof. **(Prediction 1, Part (i))** Stated formally, this part of the prediction states that if $s_p^2 > 0$ and $\beta > 0$, we have $\mathbb{G}_{bet,com} > 0$. Since $\bar{c} > 0$ in our case[14], eq. (12) immediately provides the result. □

Proof. **(Prediction 1, Part (ii))** This part of the prediction says that for $\alpha > 0$ and $s_p^2 > 0$, we have $\frac{\partial \mathbb{G}_{bet,com}}{\partial \beta} > 0$. Using eq. (12) we may compute

$$\frac{\partial}{\partial \beta} \mathbb{G}_{bet,com} = \frac{\partial}{\partial \beta} \frac{(N-1)\beta}{N(\alpha + \beta \bar{p})} s_p^2 \tag{13}$$

$$= \frac{(N-1)\alpha}{N(\alpha + \beta \bar{p})^2} s_p^2 > 0 \tag{14}$$

Given our assumptions the result follows. □

Proof. **(Prediction 2)** This prediction states that the effect of an increase in average overconfidence, d, on institutional improvement, $\mathbb{G}$, is ambiguous. Formally, for the case of betting and committee voting, we predict that for $\beta > 0$ and $s_p^2 > 0$, a change in overconfidence $d \to d + \delta d$ need not imply $\Delta \mathbb{G}_{bet,com} < 0$.

Firstly, we may compute that: $\frac{\partial \mathbb{G}_{bet,com}}{\partial \alpha} < 0$. Using eq. (12) we may compute

$$\frac{\partial}{\partial \alpha} \mathbb{G}_{bet,com} = \frac{\partial}{\partial \alpha} \frac{(N-1)\beta}{N(\alpha + \beta \bar{p})} s_p^2 \tag{15}$$

$$= -\frac{(N-1)\beta}{N(\alpha + \beta \bar{p})^2} s_p^2 < 0 \tag{16}$$

[14] Aside from the empirics, $c_i > 0$ when $\alpha, \beta > 0$.

Secondly, we recall our earlier result that:

$$\frac{\partial}{\partial \beta}\mathbb{G}_{bet,com} > 0 \tag{17}$$

Noting that:

$$\frac{\partial d}{\partial \alpha}, \ \frac{\partial d}{\partial \beta} > 0 \tag{18}$$

and our assumptions the result follows. □

B.2 Auctions

In the case of a multi-unit auction, eq. (7) showed that:

$$\mathbb{G}_{auc} = \frac{1}{|W|}\sum_{j\in W} p_j - \frac{1}{N}\sum_i p_i. \tag{19}$$

Proof. **(Prediction 1, Part (i))** This part states that when $\beta > 0$, we have $\mathbb{G}_{auc} \geq 0$. Note that when $\beta > 0$, $\frac{1}{|W|}\sum_{i\in W} p_i \geq \bar{p}$. The result then follows. □

Proof. **(Prediction 1, Part (ii))** When $\beta > 0$, we have $\frac{\partial \mathbb{G}_{auc}}{\partial \beta} = 0$. This holds because given a fixed set of probabilities, $\{p_i\}$, the ordering of the bids is invariant under a change, $\beta \to \beta + \delta\beta$ so long as $\delta\beta > -\beta$. □

Proof. **(Prediction 2)** This prediction states that an increase in overconfidence, d, has no effect on institutional improvement $\mathbb{G}$ when $\beta > 0$. In the case of auctions, we have $\frac{\partial \mathbb{G}_{auc}}{\partial \alpha} = 0$. This follows since given a fixed set of probabilities, $\{p_i\}$, the ordering of the bids is invariant under a change, $\alpha \to \alpha + \delta\alpha$. Furthermore, the proof of Pred. 1 part (ii) indicates that a change in β doesn't change $\mathbb{G}$ so long as β remains positive. Accordingly, for any change in average overconfidence, $\delta d = \delta\alpha + \delta\beta\bar{p}$ in which $\beta > 0$, we see that there is no institutional improvement. □

C Additional Figures

C.1 Screenshots of Elicitation Screens

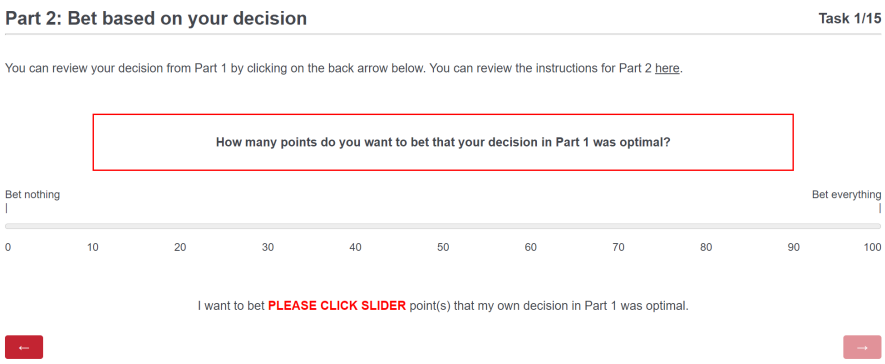


Figure 9: Screenshot of elicitation screen for the institutional decision in treatment *Betting*.

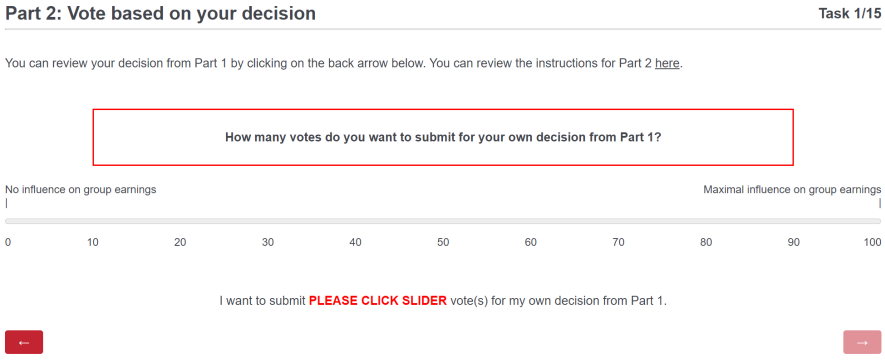


Figure 10: Screenshot of elicitation screen for the institutional decision in treatment *Committee*.

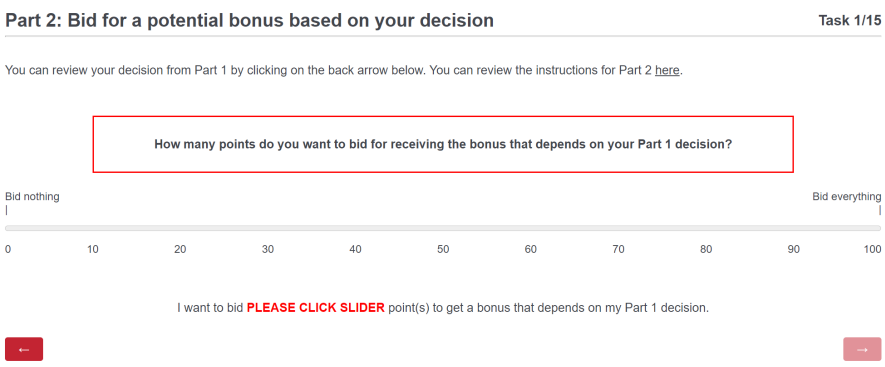


Figure 11: Screenshot of elicitation screen for the institutional decision in treatment *Auction*.

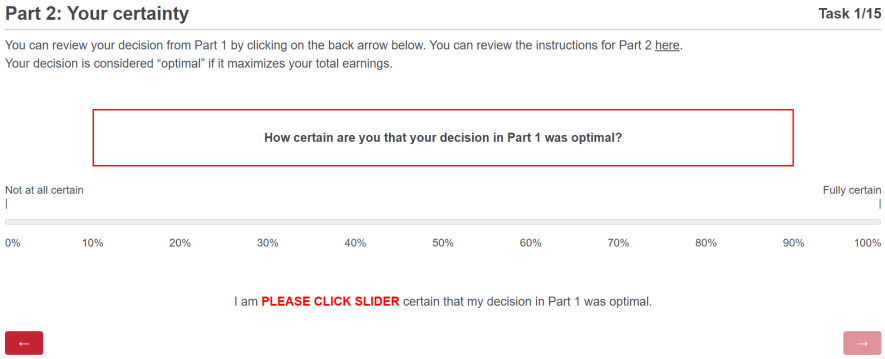


Figure 12: Screenshot of elicitation screen for confidence in the Knapsack task (treatment *Confidence*).

Predict average confidence in the cognitive tasks

For each of the tasks shown below, please predict the average confidence separately for subjects who took the optimal decision and those who did not. Confidence was elicited by asking subjects to provide their assessment of how likely it is that their decision was optimal. 0% means "Not at all certain" and 100% means "Fully certain".

- Click [here] for a reminder of the prediction task instructions and [here] for the study background and general instructions. Instructions will open in a new window.
- Click on the task name to see a screenshot of the actual task screen, including a definition of what "optimal" means in the respective task.

	Your prediction: Average confidence stated for answers that were **optimal**	**Your prediction**: Average confidence stated for answers that were **not optimal**
Wason selection task: Failure to gather valuable evidence via positive hypothesis testing bias. Adaptation of 4-card task from Wason (1968).	%	%
Balls-and-urns belief updating: Failure to calculate Bayesian posterior. Give probabilistic beliefs about which urn a colored ball is drawn from (Edwards, 1968).	%	%
Regression to the mean: Failing to account for noise in outcomes via failure to recognize regression to the mean. Adaptation of task from Kahneman and Tversky (1973).	%	%
Thinking at the margin: Thinking about average instead of marginal costs/benefits. Adaptation of marginal taxation task from Rees-Jones and Taubinsky (2020).	%	%
Acquiring-a-company: Failing to properly condition on contingencies as in the Winner's Curse. Bidding task against computer as in Charness and Levin (2009).	%	%
Exponential growth calculation: Underestimate the exponential effects of compounding.Interest rate forecasting problem adapted from Levy and Tasoff (2016).	%	%
Correlation neglect: Failing to account for non-independence of data in inference. Adaptation of tasks from Enke and Zimmermann (2019)	%	%
Base rate neglect: Ignoring base rates when computing posteriors. Adaptation of taxi-cab problem from Tversky and Kahneman (1982).	%	%
Cognitive reflection test: Following intuitive but misleading "System 1" intuitions. Adaptation of one of the problems by Frederick (2005).	%	%

Figure 13: Screenshot of elicitation screen in the expert survey for forecasts of confidence.

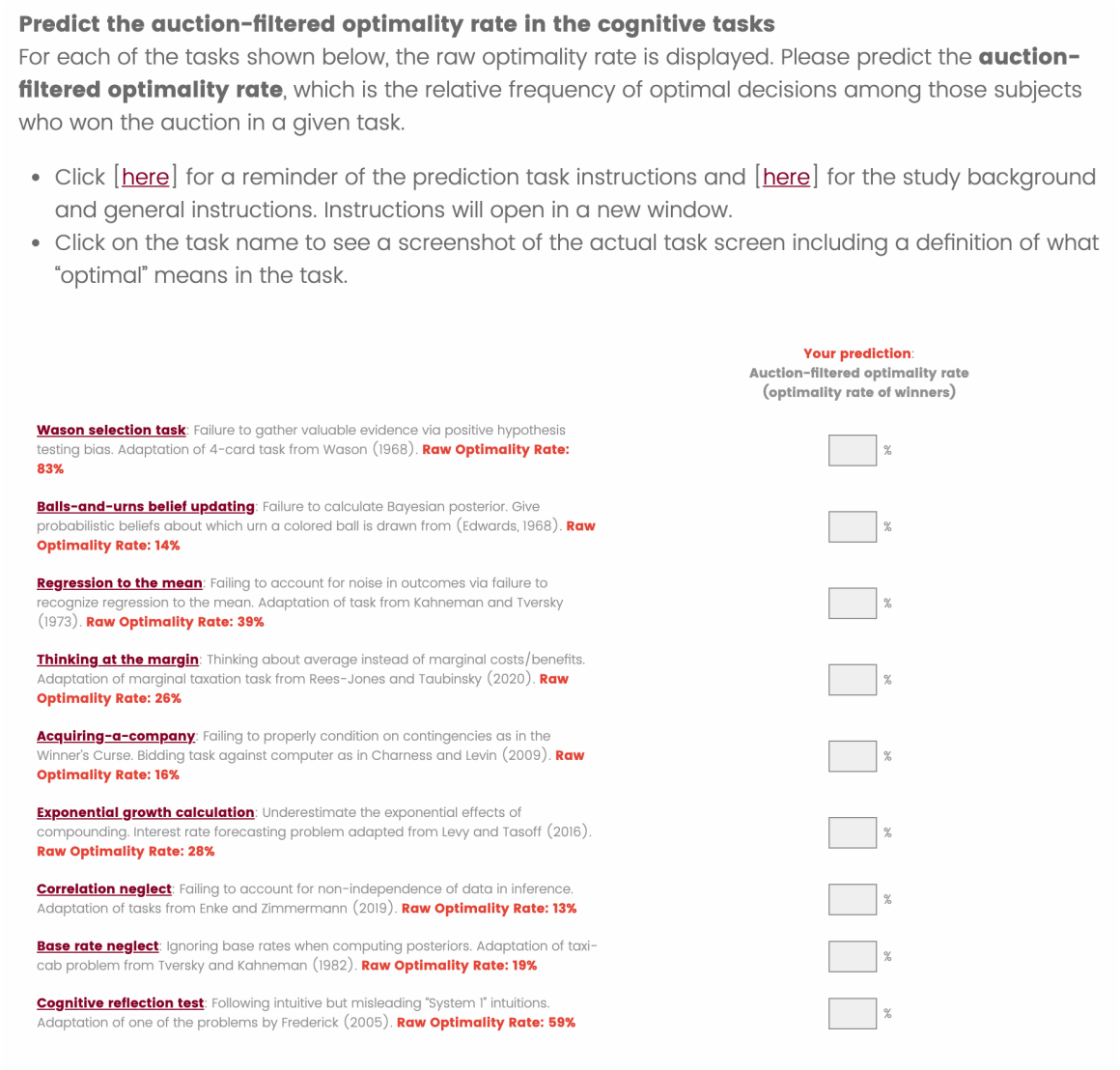


Figure 14: Screenshot of elicitation screen in the expert survey for forecasts of the change in optimality rate through the *Auction* institution.

