American Economic Review 2023, 113(7): 1933–1966
https://doi.org/10.1257/aer.20220915

Confidence, Self-Selection, and Bias in the Aggregate[†]

By Benjamin Enke, Thomas Graeber, and Ryan Oprea*

The influence of behavioral biases on aggregate outcomes depends in part on self-selection: whether rational people opt more strongly into aggregate interactions than biased individuals. In betting market, auction and committee experiments, we document that some errors are strongly reduced through self-selection, while others are not affected at all or even amplified. A large part of this variation is explained by differences in the relationship between confidence and performance. In some tasks, they are positively correlated, such that self-selection attenuates errors. In other tasks, rational and biased people are equally confident, such that self-selection has no effects on aggregate quantities. (*JEL* C91, D44, D91)

Decades of experimental research on judgment and decision-making have revealed that people 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 aggregate 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 prior research has 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 explore the idea that the psychological forces studied in behavioral economics might also affect how strongly individual errors influence economic outcomes. Specifically, we experimentally study to what degree decision makers' own beliefs about the quality of their decisions influences how strongly biases survive in economic aggregates.

Our point of departure is the observation that, in laboratory experiments, researchers "force" subjects to make cognitively difficult decisions, while real-world interactions often afford decision makers the freedom to self-select into or out of decisions. For instance, people might shy away from betting in markets when they fear that

*Enke: Harvard University, Department of Economics, and NBER (email: enke@fas.harvard.edu); Graeber: Harvard Business School (email: tgraeber@hbs.edu); Oprea: UC Santa Barbara, Department of Economics (email: roprea@gmail.com). Stefano DellaVigna was the coeditor for this article. We are grateful for comments from very constructive referees, Nick Barberis, Colin Camerer, Ernst Fehr, Dan Friedman, John List, Don Moore, Terry Odean, and Lise Vesterlund. We are also grateful to audiences at Columbia University, the University of Chicago, Harvard University, briq, the ESA, VIBES, CESifo, BEAM, and ECBE. This research was supported by the National Science Foundation under grant SES-1949366. Human subjects approval was given by UC Santa Barbara (Protocol 11-22-0691) and Harvard (Protocol 16-1753) IRB.

[†]Go to https://doi.org/10.1257/aer.20220915 to visit the article page for additional materials and author disclosure statements.

their judgments are fallible, they might be reluctant to aggressively bid in auctions over property rights for objects that they think they do not fully understand, or they might refrain from contributing their opinion to decision-making processes in groups and organizations when they suspect that they don't understand the matter at hand. Thus, self-selection might severely filter individual-level irrationalities, relative to the unfiltered measures we observe in the lab.[1]

Our research is built on the observation that the argument above relies on the assumption that *selection is positive*, meaning that more rational people self-select more strongly into decisions that affect aggregate quantities. Yet, to date, we know relatively little about people's selection decisions across the multitude of cognitive biases studied by behavioral economists. A first-order statistic that may determine the economic effects of self-selection is the *confidence-performance correlation* of a cognitive error: the correlation between objective performance and confidence in the population. Using a simple model, we illustrate that if error-prone individuals are relatively less confident in the optimality of their decisions, then aggregate interactions (in, e.g., markets or organizations) will tend to effectively filter a bias. If, on the other hand, performance and confidence are unrelated, or even negatively correlated, then self-selection will not filter a bias, or might even amplify its effect. Our simple model clarifies that what should matter for the economic effects of self-selection is not so much average overconfidence (i.e., overestimation, overplacement, or overprecision in the taxonomy of Moore and Healy 2008) but instead the correlation between performance and confidence. For example, as long as performance and confidence are positively correlated, self-selection will attenuate errors even in the extreme scenario that every single decision maker is overconfident.

As we discuss in detail below, a large literature in economics and psychology has studied to what degree various types of overconfidence vary across different types of tasks. By contrast, there is relatively little evidence on how the correlation between confidence and performance—which is the relevant object for studying and predicting the effects of self-selection—varies across cognitive biases. This gap in our knowledge appears crucial because there is immense diversity in the types of cognitive biases documented in the literature, in terms of both their domains (information processing, financial decision-making, strategic sophistication, etc.) and their underlying psychology (complexity, misleading intuitions, inattention, etc.). Given this diversity, it is conceivable that the confidence-performance correlation varies markedly across types of errors.