C.2 Additional Figures for Between-Subjects Treatments

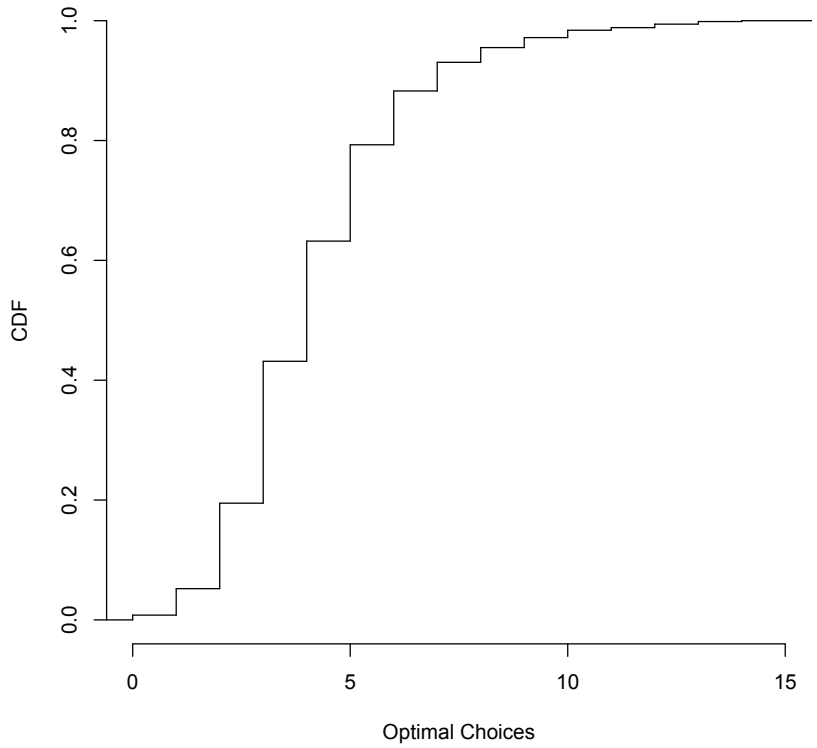


Figure 15: Empirical cumulative distribution of total number of optimal Part 1 responses, across subjects in treatments *Betting*, *Auction*, *Committee* and *Confidence*. The figure pools data from all 15 tasks across all four between-subject treatments.

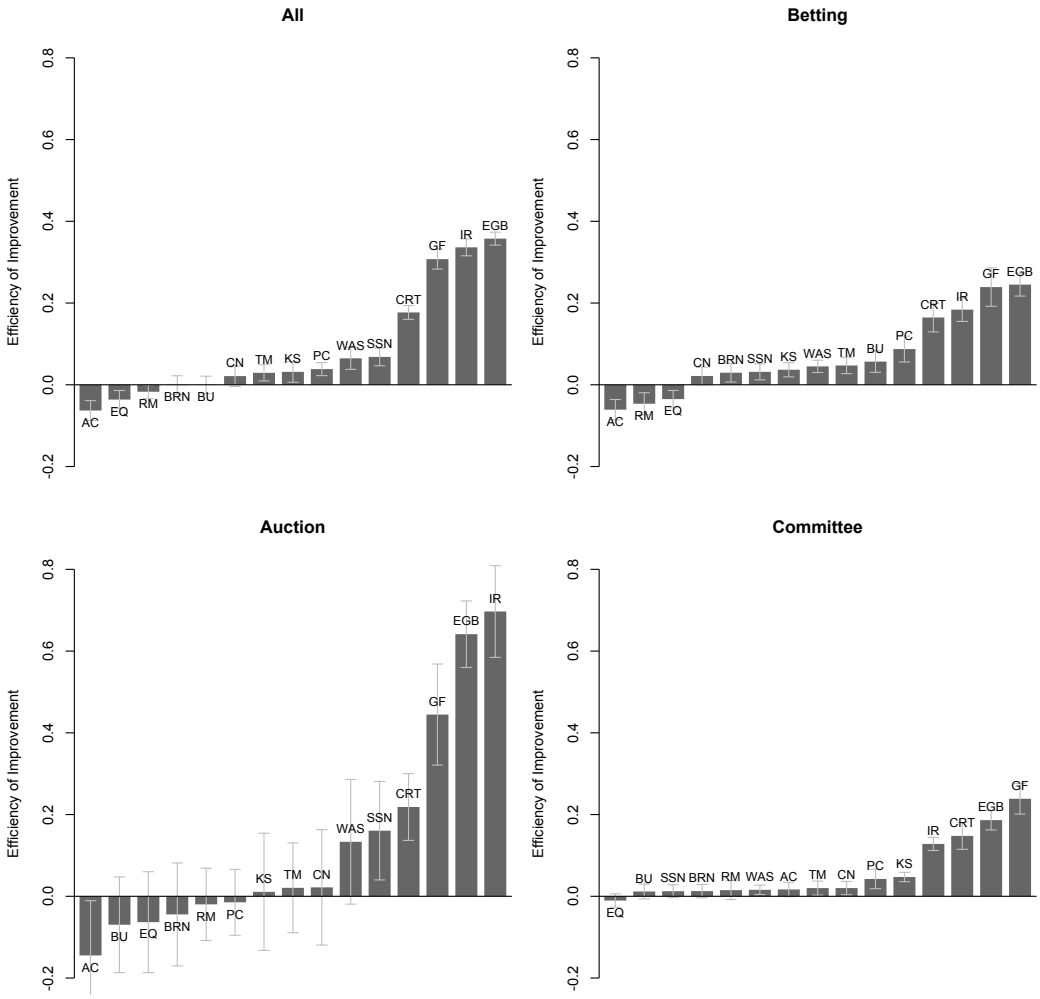


Figure 16: Performance improvement (efficiency measure) through institutions across tasks. Efficiency is computed as the aggregate performance rate after institutional filtering minus the the fraction of optimal Part 1 responses, all divided by the difference between the maximum possible improvement (given Part 1 responses and the structure of the institution) and the fraction of optimal Part 1 responses. The aggregate performance rate is based on 10,000 randomly constructed 10-subject cohorts for each institution, taking the mean over all samples. Each participant completed all 15 tasks in random order. Based on $N = 323$ participants in the Auction condition, $N = 387$ in Betting and $N = 337$ in Committee. One-standard error bars are conservatively calculated as the ratio of the standard deviations of efficiencies over these random cohorts divided by the square root of the number of cohorts available in the dataset (e.g., 387/10=38.7 in Betting). Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

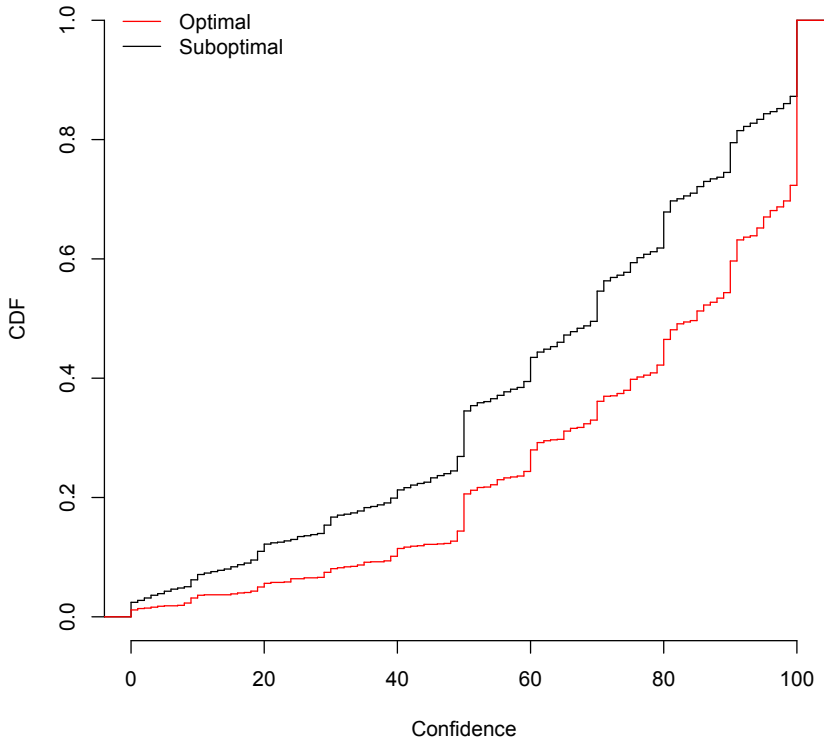


Figure 17: Part 2 stated confidence by Part 1 optimality. Based on $N = 5,010$ Part 2 decisions in the *Confidence* condition, pooled across 15 different cognitive tasks. The sample of Part 2 decisions is split by whether the corresponding Part 1 decision was optimal and empirical distribution functions are displayed.

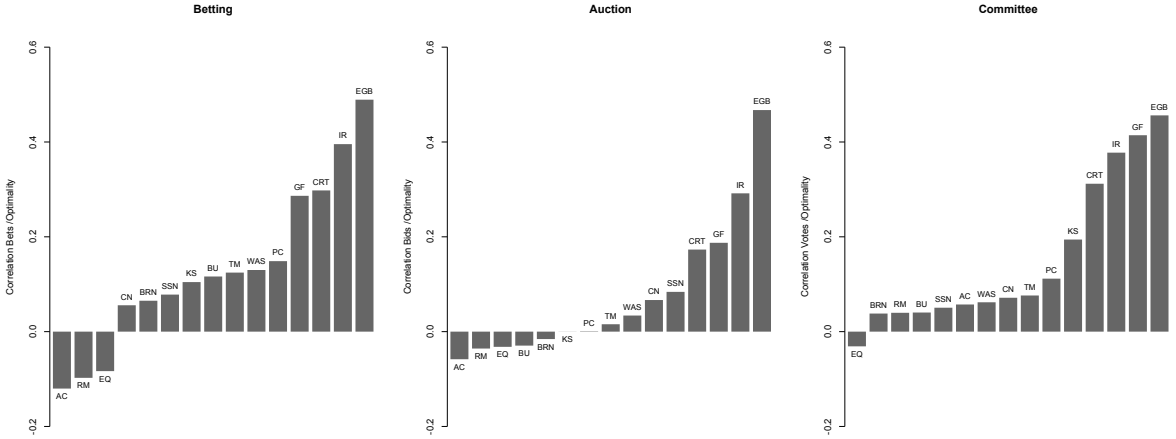


Figure 18: Correlation between Part 1 optimality and institutional decision by task. Based on $N = 323$ participants in the Auction condition, $N = 387$ in Betting and $N = 337$ in Committee. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

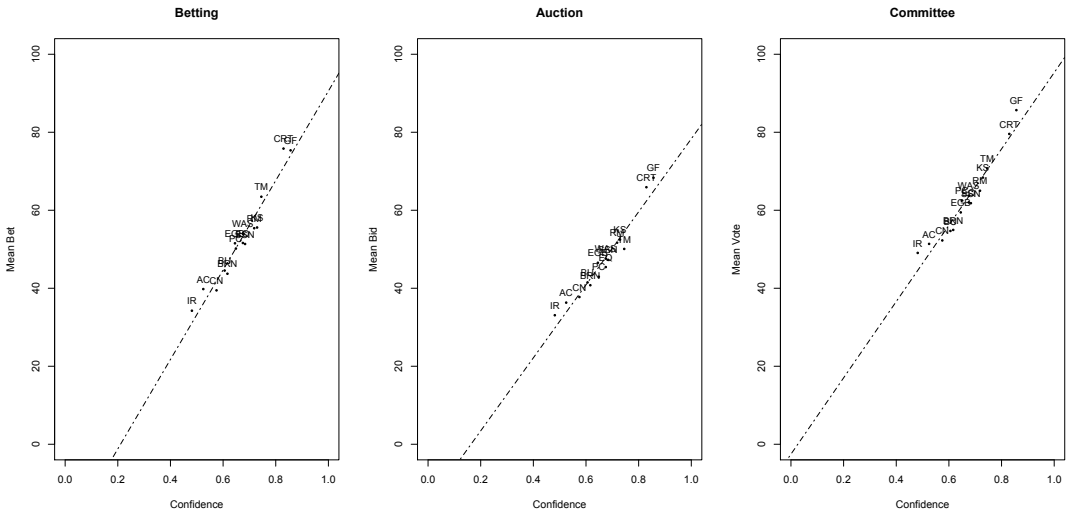


Figure 19: Average confidence and average institutional behavior. For each task, we compute average confidence stated in treatment *Confidence* and plot this against the average Part 2 decision taken in each of the three between-subjects institutional treatments.

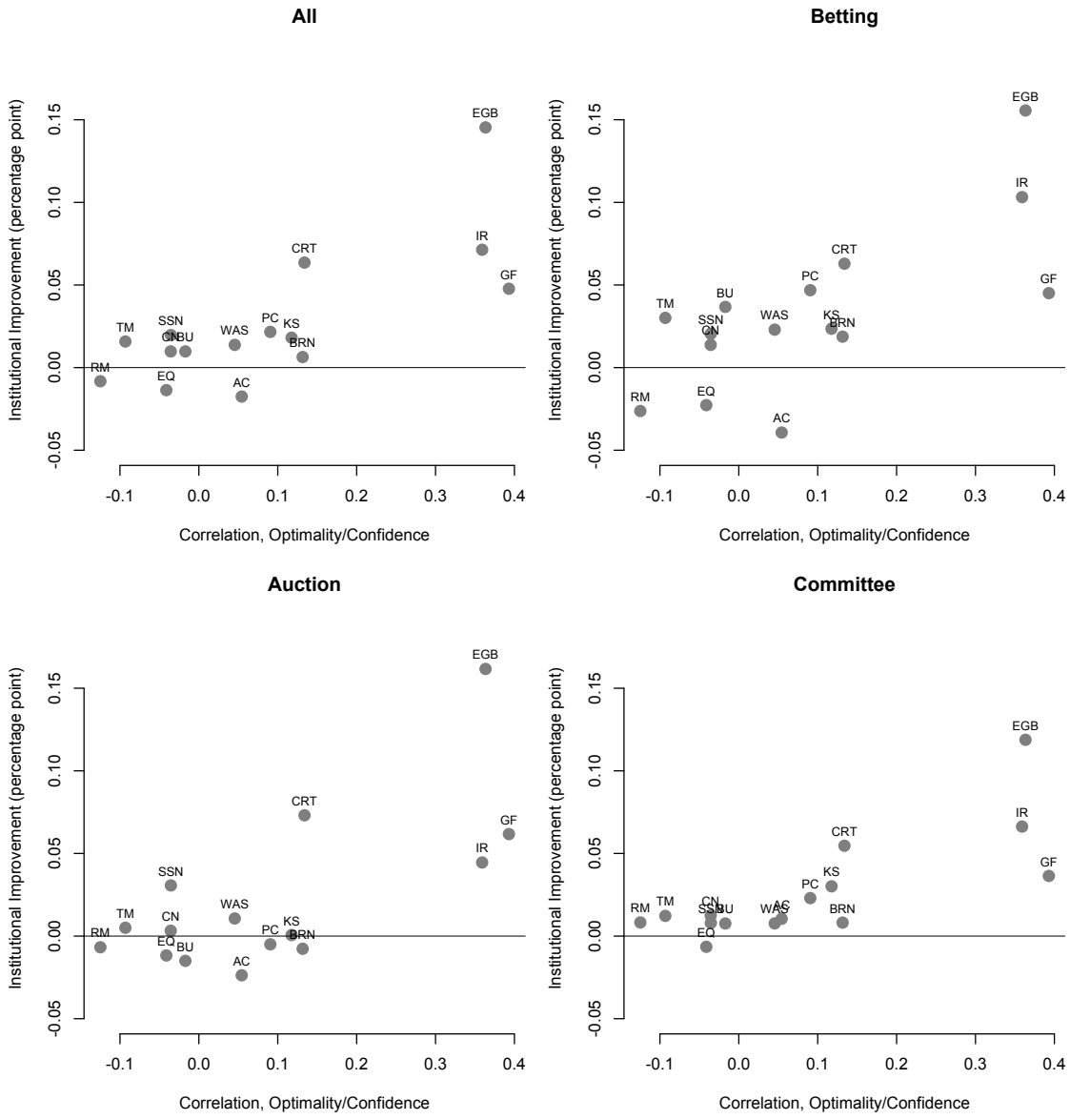


Figure 20: Confidence-optimality correlation and institutional improvement for the separate institutions in the between-subjects treatments. The horizontal axis shows the within-task correlation between confidence and optimality in a given task in treatment *Confidence*. The vertical axis shows the performance improvement that is implied by an institution for the respective cognitive task in treatments *Betting*, *Auction* and *Committee*. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

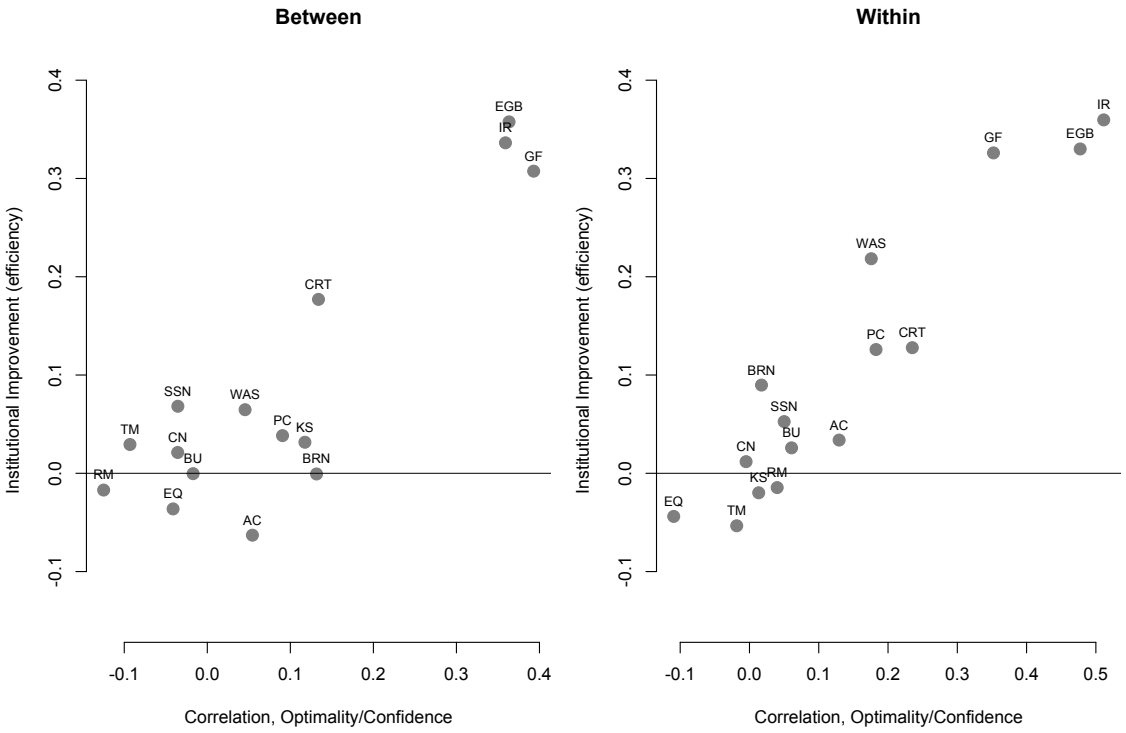


Figure 21: Confidence-optimality correlation and institutional improvement using the efficiency measure. The left panel shows the results for the between-subjects treatments and the right panel those for the within-subjects treatments. In the left panel, the horizontal axis shows the within-task correlation between confidence and optimality in a given task in treatment *Confidence*. The vertical axis shows the efficiency of an institution for the respective cognitive task. Efficiency is computed as the aggregate performance rate after institutional filtering minus the the fraction of optimal Part 1 responses, all divided by the difference between the maximum possible improvement (given Part 1 responses and the structure of the institution) and the fraction of optimal Part 1 responses. The data are from treatments *Betting*, *Auction* and *Committee*. In the right panel, we show analogous quantities, except that they are all derived from treatments *Betting Within*, *Auction Within* and *Committee Within*. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

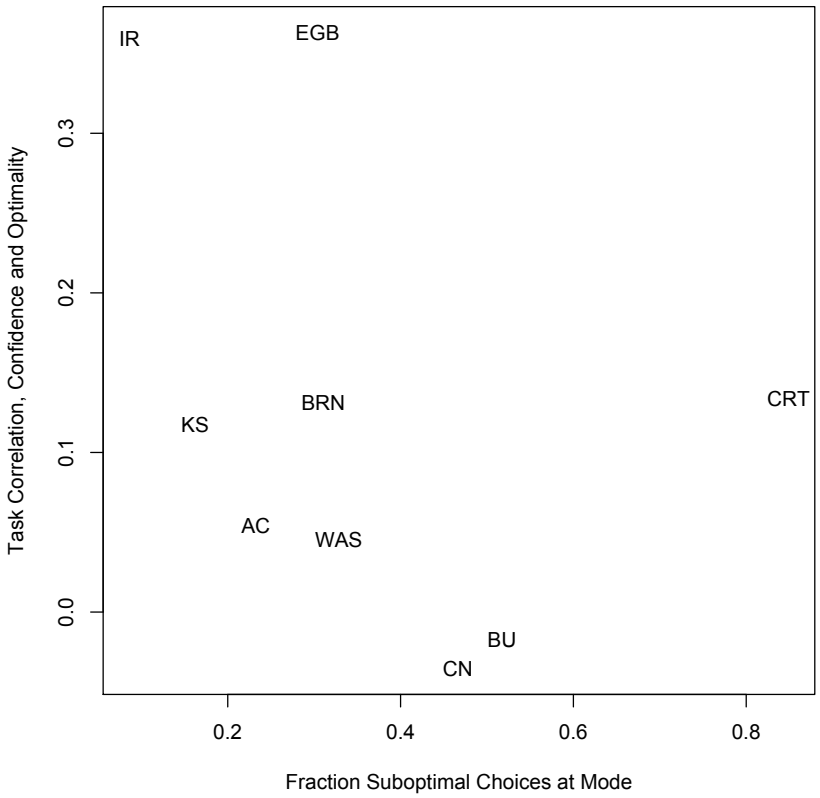


Figure 22: The peakedness of the distribution of suboptimal answers and institutional improvement in the between-subjects treatments. The horizontal axis displays the fraction of subjects playing the modal suboptimal Part 1 answer. The vertical axis displays the confidence-optimality correlation in treatment *Confidence*. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; IR=Backwards induction; KS=Knapsack; WAS=Wason task.

C.3 Additional Figures for Within-Subjects Treatments

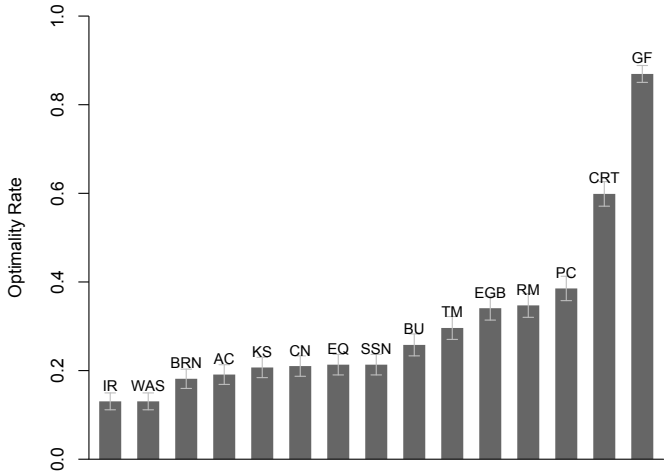


Figure 23: Fraction of optimal Part 1 responses across tasks in treatments *Betting Within*, *Auction Within* and *Committee Within*. The tasks and optimal responses are described in Appendices A and E. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Thinking at the margin; WAS=Wason task. Error bars are the standard error of the binomial mean.

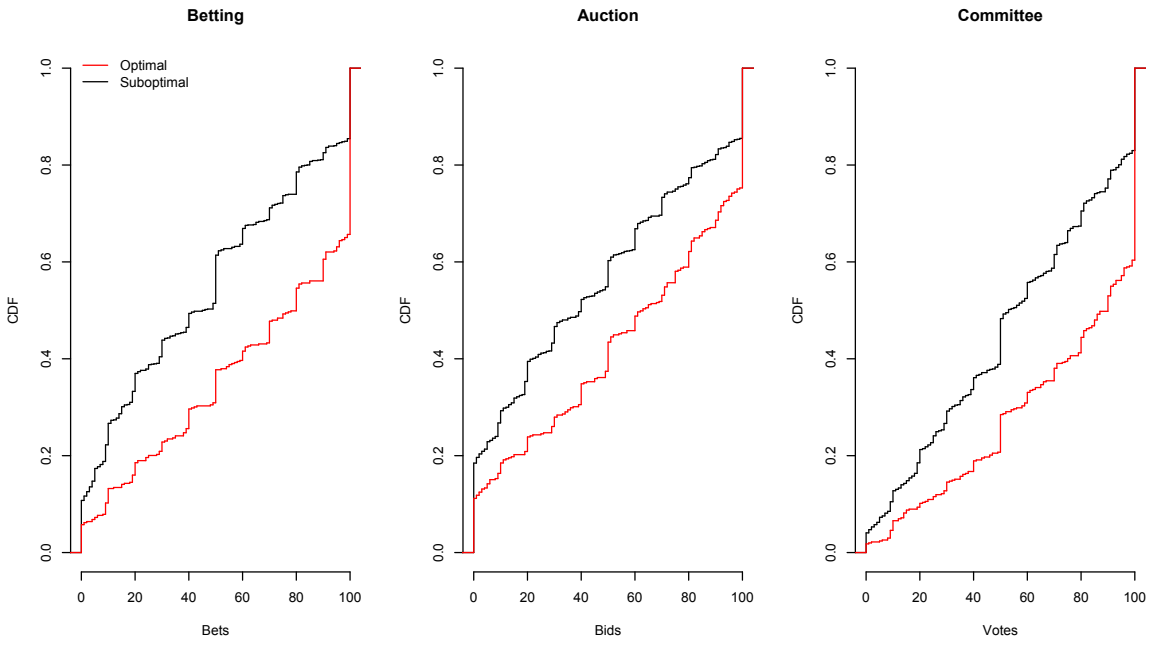


Figure 24: Part 2 institutional behavior by Part 1 optimality. Based on $N = 1575$ Part 2 decisions in the *Auction Within* condition, $N = 1575$ in *Betting Within* and $N = 1560$ in *Committee Within*, pooled across 15 different cognitive tasks. For each institution, the sample of Part 2 decisions is split by whether the corresponding Part 1 decision was optimal and empirical distribution functions are displayed.