We implement a series of preregistered experiments to study the nature of self-selection under social institutions, and how this relates to the distribution of confidence. The main features of our experiment are (i) a broad set of 15 classical cognitive tasks and associated biases; (ii) three different self-selective "social institutions" in which subjects interact to produce aggregate outcomes; and (iii) direct measurements of across-subject confidence-performance correlations in each cognitive task. 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 (Enke et al. 2023).

[1] Gary Becker once opined in an interview that division of labor and resulting self-selection "strongly attenuates if not eliminates" effects of bounded rationality on economic aggregates (Stewart 2005).

We consider 15 cognitive tasks, culled from the literature 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 ten-subject cohorts to participate in one of three maximally simple canonical social institutions.

In a Betting treatment, ten subjects participate in a parimutuel betting market, in which they bet on the optimality of their part 1 decisions. In an 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 a Committee treatment, we have all ten subjects decide how intensively to vote for their own part 1 decision to influence a common group decision. In each of these institutions, subjects make a single decision that determines their degree of self-selection. By betting, bidding, or voting less intensively, subjects can partly or fully select out of influencing institutional outcomes in a continuous way. We built our design around three distinct social institutions not to intensively compare them but rather to ensure that our conclusions on self-selection across biases aren't overfit to any one idiosyncratic institutional setting. Indeed, we deliberately implemented maximally simple and static versions of these institutions, in order to be able to study an unusually large number of biases—a design choice that we believe can be profitably relaxed in future work.

In each of these institutions, the difference between how intensively cognitively biased versus unbiased subjects bid, bet, or vote 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 aggregate outcomes to average rates of bias (measured in part 1), we can measure 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.[2]

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 of this positive selection, on average, biases are filtered in all three institutions, producing institutional aggregates that are less biased than subjects 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 due to negative selection.

The heterogeneity in institutional filtering across cognitive tasks is very similar across the three different institutions. Although levels of improvement vary 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.

[2] As discussed in the conclusion, we only focus on how self-selection affects the rationality/efficiency of the aggregate quantity that an institution produces, rather than on how it affects aggregate welfare.

Our motivating hypothesis (preregistered prior to the experiment) is that this variation can be partly explained by a specific summary statistic of the distribution of confidence. As derived in a simple framework, our key prediction is that institutional filtering by self-selection critically depends on the cross-subject correlation between performance and confidence. 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 question of independent interest that has received little attention in behavioral economics so far: how strongly are confidence and performance correlated for different biases commonly studied in economics?

We find strong heterogeneity in the size and sign of the confidence-performance correlation 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).

As predicted by our simple framework, this correlation is strongly predictive of institutional filtering across cognitive tasks ($r \approx 0.76 - 0.91$). In tasks in which the confidence-performance correlation is strongly positive, self-selection in social institutions effectively filters errors. These results suggest that in order to understand and predict to what degree cognitive errors will be "filtered out" of aggregate quantities through self-selection, we must understand the precise distribution of confidence, rather than the average level of overconfidence in the population.

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 confidence-performance correlations. Nonetheless, in an exploratory analysis, we use the "peakedness" of the distribution of answers to classify tasks according to whether cognitive errors reflect strong misleading intuitions or a high degree of complexity. We identify some tentative evidence that cognitive tasks that evoke strong intuitions (such as correlation neglect) are associated with lower confidence-performance correlations than tasks that do not evoke a strong gut feeling (such as backward induction).

The importance of directly measuring the performance-confidence correlation is reinforced by the observation that it is nontrivial for economists to forecast the magnitude of institutional filtering or the size of the confidence-performance correlation ex ante. To underscore this point, we ran a survey asking a panel of experts to guess, for a variety of cognitive tasks, (i) to what degree performance and confidence go hand-in-hand, and (ii) to which degree one of our institutions (auctions) filters errors. We find that while the experts are not far off, they consistently overestimate both the degree of institutional filtering and the confidence-performance correlation. Moreover, the experts underpredict variation 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. (ii) We document that understanding or predicting institutional filtering of a given cognitive bias requires that we take into account the confidence-performance correlation in the population (rather than, e.g., the frequency of errors or the average level of overconfidence). This is especially valuable from a methodological perspective because it suggests a simple blueprint: researchers who study cognitive biases can 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 some of the first systematic evidence on the confidence-performance correlation across a large set of widely studied behavioral economics biases.