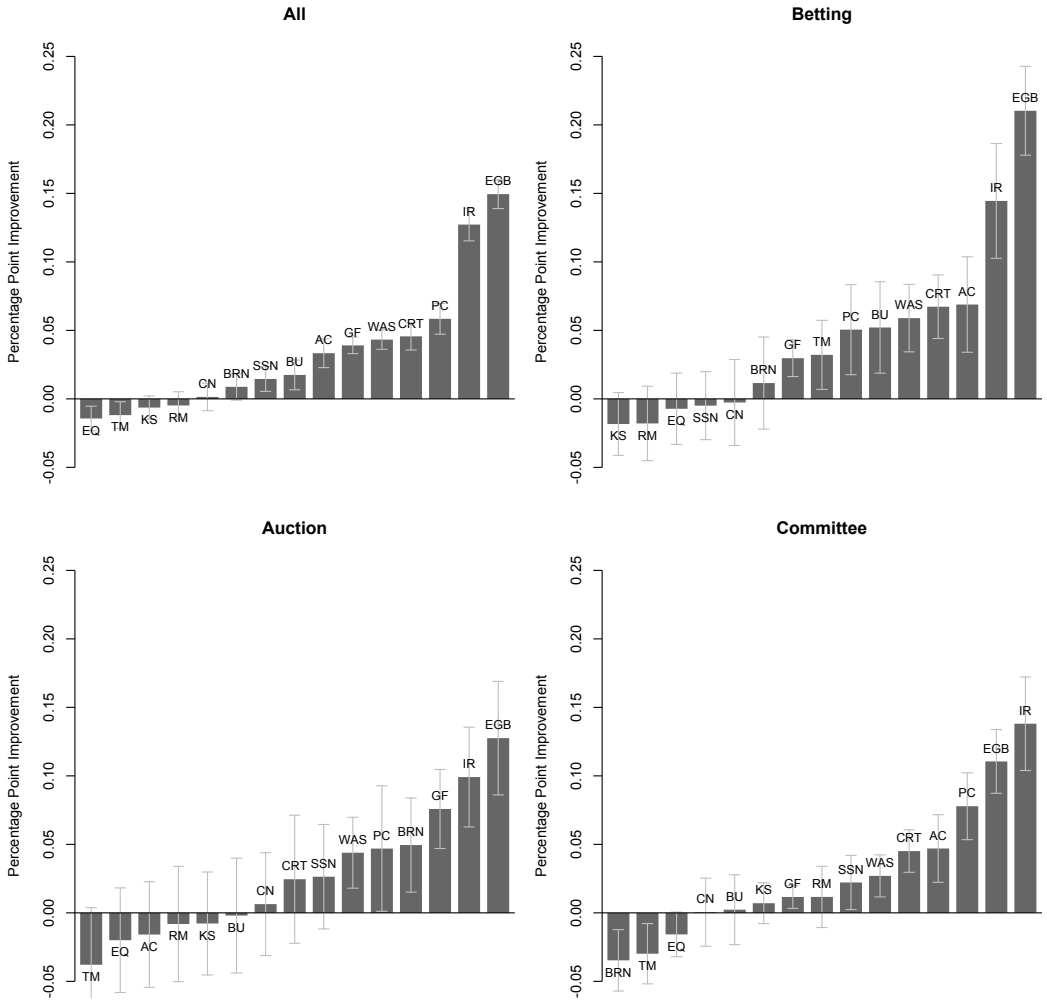


Figure 25: Performance improvement through institutions across tasks in treatments *Betting Within*, *Auction Within* and *Committee Within*. Percentage point improvement is computed as the aggregate performance rate after institutional filtering minus the fraction of optimal Part 1 responses. The aggregate performance rate is based on 10,000 randomly constructed 10-subject cohorts for each institution, taking the mean over all samples. Each participant completed all 15 tasks in random order. Based on $N = 105$ participants in the Auction condition, $N = 105$ in Betting and $N = 104$ in Committee. One-standard error bars are conservatively calculated as the ratio of the standard deviations of improvements over these random cohorts divided by the square root of the number of cohorts available in the dataset (e.g., 105/10=10.5 in Betting). Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

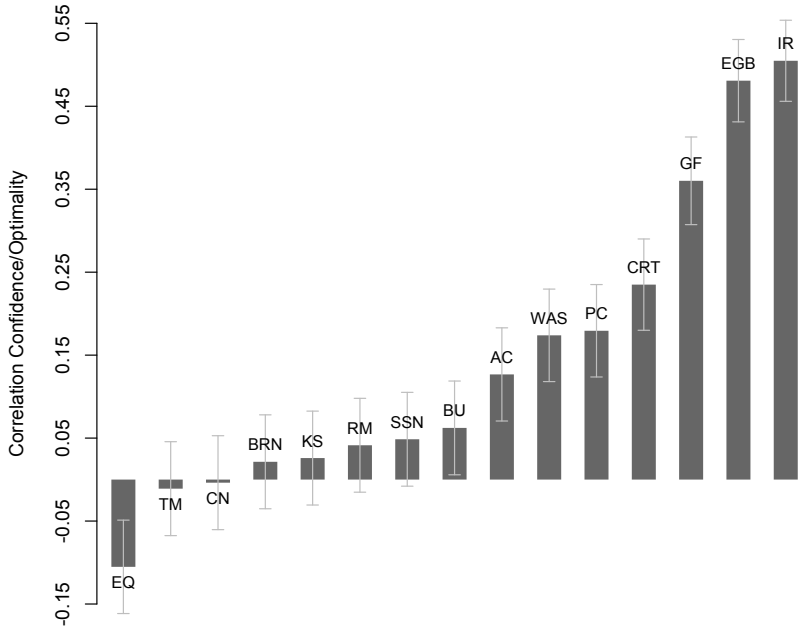


Figure 26: Within-task correlation between confidence and Part 1 optimality across tasks in treatments *Betting Within*, *Auction Within* and *Committee Within*. Displayed are Pearson correlation coefficients, based on $N = 313$ participants. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

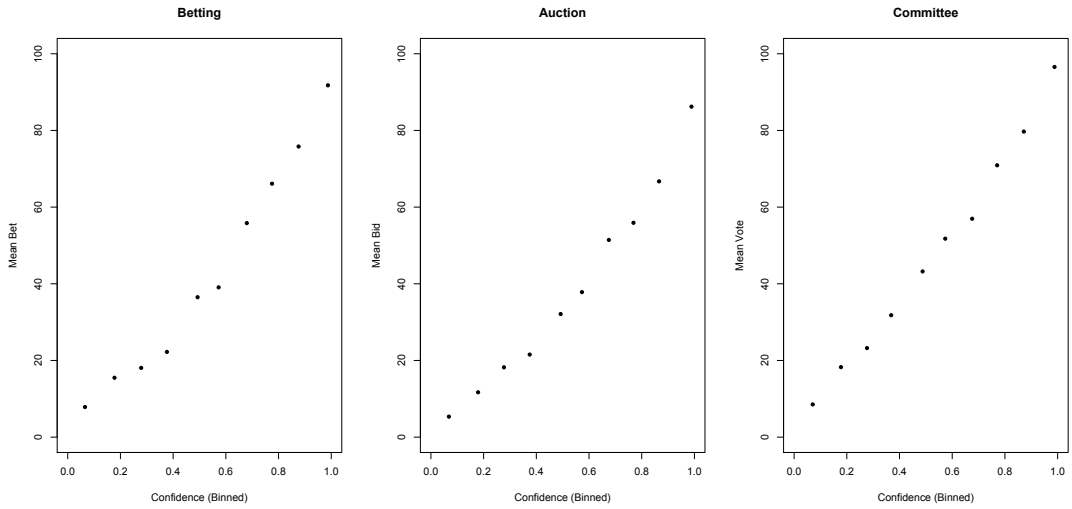


Figure 27: Binned scatterplots of institutional decisions against stated confidence in the within-subjects treatments, separately for each institution. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

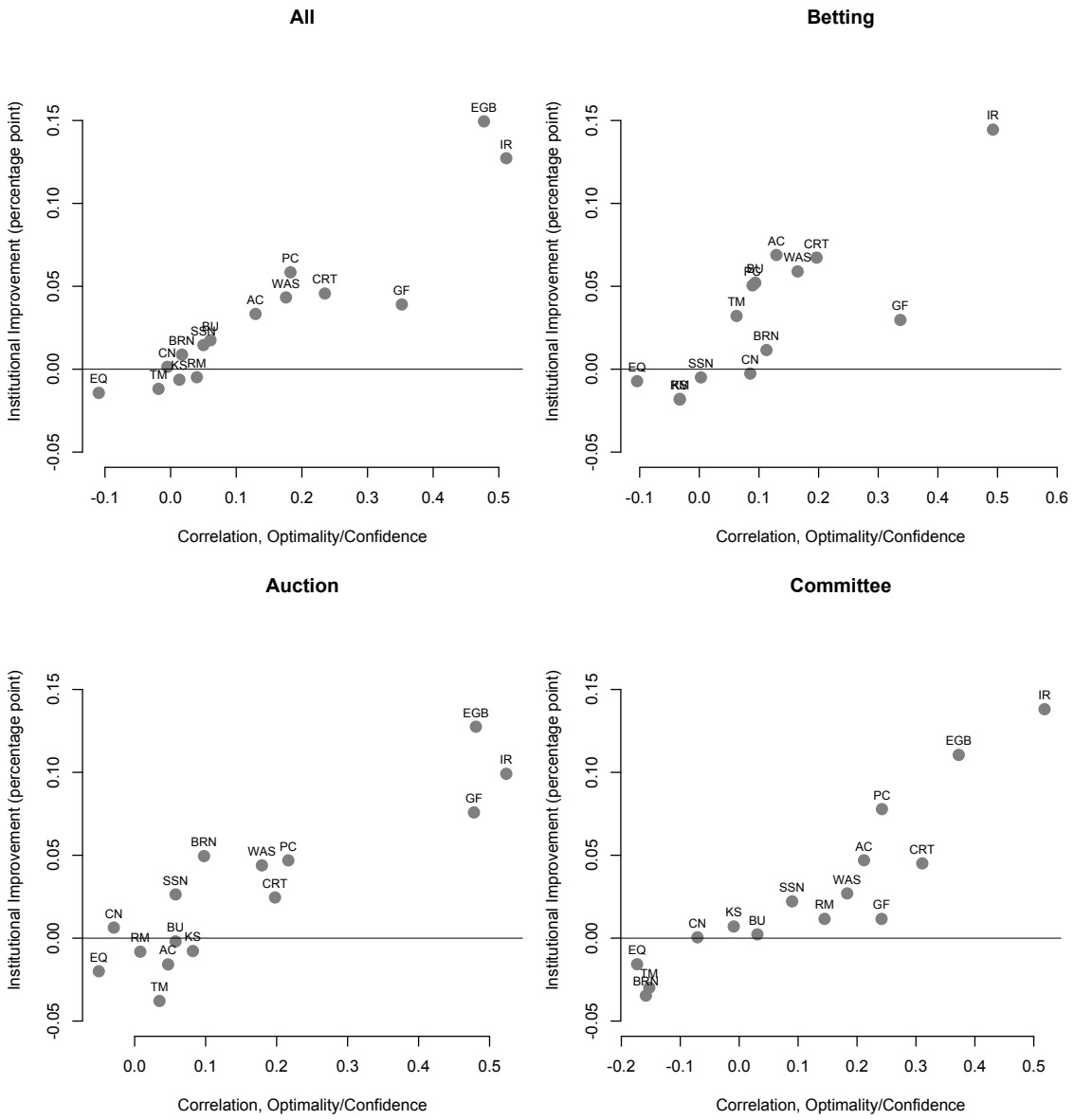


Figure 28: Confidence-optimality correlation and institutional improvement in the within-subjects treatments. The horizontal axis shows the within-task correlation between confidence and optimality in a given task. The vertical axis shows the performance improvement that is implied by an institution for the respective cognitive task. The data are from treatments *Betting Within*, *Auction Within* and *Committee Within*. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Marginal thinking; WAS=Wason task.

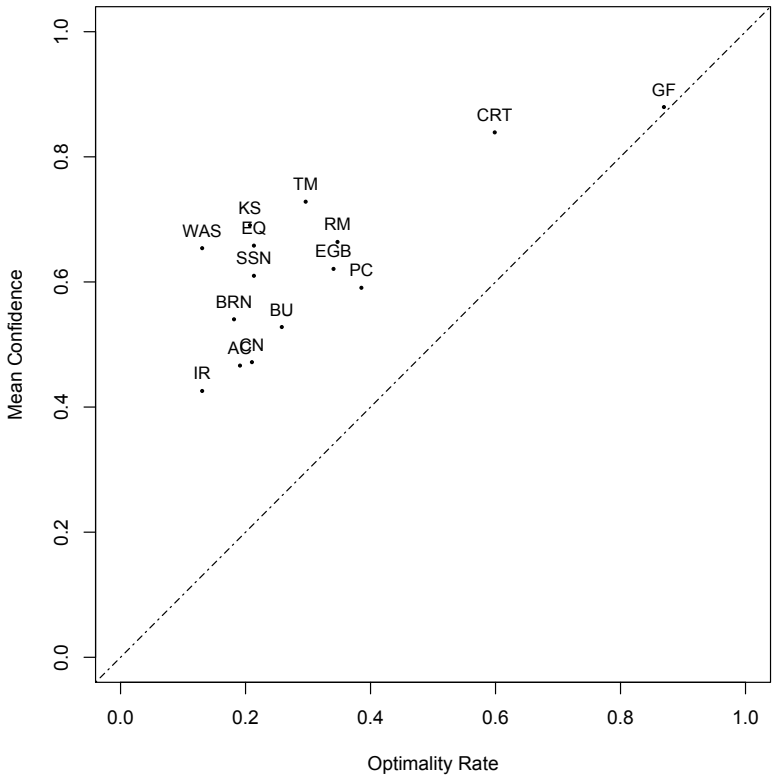


Figure 29: Fraction of optimal Part 1 responses and average confidence in treatments *Betting Within*, *Auction Within* and *Committee Within*, separately for each task. Task codes: AC=Acquiring-a-company; BRN=Base rate neglect; BU=Balls-and-urns belief updating; CN=Correlation neglect; CRT=Cognitive reflection test; EGB=Exponential growth calc.; EQ=Predict others' play; GF: Predict sequence of draws; IR=Backwards induction; KS=Knapsack; PC=Portfolio choice; RM=Attribution; SSN=Account for sample size; TM=Minimize taxation; WAS=Wason task.

D Additional Tables

Table 3: Determinants and correlates of overconfidence

	Overconfidence
Optimality Rate	−84.933***
	(7.111)
Age	0.025
	(0.065)
Male	8.928***
	(1.837)
College?	4.036**
	(1.923)
Income	0.00001
	(0.00002)
Black	3.831
	(4.147)
Hispanic	−14.553***
	(5.197)
Other Race	−1.489
	(5.860)
White	−1.099
	(3.285)
Constant	55.441***
	(4.315)
Observations	334
R^2	0.362
Adjusted R^2	0.344
Residual Std. Error	16.076 (df = 324)
F Statistic	20.413*** (df = 9; 324)

Notes. OLS estimates of overconfidence on demographic variables. The unit of the obervation is an individual subject and the independent variable is the difference between the subject's average confidence and optimality rate. Dependent variables include Age in years, an indicator for being male, an indicator for having graduated from college, income and four indicator variables for race (Asian is the excluded variable). . * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

E Experimental Instructions

E.1 Treatment *Confidence*

Instruction screens

Instructions (1/2)

Please read these instructions carefully. There will be comprehension checks. If you fail those, you will not be able to participate in the study and earn a bonus.

This study consists of a total of 15 tasks. Each of these tasks consists of two parts:

Part 1: You will make a decision by answering a question. Your decision potentially determines your bonus payment. In each question, there is going to be an optimal decision, by which we mean a decision that maximizes your earnings, on average.

Part 2: We will ask you how certain you are that your decision in Part 1 was optimal. Your response to this question does not affect your bonus.

Your bonus

As noted above, there are a total of 15 tasks in this study. In each task, you will make one earnings-relevant decision (Part 1). At the end of the study, the computer will randomly select one of your 15 decisions to determine your bonus. Because we only pay you based on a single decision of yours, there is no point for you in strategizing across decisions or tasks. You should simply always take the decision that you think is best.

→

Instructions (2/2)

Part 1: Your decision

As we described on the previous screen, in each task you will first make a decision (Part 1).

For example, in Part 1 we might ask you a question like "How many cities with more than 2 million people are there in the United States? You will receive 100 points if your decision is correct." Your Part 1 decision is simply your answer to this question, and your decision is "optimal" if it is correct.

Part 2: Your certainty

In Part 2, we ask you "How certain are you that your decision was optimal?" When we ask this question, we are interested in your assessment of how likely it is (in %) that your decision was optimal. You use a slider like the one below to give your answer. If you are completely sure your answer was correct, you should set the slider all the way to the right (100%). If you are certain your answer was not correct, you should set it all the way to the left (0%). In general, the more likely you think it is that you answered the Part 1 question correctly, the further to the right you should set your Part 2 slider.

You need to click on the slider to see the handle.

EXAMPLE:

Example of a task

Here is an example of how a task proceeds. Once the study begins, you will see these two parts on consecutive screens. We just summarize them on one screen here to give you an overview of how things work.

Part 1: Your decision

How many cities with more than 2 million people are there in the United States? You will receive 100 points if your decision is correct.

Part 2: Your certainty

Your decision is considered “optimal” if it is correct.

How certain are you that your decision in Part 1 was optimal?

Not at all certain | Fully certain |

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

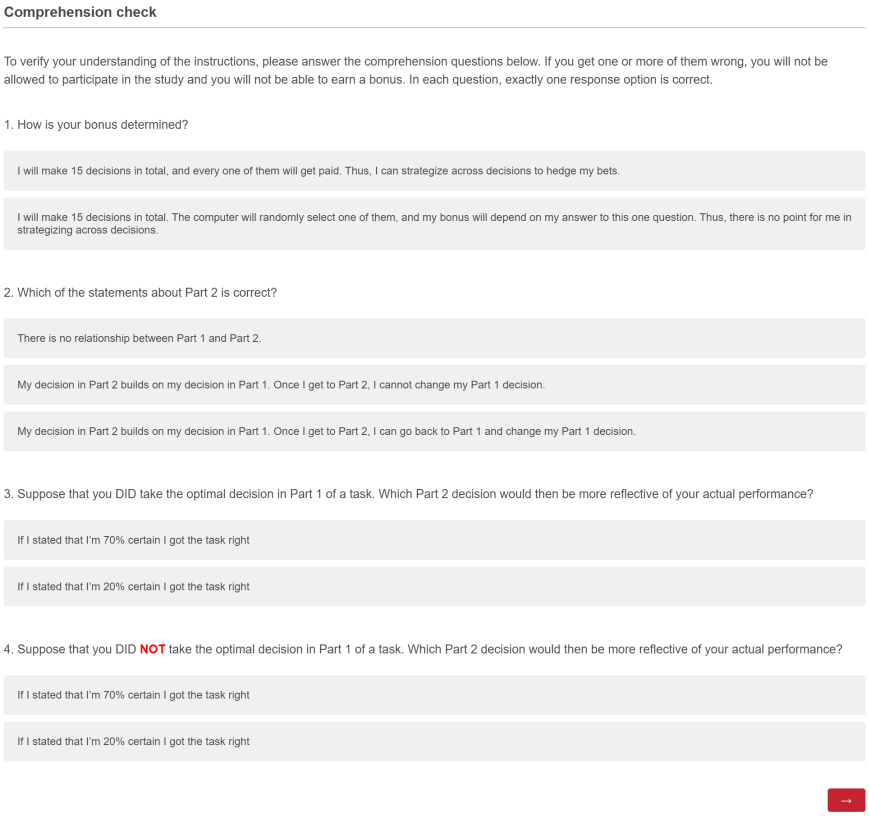
I am **PLEASE CLICK SLIDER** certain that my decision in Part 1 was optimal.

Once you click the next button, you will not be able to go back to the instructions and the comprehension check questions will start.

←

→

Comprehension questions



Comprehension check

To verify your understanding of the instructions, please answer the comprehension questions below. If you get one or more of them wrong, you will not be allowed to participate in the study and you will not be able to earn a bonus. In each question, exactly one response option is correct.

1. How is your bonus determined?

I will make 15 decisions in total, and every one of them will get paid. Thus, I can strategize across decisions to hedge my bets.

I will make 15 decisions in total. The computer will randomly select one of them, and my bonus will depend on my answer to this one question. Thus, there is no point for me in strategizing across decisions.

2. Which of the statements about Part 2 is correct?

There is no relationship between Part 1 and Part 2.

My decision in Part 2 builds on my decision in Part 1. Once I get to Part 2, I cannot change my Part 1 decision.

My decision in Part 2 builds on my decision in Part 1. Once I get to Part 2, I can go back to Part 1 and change my Part 1 decision.

3. Suppose that you DID take the optimal decision in Part 1 of a task. Which Part 2 decision would then be more reflective of your actual performance?

If I stated that I'm 70% certain I got the task right

If I stated that I'm 20% certain I got the task right

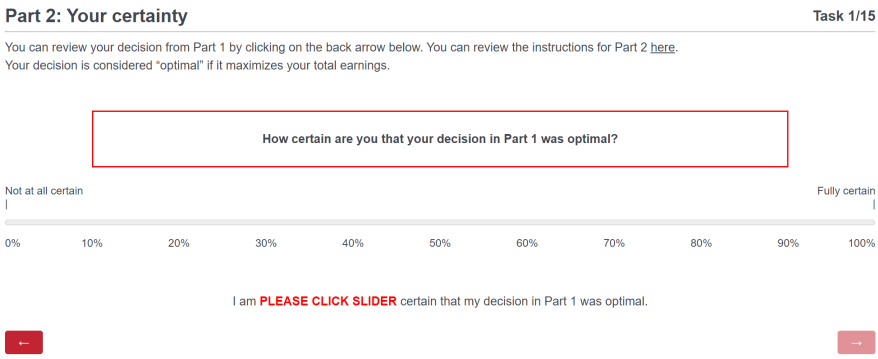
4. Suppose that you DID **NOT** take the optimal decision in Part 1 of a task. Which Part 2 decision would then be more reflective of your actual performance?

If I stated that I'm 70% certain I got the task right

If I stated that I'm 20% certain I got the task right

→

Example screen



Part 2: Your certainty

Task 1/15

You can review your decision from Part 1 by clicking on the back arrow below. You can review the instructions for Part 2 here.
Your decision is considered "optimal" if it maximizes your total earnings.

How certain are you that your decision in Part 1 was optimal?

Not at all certain

Fully certain

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

I am **PLEASE CLICK SLIDER** certain that my decision in Part 1 was optimal.

←

→

E.2 Treatment *Betting*

Instructions (1/2)

Please read these instructions carefully. There will be comprehension checks. If you fail those, you will not be able to participate in the study and earn a bonus.

This study consists of a total of 15 tasks. Each of these tasks consists of two parts:

Part 1: You will make a decision by answering a question. Your decision potentially determines your bonus payment. In each question, there is going to be an optimal decision, by which we mean a decision that maximizes your earnings, on average.

Part 2: You will make another decision that relates to the decision you made in Part 1. This decision will also potentially determine your bonus.

Your bonus

As noted above, there are a total of 15 tasks in this study. In each task, you will make two earnings-relevant decisions (Part 1 and Part 2), for a total of 30 decisions. At the end of the study, the computer will randomly select one of your 30 decisions to determine your bonus. Because we only pay you based on a single decision of yours, there is no point for you in strategizing across decisions or tasks. You should simply always take the decision that you think is best.

→

Instructions (2/2)

Part 1: Your decision

As we described on the previous screen, in each task you will first make a decision (Part 1).

For example, in Part 1 we might ask you a question like "How many cities with more than 2 million people are there in the United States? You will receive 100 points if your decision is correct." Your Part 1 decision is simply your answer to this question, and your decision is "optimal" if it is correct.

Part 2: Bet based on your decision

Once you get to Part 2, you cannot change your Part 1 decision. However, your decision in Part 2 builds on your decision in Part 1. This will work as follows:

- All participants in this study will make a decision in Part 1 by answering the exact same question as you did. In Part 2, 10 participants (including yourself) take part in a betting market that relates to their decision from Part 1.
- Each participant is given a budget of 100 points to participate in the betting market. You can use the slider below to decide **how many points to bet that your decision in Part 1 was optimal. Every point that you don't bet you get to keep**.
- Each of the other 9 participants will also decide how many of their 100 points to bet on the optimality of their decision.
- Based on how much everyone bets, and whether their decisions from Part 1 were actually optimal or not, we determine your bonus as follows:
  - **If your decision in Part 1 was not optimal, every point you bet will be lost.**
  - **If your decision in Part 1 was optimal, every point you bet will yield a positive profit for you**. In this case, your bonus is given by:

    Bonus = Number of points you bet * (Number of points bet by all participants) / (Number of points bet by participants whose Part 1 decision was optimal)

- **While this may sound complicated, what it means is relatively simple: if your decision in Part 1 was optimal, you're guaranteed to earn back at least what you bet, and probably more**.

You need to click on the slider to see the handle.

EXAMPLE:

I want to bet **PLEASE CLICK SLIDER** point(s) that my own decision in Part 1 was optimal.

← →

Example of a task

Here is an example of how a task proceeds. Once the study begins, you will see these two parts on consecutive screens. We just summarize them on one screen here to give you an overview of how things work.

Part 1: Your decision

How many cities with more than 2 million people are there in the United States? You will receive 100 points if your decision is correct.

Part 2: Bet based on your decision

How many points do you want to bet that your decision in Part 1 was optimal?

Bet nothing | Bet everything |

0 10 20 30 40 50 60 70 80 90 100

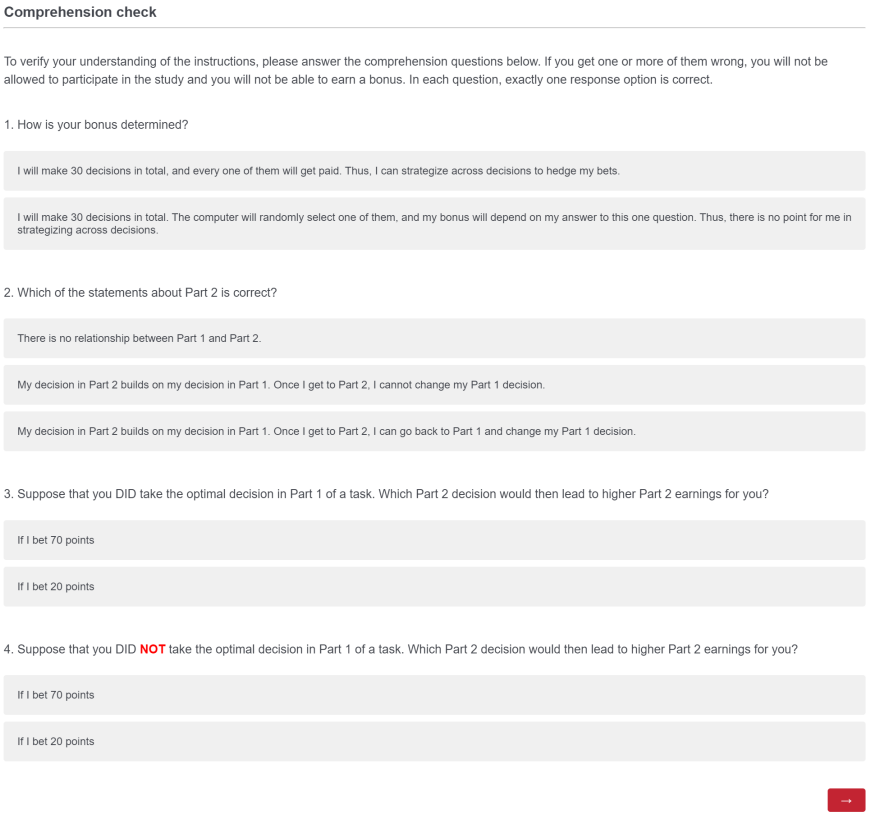
I want to bet **PLEASE CLICK SLIDER** point(s) that my own Part 1 decision was optimal.