Our paper ties into several literatures. First, our work relates to an ongoing discussion about when behavioral anomalies affect aggregate quantities (e.g., Russell and Thaler 1985; List 2003; Barberis and Thaler 2003; Fehr and Tyran 2005; Sonnemann et al. 2013). Various experimental contributions have studied the effect of social institutions such as markets and groups on several biases and economic behaviors (e.g., Camerer 1987; Friedman 2010; Charness and Sutter 2012).[3]

Second, our paper relates to work on self-selection. Most closely related is the literature on excess entry (Camerer and Lovallo 1999; Cain, Moore, and Haran 2015; Hollard and Perez 2021), which studies the link between individual confidence and market (or game) entry.[4] Our main contribution to this line of work is to study the effectiveness of social institutions more systematically for a broad set of cognitive tasks, and to show that the performance-confidence correlation is an effective way to conceptualize and empirically predict how and why institutional effects differ strongly across cognitive biases.

Third, our work relates to the literature on how different aspects of confidence and self-awareness vary across cognitive tasks (Moore and Healy 2008). While we highlight and measure the confidence-performance correlation, various earlier literatures have studied how types of average overconfidence vary across tasks, such as in the hard-easy effect (e.g., Koriat, Lichtenstein, and Fischhoff 1980; Erev, Wallsten, and Budescu 1994; Moore and Cain 2007).[5] A prominent example of this body of research is work in psychology on the so-called "bias blind spot": the tendency for people to believe they are less susceptible to behavioral biases than other people are (e.g., Scopelliti et al. 2015; Pronin, Gilovich, and Ross 2004; Pronin, Lin, and Ross 2002; Pronin 2007). A typical paradigm in this literature asks subjects to self-report on a qualitative scale whether they or others are more likely to fall prey to some bias that is described to them verbally. These papers document that there is substantial variation in average overplacement ("relative overconfidence") across tasks. A main difference between our work and this line of research is that we study (and argue for the importance of) the confidence-performance correlation. Indeed, a key conceptual point that emerges from our theoretical and empirical analysis is that the average degree of overconfidence or overplacement (the focus of virtually all of this literature) is largely irrelevant for understanding whether self-selection into institutions attenuates cognitive biases.

Most closely related to our paper is work that, like us, directly examines the confidence-performance correlation, which is variably referred to as "relative accuracy," "monitoring resolution," or "discrimination" in the psychology literature

[3] While we study self-selection (voluntary choice to remove oneself from, e.g., markets), Kendall and Oprea (2018) study the "market selection hypothesis," which asks whether markets reduce biases because error-prone people eventually run out of investment funds.

[4] Odean (1998, 1999) studies the distribution of overconfidence across investor types and its implications for overtrading and other market anomalies.

[5] Loosely related is also work on (lack of) awareness of present bias (O'Donoghue and Rabin 1999) and its implications for market outcomes and welfare (e.g., Carrera et al. 2022; John 2020).

(e.g., Yaniv, Yates, and Smith 1991; Nelson 1984).[6] Psychological experiments show a link between the confidence-performance correlation and the effectiveness of groups in aggregating individual knowledge (e.g., Sniezek and Van Swol 2001; Silver, Mellers, and Tetlock 2021; see also the overview in online Appendix Table 3). The main differences relative to our work are that (i) this literature studies knowledge questions ("wisdom of crowd" effects) rather than cognitive biases as we do here, and (ii) it does not study the formal institutions of interest to economists like markets.

The paper proceeds as follows. Section I lays out our experimental design and Section II derives our predictions. Section III presents results on institutional improvements across tasks and Section IV examines the role of the confidence-performance correlation. Section V reports on our expert survey. Section VI concludes.

I. Experimental Design

A. *Overview*

Our goal is to design an experiment to answer three questions. First, to what degree does self-selection in basic economic institutions filter out the effects of different cognitive biases? Second, how does the confidence-performance correlation vary across biases that are commonly studied by behavioral and experimental economists? And third, how strongly is the institutional filtering of different biases predicted by variation in this confidence-performance correlation?

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:** The subject participates in an anonymous social institution that involves a ten-person cohort: 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.

[6] Much of this work originates from earlier theoretical work that decomposed the quality of forecasts into components that include the confidence-performance correlation (Yates 1982; Murphy 1973).

B. *Part 1: Cognitive Tasks*

We selected 15 cognitive tasks based on four principles. First, we wanted tasks with associated biases that 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 focused on tasks associated with cognitive rather than motivational biases such as present bias. Fourth, 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.