Once you click the next button, you will not be able to go back to the instructions and the comprehension check questions will start.

←

→

Comprehension questions



Comprehension check

To verify your understanding of the instructions, please answer the comprehension questions below. If you get one or more of them wrong, you will not be allowed to participate in the study and you will not be able to earn a bonus. In each question, exactly one response option is correct.

1. How is your bonus determined?

I will make 30 decisions in total, and every one of them will get paid. Thus, I can strategize across decisions to hedge my bets.

I will make 30 decisions in total. The computer will randomly select one of them, and my bonus will depend on my answer to this one question. Thus, there is no point for me in strategizing across decisions.

2. Which of the statements about Part 2 is correct?

There is no relationship between Part 1 and Part 2.

My decision in Part 2 builds on my decision in Part 1. Once I get to Part 2, I cannot change my Part 1 decision.

My decision in Part 2 builds on my decision in Part 1. Once I get to Part 2, I can go back to Part 1 and change my Part 1 decision.

3. Suppose that you DID take the optimal decision in Part 1 of a task. Which Part 2 decision would then lead to higher Part 2 earnings for you?

If I bet 70 points

If I bet 20 points

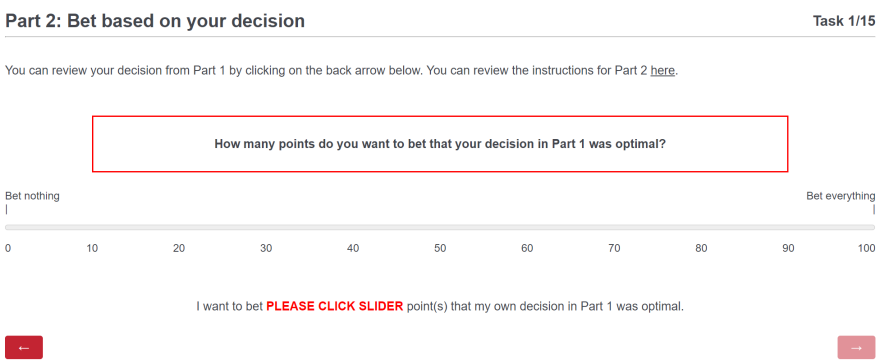
4. Suppose that you DID **NOT** take the optimal decision in Part 1 of a task. Which Part 2 decision would then lead to higher Part 2 earnings for you?

If I bet 70 points

If I bet 20 points

→

Example screen



Part 2: Bet based on your decision — Task 1/15

You can review your decision from Part 1 by clicking on the back arrow below. You can review the instructions for Part 2 here.

How many points do you want to bet that your decision in Part 1 was optimal?

Bet nothing — Bet everything

0 10 20 30 40 50 60 70 80 90 100

I want to bet **PLEASE CLICK SLIDER** point(s) that my own decision in Part 1 was optimal.

←

→

E.3 Treatment *Auction*

Instructions (1/2)

Please read these instructions carefully. There will be comprehension checks. If you fail those, you will not be able to participate in the study and earn a bonus.

This study consists of a total of 15 tasks. Each of these tasks consists of two parts:

Part 1: You will make a decision by answering a question. Your decision potentially determines your bonus payment. In each question, there is going to be an optimal decision, by which we mean a decision that maximizes your earnings, on average.

Part 2: You will make another decision that relates to the decision you made in Part 1. This decision will also potentially determine your bonus.

Your bonus

As noted above, there are a total of 15 tasks in this study. In each task, you will make two earnings-relevant decisions (Part 1 and Part 2), for a total of 30 decisions. At the end of the study, the computer will randomly select one of your 30 decisions to determine your bonus. Because we only pay you based on a single decision of yours, there is no point for you in strategizing across decisions or tasks. You should simply always take the decision that you think is best.

→

Instructions (2/2)

Part 1: Your decision

As we described on the previous screen, in each task you will first make a decision (Part 1).

For example, in Part 1 we might ask you a question like "How many cities with more than 2 million people are there in the United States? You will receive 100 points if your decision is correct." Your Part 1 decision is simply your answer to this question, and your decision is "optimal" if it is correct.

Part 2: Bid for a potential bonus based on your decision

Once you get to Part 2, you cannot change your Part 1 decision. However, your decision in Part 2 builds on your decision in Part 1. This will work as follows:

- All participants in this study will make a decision in Part 1 by answering the exact same question as you did. In Part 2, 10 participants (including yourself) will take part in an **auction** that relates to your decisions in Part 1.
- The winners of this auction will be eligible for an additional bonus. However, as explained below, the winners of the auction only receives a bonus from winning the auction if they **also** made the optimal decision in Part 1.
- You and the other participants are each given a budget of 100 points to participate in the auction. You can use the slider below to decide **how many points (maximum 100) to bid. Every point that you don't bid you get to keep no matter what**.
- The other participants will also decide how many of their 100 points to bid to get the potential bonus.
- Your Part 2 bonus is then determined according to the outcome of the auction:
  - The five highest bids win the auction. That is, the auction will have FIVE winners. If multiple participants make exactly the same fifth-highest bid, the winner will be chosen randomly from among those bidders.
  - If you are NOT among the five highest bidders, you simply keep your entire initial budget of 100 points, and you won't have to pay the bid you make.
  - If you ARE among the five highest bidders, you will have to pay the amount you bid out of your 100-point budget. In addition, you will receive a bonus:
    - If your Part 1 decision was optimal, your bonus from being a winner in the auction is 100 points.
    - If your Part 1 decision was not optimal, your bonus from being a winner in the auction is 0 points.
- **While this may sound complicated, what it means is relatively simple: if your decision in Part 1 was optimal, you receive a bonus of 100 points from being one of the winners in the auction in Part 2, but you will also have to pay your winning bid. If, on the other hand, your decision in Part 1 was not optimal, you don't receive a bonus from being one of the winners in the auction, but you will still have to pay your winning bid.**

You need to click on the slider to see the handle.

EXAMPLE:

I want to bid **PLEASE CLICK SLIDER** point(s) to get a bonus that depends on my Part 1 decision.

←

→

Example of a task

Here is an example of how a task proceeds. Once the study begins, you will see these two parts on consecutive screens. We just summarize them on one screen here to give you an overview of how things work.

Part 1: Your decision

How many cities with more than 2 million people are there in the United States? You will receive 100 points if your decision is correct.

Part 2: Bid for a potential bonus based on your decision

How many points do you want to bid for receiving the bonus that depends on your Part 1 decision?

Bid nothing | Bid everything |

0 10 20 30 40 50 60 70 80 90 100

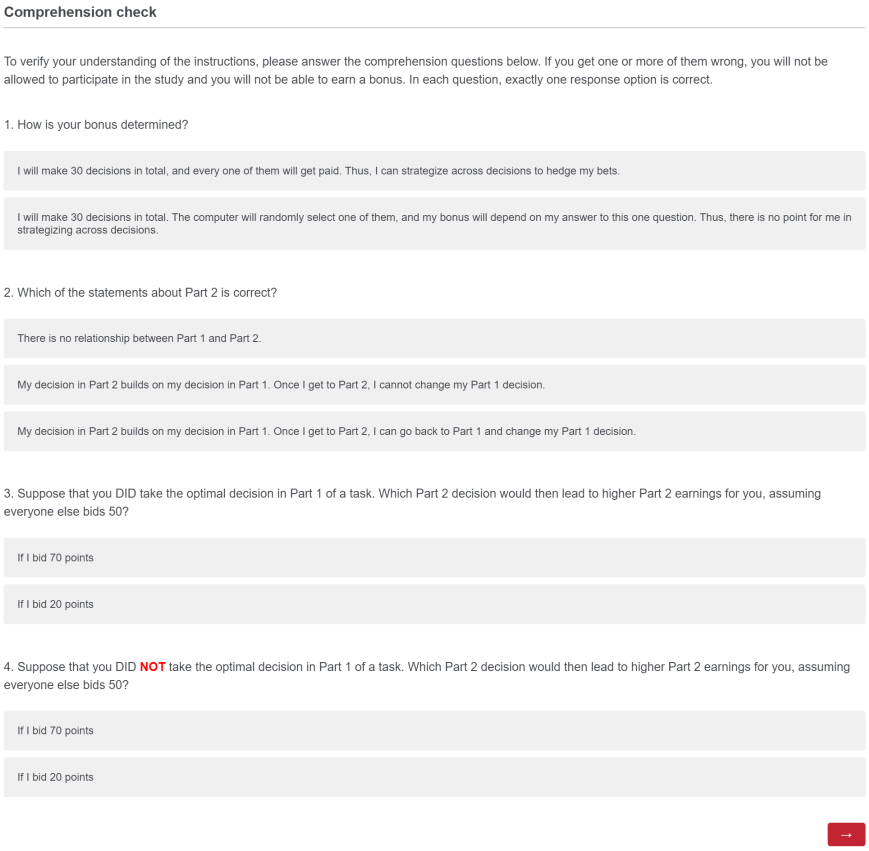
I want to bid **PLEASE CLICK SLIDER** point(s) to get a bonus that depends on my Part 1 decision.

Once you click the next button, you will not be able to go back to the instructions and the comprehension check questions will start.

←

→

Comprehension questions



Comprehension check

To verify your understanding of the instructions, please answer the comprehension questions below. If you get one or more of them wrong, you will not be allowed to participate in the study and you will not be able to earn a bonus. In each question, exactly one response option is correct.

1. How is your bonus determined?

I will make 30 decisions in total, and every one of them will get paid. Thus, I can strategize across decisions to hedge my bets.

I will make 30 decisions in total. The computer will randomly select one of them, and my bonus will depend on my answer to this one question. Thus, there is no point for me in strategizing across decisions.

2. Which of the statements about Part 2 is correct?

There is no relationship between Part 1 and Part 2.

My decision in Part 2 builds on my decision in Part 1. Once I get to Part 2, I cannot change my Part 1 decision.

My decision in Part 2 builds on my decision in Part 1. Once I get to Part 2, I can go back to Part 1 and change my Part 1 decision.

3. Suppose that you DID take the optimal decision in Part 1 of a task. Which Part 2 decision would then lead to higher Part 2 earnings for you, assuming everyone else bids 50?

If I bid 70 points

If I bid 20 points

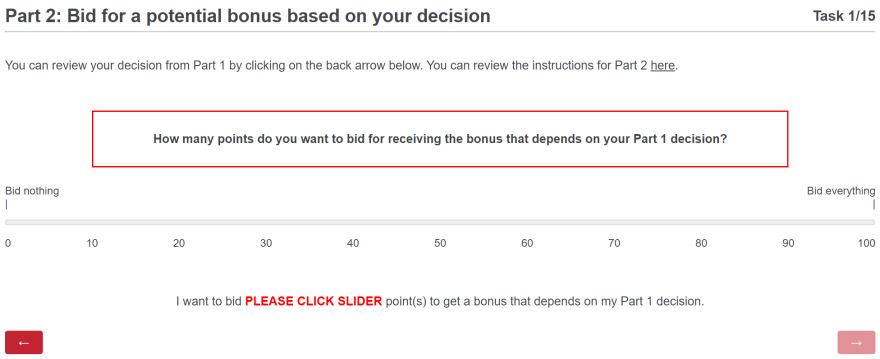
4. Suppose that you DID **NOT** take the optimal decision in Part 1 of a task. Which Part 2 decision would then lead to higher Part 2 earnings for you, assuming everyone else bids 50?

If I bid 70 points

If I bid 20 points

→

Example screen



Part 2: Bid for a potential bonus based on your decision Task 1/15

You can review your decision from Part 1 by clicking on the back arrow below. You can review the instructions for Part 2 here.

How many points do you want to bid for receiving the bonus that depends on your Part 1 decision?

Bid nothing Bid everything

0 10 20 30 40 50 60 70 80 90 100

I want to bid **PLEASE CLICK SLIDER** point(s) to get a bonus that depends on my Part 1 decision.

← →

E.4 Treatment *Committee*

Instructions (1/2)

Please read these instructions carefully. There will be comprehension checks. If you fail those, you will not be able to participate in the study and earn a bonus.

This study consists of a total of 15 tasks. Each of these tasks consists of two parts:

Part 1: You will make a decision by answering a question. Your decision potentially determines your bonus payment. In each question, there is going to be an optimal decision, by which we mean a decision that maximizes your earnings, on average.

Part 2: You will make another decision that relates to the decision you made in Part 1. This decision will also potentially determine your bonus.

Your bonus

As noted above, there are a total of 15 tasks in this study. In each task, you will make two earnings-relevant decisions (Part 1 and Part 2), for a total of 30 decisions. At the end of the study, the computer will randomly select one of your 30 decisions to determine your bonus. Because we only pay you based on a single decision of yours, there is no point for you in strategizing across decisions or tasks. You should simply always take the decision that you think is best.

→

Instructions (2/2)

Part 1: Your decision

As we described on the previous screen, in each task you will first make a decision (Part 1).

For example, in Part 1 we might ask you a question like "How many cities with more than 2 million people are there in the United States? You will receive 100 points if your decision is correct." Your Part 1 decision is simply your answer to this question, and your decision is "optimal" if it is correct.

Part 2: Vote based on your decision

Once you get to Part 2, you cannot change your Part 1 decision. However, your decision in Part 2 builds on your decision in Part 1. This will work as follows:

- All participants in this study will make a decision in Part 1 by answering the exact same question as you did. In Part 2, we will combine the decisions of 10 participants (including yourself) through a voting procedure to determine the group's Part 2 earnings.
- You can submit **up to 100 votes for your own decision from Part 1**. The more votes you choose to submit, the more your decision will influence the group's earnings.
  - For instance, if you choose to submit 0 votes, you are choosing that your decision from Part 1 does not have any influence on the group's earnings. If you choose to submit 100 votes, you are choosing to have as much influence as possible. You can choose how many votes you'd like to submit using a slider like the one below
- Each of the other 9 participants will also decide how many of their 100 votes to submit for their own decision from Part 1. **The group's earnings are higher the more total votes (across all participants in your group) get submitted for the optimal decision, and lower the more total votes get submitted for decisions that are not optimal.**
- **Important: You and all other participants will all earn an identical amount if Part 2 is selected to count for payment, regardless of how many votes you each individually choose to submit**. Everyone's bonus from Part 2 is only determined by the **fraction of total submitted votes** that are for the optimal decision. Therefore, from the perspective of your earnings, it doesn't matter whether you or other people submit votes – **all that matters is that the votes that do get submitted (whoever they are from) are for the optimal decision**. In case you're interested, the specific formula we use to determine everyone's bonus is given by

  $$\text{Bonus} = 100 * (\text{Number of votes for the optimal decision}) / (\text{Total number of votes})$$

You need to click on the slider to see the handle.

EXAMPLE:

I want to submit **PLEASE CLICK SLIDER** vote(s) for my own Part 1 decision.

← →

Example of a task

Here is an example of how a task proceeds. Once the study begins, you will see these two parts on consecutive screens. We just summarize them on one screen here to give you an overview of how things work.

Part 1: Your decision

How many cities with more than 2 million people are there in the United States? You will receive 100 points if your decision is correct.

Part 2: Vote based on your decision

How many votes do you want to submit for your own decision from Part 1?

No influence on group earnings | Maximal influence on group earnings |

0 10 20 30 40 50 60 70 80 90 100

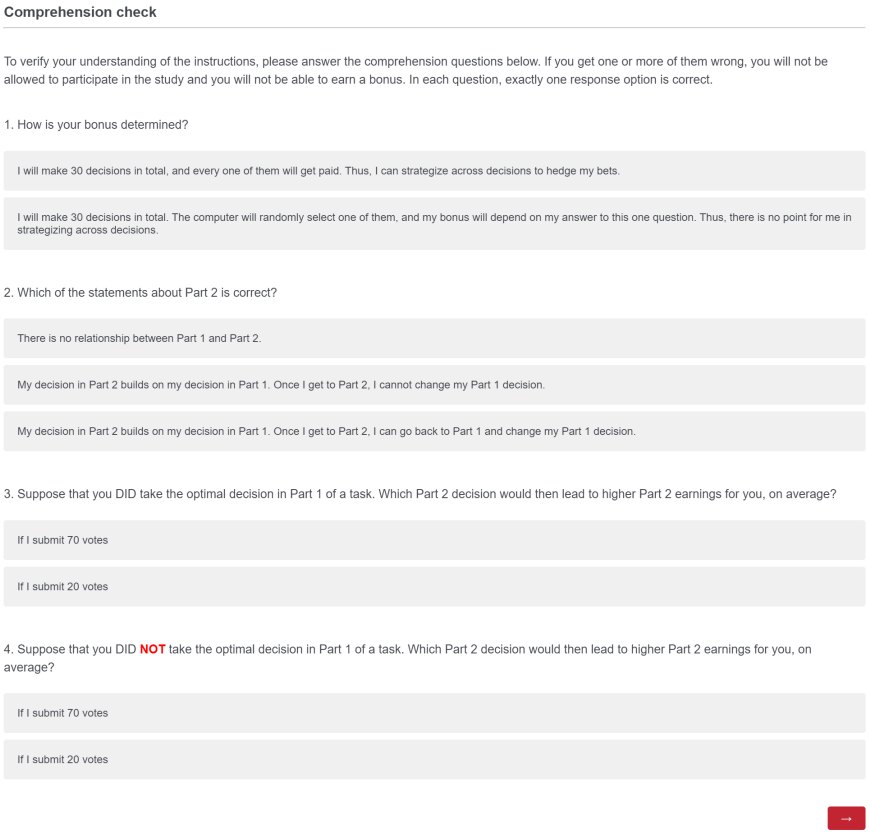
I want to submit **PLEASE CLICK SLIDER** vote(s) for my own Part 1 decision.

Once you click the next button, you will not be able to go back to the instructions and the comprehension check questions will start.

←

→

Comprehension questions



Comprehension check

To verify your understanding of the instructions, please answer the comprehension questions below. If you get one or more of them wrong, you will not be allowed to participate in the study and you will not be able to earn a bonus. In each question, exactly one response option is correct.

1. How is your bonus determined?

I will make 30 decisions in total, and every one of them will get paid. Thus, I can strategize across decisions to hedge my bets.

I will make 30 decisions in total. The computer will randomly select one of them, and my bonus will depend on my answer to this one question. Thus, there is no point for me in strategizing across decisions.

2. Which of the statements about Part 2 is correct?

There is no relationship between Part 1 and Part 2.

My decision in Part 2 builds on my decision in Part 1. Once I get to Part 2, I cannot change my Part 1 decision.

My decision in Part 2 builds on my decision in Part 1. Once I get to Part 2, I can go back to Part 1 and change my Part 1 decision.

3. Suppose that you DID take the optimal decision in Part 1 of a task. Which Part 2 decision would then lead to higher Part 2 earnings for you, on average?

If I submit 70 votes

If I submit 20 votes

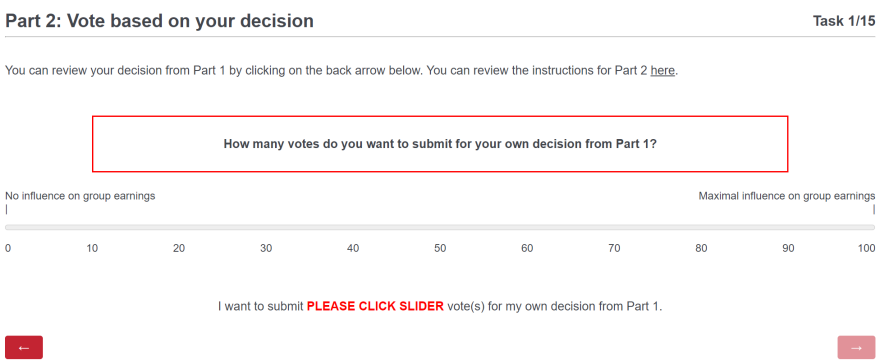
4. Suppose that you DID **NOT** take the optimal decision in Part 1 of a task. Which Part 2 decision would then lead to higher Part 2 earnings for you, on average?

If I submit 70 votes

If I submit 20 votes

→

Example screen



Part 2: Vote based on your decision

Task 1/15

You can review your decision from Part 1 by clicking on the back arrow below. You can review the instructions for Part 2 here.

How many votes do you want to submit for your own decision from Part 1?

No influence on group earnings

Maximal influence on group earnings

0 10 20 30 40 50 60 70 80 90 100

I want to submit **PLEASE CLICK SLIDER** vote(s) for my own decision from Part 1.

←

→

E.5 Task Descriptions

Acquiring a compnay

Part 1: Your decision — **Task 1/15**

- You have a budget of 180 points. You can either keep it or use it to buy a company.
- Bob is selling his company. The **VALUE** of Bob's company to him is either **20** or **120 points**, but you do not know which. There is a **50% chance** it is worth 20 points to him and a **50% chance** it is worth 120 points to him.
- **Bob's company has a higher value to you than to Bob**. If you acquire his company, it will pay you 1.5 times its value to Bob. Therefore, if the value of the company turns out to be 20 points for Bob, it would be worth 30 points for you. If the value of the company turns out to be 120 points for Bob, it would be worth 180 points for you.
- The realized value is determined randomly by the computer, and you will not know the value until after you've made your decision.
- You can make a **PRICE** offer to Bob of up to 180 points.
- Your earnings will be determined as follows:
  - If you offer a PRICE **that is at least as high** as Bob's realized VALUE, Bob will accept your offer, and your earnings will be
    *Earnings = (Your budget) + 1.5 * (Bob's **VALUE**) – (the **PRICE** you offered)*
  - If you offer a PRICE **less** than Bob's realized VALUE, you will not acquire his company and your profits will be
    Earnings = Your budget

How much do you bid for Bob's company?

point(s)

→

Knapsack

Part 1: Your decision — **Task 1/15**

- There are 12 **ITEMS** shown in the Table below. Your task is to choose one or more of these items.
- Each item has a **VALUE** in points to you. Your earnings for this task are given by the **SUM OF VALUES** of the items you choose.
- However, each item also has a **WEIGHT**. The total **SUM OF WEIGHTS** of the items you choose **CANNOT EXCEED 14**. If your selection exceeds this weight limit, you will earn nothing.

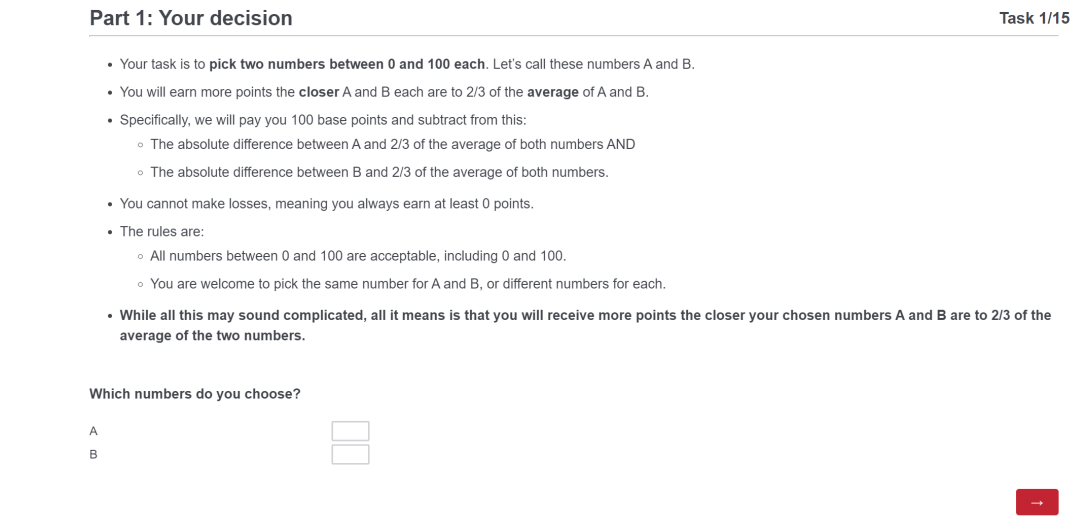
Which items do you choose? (Please click on the columns)

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12
Value	2	3	4	5	6	9	8	7	6	5	8	9
Weight	3	4	6	3	5	13	6	9	2	4	7	7

Current sum of weights chosen **(cannot exceed 14)**: **0**

→

Iterated reasoning / backward induction



Part 1: Your decision — Task 1/15

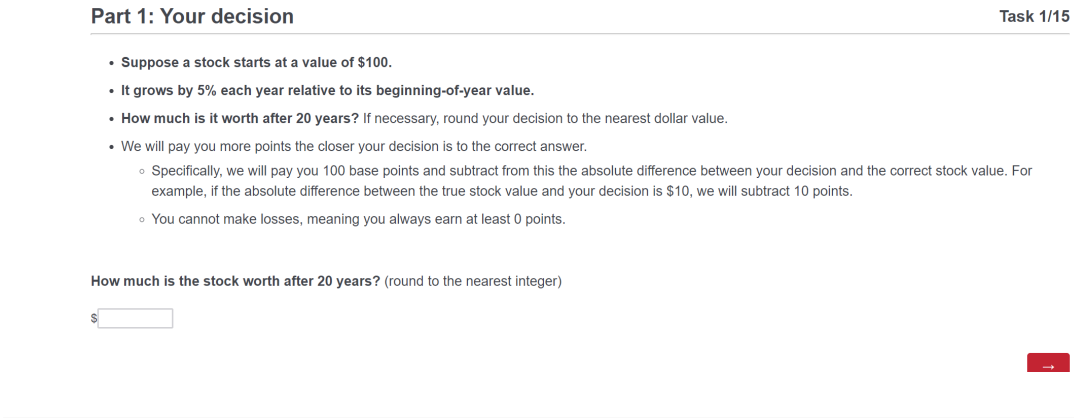
- Your task is to **pick two numbers between 0 and 100 each**. Let's call these numbers A and B.
- You will earn more points the **closer** A and B each are to 2/3 of the **average** of A and B.
- Specifically, we will pay you 100 base points and subtract from this:
  - The absolute difference between A and 2/3 of the average of both numbers AND
  - The absolute difference between B and 2/3 of the average of both numbers.
- You cannot make losses, meaning you always earn at least 0 points.
- The rules are:
  - All numbers between 0 and 100 are acceptable, including 0 and 100.
  - You are welcome to pick the same number for A and B, or different numbers for each.
- **While all this may sound complicated, all it means is that you will receive more points the closer your chosen numbers A and B are to 2/3 of the average of the two numbers.**

Which numbers do you choose?

A
B

→

Exponential growth bias



Part 1: Your decision — Task 1/15

- **Suppose a stock starts at a value of $100.**
- **It grows by 5% each year relative to its beginning-of-year value.**
- **How much is it worth after 20 years?** If necessary, round your decision to the nearest dollar value.
- We will pay you more points the closer your decision is to the correct answer.
  - Specifically, we will pay you 100 base points and subtract from this the absolute difference between your decision and the correct stock value. For example, if the absolute difference between the true stock value and your decision is $10, we will subtract 10 points.
  - You cannot make losses, meaning you always earn at least 0 points.

How much is the stock worth after 20 years? (round to the nearest integer)

$

→

Correlation neglect

Part 1: Your decision

Task 1/15

- There are three people: Ann, Bob and Charlie. Each of them is interested in estimating the weight of a water bucket in pounds.
- Ann and Bob both **get to take a peek at the bucket**. They are equally good at estimating weight. Each of them gets weight estimates right, on average, but sometimes makes **random mistakes**. Ann and Bob are equally likely to make mistakes in any given estimate they make.
- Ann and Bob both share their estimates with Charlie, who has never seen the bucket. Because he has never seen the bucket, Charlie computes his best estimate of the weight of the bucket as the average of the estimates of Ann and Bob.
- **You have never seen the bucket either, but you're asked to produce an estimate of its weight.** You now talk to Ann and Charlie. They share the following estimates with you:
  - Ann's estimate: 70
  - Charlie's estimate: 40
- **Your task is to estimate the weight of the bucket.**
- We will pay you more points the closer your decision is to the statistically-correct estimate given the information you are provided.
  - Specifically, we will pay you 100 points if your decision corresponds to this correct answer. We subtract 3 points for every number you are away from the correct answer.
  - You cannot make losses, meaning you always earn at least 0 points.

What is your best estimate of the weight of the bucket? (round to the nearest integer)

→

CRT

Part 1: Your decision

Task 1/15

It takes 6 machines 6 days to produce 6 cars. **How long would it take 12 machines to produce 12 cars?** (round to the nearest integer)

day(s)

We will pay you 100 points if you get it right, and nothing otherwise.

Wason

Part 1: Your decision

Task 1/15

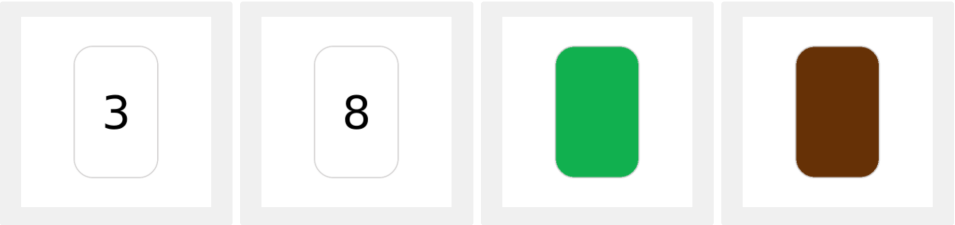
- Suppose your friend has a special deck of cards.
- His special deck of cards all have **numbers** (odd or even) on one side and **colors** (brown or green) on the other side. Suppose that the **4 cards** below are from his deck.
- Your friend claims that: "**In my deck of cards, all of the cards with an even number on one side are green on the other.**"

We will pay you 100 points if:

- you turn over **ALL** of the card(s) that can be helpful in determining whether your friend's statement is true AND
- you do not turn over **ANY** of the cards that **CANNOT** be helpful in assessing whether his statement is true.

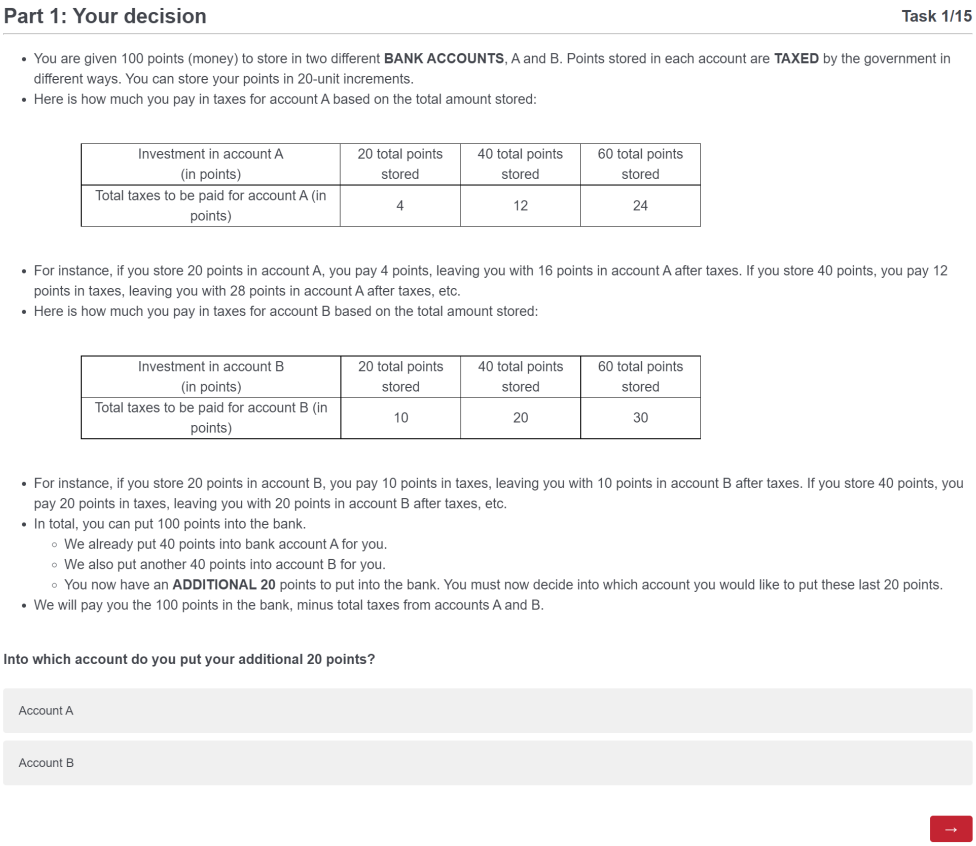
In other words, you won't earn 100 points if you turned over a card that isn't actually helpful in determining whether your friend's statement is true. Likewise, you won't earn 100 points if you failed to turn over a card that is actually helpful.

Which card(s) do you want to turn over?



→

Thinking at margin



Part 1: Your decision Task 1/15

- You are given 100 points (money) to store in two different **BANK ACCOUNTS**, A and B. Points stored in each account are **TAXED** by the government in different ways. You can store your points in 20-unit increments.
- Here is how much you pay in taxes for account A based on the total amount stored:

Investment in account A (in points)	20 total points stored	40 total points stored	60 total points stored
Total taxes to be paid for account A (in points)	4	12	24

- For instance, if you store 20 points in account A, you pay 4 points, leaving you with 16 points in account A after taxes. If you store 40 points, you pay 12 points in taxes, leaving you with 28 points in account A after taxes, etc.
- Here is how much you pay in taxes for account B based on the total amount stored:

Investment in account B (in points)	20 total points stored	40 total points stored	60 total points stored
Total taxes to be paid for account B (in points)	10	20	30

- For instance, if you store 20 points in account B, you pay 10 points in taxes, leaving you with 10 points in account B after taxes. If you store 40 points, you pay 20 points in taxes, leaving you with 20 points in account B after taxes, etc.
- In total, you can put 100 points into the bank.
  - We already put 40 points into bank account A for you.
  - We also put another 40 points into account B for you.
  - You now have an **ADDITIONAL 20** points to put into the bank. You must now decide into which account you would like to put these last 20 points.
- We will pay you the 100 points in the bank, minus total taxes from accounts A and B.

Into which account do you put your additional 20 points?

Account A

Account B

→

We noticed a typo in the instructions of this task after beginning the data collection: In the first sentence, instead of “100 points”, the instructions read “60 points”. Despite this typo, we believe that it was still possible to follow the task description and arrive at the correct decision. The optimization rate in this task did not change after fixing this typo: it was 28.28% (N=1,372) before and 28.24% (N=170) after correcting the mistake. In all our analysis, we thus pool these data.

Portfolio choice and 1/N

Part 1: Your decision

Task 1/15

- **In this task, you'll be asked to choose an investment portfolio that consists of different stocks.**
- There are **four stocks** that pay you different amounts of money depending on the color of a ball that a computer will randomly draw. Each of the colors red, blue, and green is equally likely to get selected by the computer.
- The table below shows you the **payment rate** of each stock, depending on which ball the computer randomly draws. For example, a realized return of 10% means that if you invest 20 points, you end up with 22 points. Likewise, a realized return of -10% means that if you invest 20 points, you end up with 18 points.

Color of ball drawn	Return of **Stock A**	Return of **Stock B**	Return of **Stock C**	Return of **Stock D**
Red	13%	-2%	-9%	17%
Blue	-8%	8%	12%	-9%
Green	8%	6%	7%	7%

- In total, you need to invest **100 points** across these stocks. You can select one of the portfolios (combinations of stock purchases) below.
- The computer will randomly draw a ball and pay you the total amount earned across the stocks in your portfolio.

Portfolio	Points in **Stock A**	Points in **Stock B**	Points in **Stock C**	Points in **Stock D**
I	50	25	0	25
II	25	25	25	25

Which investment portfolio do you choose?

Portfolio I

Portfolio II

→

Balls and urns

Part 1: Your decision

Task 1/15

- There are **two bags**. One bag contains 70 red chips and 30 blue chips. The other one contains 30 red chips and 70 blue chips.
- We secretly flipped a (fair) coin. If it came up **HEADS**, we chose the bag with more red chips. If the coin came up **TAILS**, we chose the bag with more blue chips. **You do not observe which bag was selected**.
- Next, **we drew one chip at random from the bag selected by the coin toss**. You will learn the color of this randomly-drawn chip below. Then, you need to guess (in percent) which bag was selected.
- We will pay you more points the closer your decision is to the statistically-correct percentage chance given the information you are provided.
  - Specifically, we will pay you 100 points if your decision corresponds to this correct answer. We subtract 3 points for every percentage point you are away from the correct answer.
  - You cannot make losses, meaning you always earn at least 0 points.

You are told that one red chip has randomly been drawn from the secretly selected bag. What do you think is the likelihood (percentage chance) that the selected bag is the one with more red chips? (round to the nearest integer)

____ %

→

Sample size neglect

Part 1: Your decision

Task 1/15

- There are **two factories** that make office chairs. The **larger** factory produces **45 chairs** each day, and the smaller factory produces **15 chairs** each day.
- For both factories, there is a **10% random chance** that any given chair is **defective**. However, since this is random, the exact percentage varies from day to day. Sometimes it may be higher than 10%, sometimes lower.
- For a period of 1 year, each factory recorded the days on which **MORE THAN 20%** of the chairs were defective.

Which factory do you think recorded more days on which more than 20% of the chairs were defective?

The larger factory

The smaller factory

About the same (that is, within 2% of each other)

We will pay you 100 points if your decision corresponds to the statistically-correct option given the information you are provided, and nothing otherwise.

→

Base rate neglect

Part 1: Your decision

Task 1/15

- Assume that, on average, out of every 100 bicycles produced by a bike manufacturer, **90 are good** and **10 are defective**.
- There is a quality control machine that classifies bicycles as either good or defective at the end of the production process. This quality control machine makes classification mistakes from time to time. On average, the machine **correctly** classifies a bicycle (as good or defective) **75 out of 100 times**, but **incorrectly** classifies it **25 out of 100 times**.
- Now a bicycle produced by the manufacturer has randomly been selected. Next, this specific bicycle was run though the quality control machine, and you will learn about the machine's classification below. Based on this classification, your task is to state the likelihood (percentage chance) that this specific bicycle is actually defective.
- We will pay you more points the closer your decision is to the statistically-correct percentage chance given the information you are provided.
  - Specifically, we will pay you 100 points if your decision corresponds to this correct answer. We subtract 3 points for every percentage point you are away from the correct answer.
  - You cannot make losses, meaning you always earn at least 0 points.

You learn that the randomly selected bicycle has been **classified as defective** by the quality control machine. **What do you think is the likelihood (percentage chance) that it is actually defective?** (round to the nearest integer)

%

→

Gambler's fallacy

Part 1: Your decision

Task 1/15

Imagine you are tossing a **fair coin**. After eight tosses you observe the following result (where **T** stands for **TAILS** and **H** stands for **HEADS**):

T – T – T – H – T – H – H – H

Which event is more likely to happen on the next coin toss?

- Heads is more likely
- Tails is more likely
- Both are equally likely

We will pay you 100 points if your decision corresponds to the statistically-correct option given the information you are provided, and nothing otherwise.

→

Regression to the mean / misattribution

Part 1: Your decision

Task 1/15

- The **average** score on a standard **IQ test** is **100**. Suppose a **randomly** selected individual has obtained a score of **140**.
- Suppose further that an IQ score is the sum of both **true ability** and **random good or bad luck**. The luck component can be positive or negative but equals zero on average (over all people).

Which of the following statements is correct?

- This person's true IQ is more likely to be above than below 140.
- This person's true IQ is more likely to be below than above 140.
- This person's true IQ is equally likely to be above or below 140.

We will pay you 100 points if your decision corresponds to the statistically-correct statement given the information you are provided, and nothing otherwise.

→