Our objective was not to select a set of tasks that is representative of the range of tasks we believe people encounter in everyday life. Instead, we deliberately sought out tasks that produce well-documented biases, i.e., on which a considerable fraction of people perform poorly. This approach is warranted here because our research question (how institutions filter biases) is predicated on the existence of biases in the first place. Nonetheless, we strove to select biases that are broadly representative of those studied in the *cognitive bias* literature.[7]

We summarize the tasks in Table 1 and provide more details in online Appendix A. Task instructions are provided in online Appendix F. 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, logic problems, errors in strategic reasoning (behavioral game theory), failure to identify constrained optima, and various errors related to financial decision-making.

A central distinction between our selection of tasks and previous work comparing collections of biases is that we deliberately avoided motivational, or preference-related, behavioral regularities, which motivate the majority of tasks studied in Stango and Zinman (2020) and Chapman et al. (2018). We avoided preference-oriented anomalies because our research questions require tasks in which there is a clearly identifiable "right answer," which is typically not available for preference anomalies. Like Stango and Zinman (2020), our set of tasks includes the gambler's fallacy and exponential growth bias.

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. Clearly, the requirement that a response be exactly optimal is more demanding in some tasks than in others for a variety of reasons, and we discuss the corresponding considerations in Section IF.

[7] The confidence-performance correlation has been shown to depend on task features such as average performance (see, e.g., Koriat 2012). Consequently, we caution against interpreting our findings across the set of 15 tasks as being representative for alternative sets of tasks.

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.[d] Adaptation of taxi-cab problem from Tversky and Kahneman (1982).
Correlation neglect (CN)	Failing to account for nonindependence of data in inference.[d] Adaptation of tasks from Enke and Zimmermann (2019).
Balls-and-urns belief updating (BU)	Failure to calculate Bayesian posterior.[d] State probabilistic beliefs about which urn a colored ball is drawn from.
Gambler's fallacy (GF)	Failing to properly attribute independence to i.i.d. draws.[b] 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.[b] 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.[b] Adaptation of task from Kahneman and Tversky (1973).
Acquiring a company (AC)	Failing to properly condition on contingencies, à la the winner's curse.[a] Bidding task against computer as in Charness and Levin (2009).
Logic	
Wason task (WAS)	Failure to gather valuable evidence/positive hypothesis testing.[b] Adaptation of four-card task from Wason (1968).
Cognitive reflection test (CRT)	Following intuitive but misleading "system 1" intuitions.[b] Adaptation of Frederick (2005).
Strategic reasoning	
Backw. ind./iter. reason. (IR)	Limited depth of reasoning in recursive reasoning problems.[a] One-player beauty contest game, à la Bosch-Rosa and Meissner (2020).
Equilibrium reason. (EQ)	Failure to forecast effects of incentives in dominance solvable games.[b] Identify higher earning payoff matrix, adapted from Dal Bó, Dal Bó, and Eyster (2018).
Constrained optimization	
Knapsack (KS)	Failure to identify optimal bundle in constrained optimization problem.[a] Knapsack problems taken from Murawski and Bossaerts (2016).
Financial reasoning	
Thinking at the margin (TM)	Thinking about average instead of marginal costs/benefits.[a] Adaptation of marginal tax task from Rees-Jones and Taubinsky (2020).
Portfolio choice (PC)	Failure to construct efficient portfolios due to 1/N heuristic.[a] Choose optimal portfolio versus dominated 1/N portfolio.
Exponential growth bias (EGB)	Underestimate the exponential effects of compounding.[c] Interest rate forecasting problem adapted from Levy and Tasoff (2016).

Notes: Symbols indicate part 1 payoff function in experimental currency units (ECUs).
[a] Payoffs correspond to implied game payoffs as described in the task;
[b] 100 ECUs if optimal choice, nothing otherwise;
[c] $100 - d$; and
[d] $100 - 3d$, where d = difference between response and expected payoff-maximizing/Bayesian response.

C. *Part 2: Social Institutions*

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. Our goal is not to exhaustively cover every conceivable type of institution but, rather, to focus on a few maximally simple institutions that are widely studied by economists. 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 bettors in speculative markets will be incentivized to bet more aggressively than less well-informed bettors, 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 confidence and performance are correlated, 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 the confidence-performance correlation is strongly positive, 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:

- **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. This self-selection could cause the committee'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. 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.

Implementation and Institutional Details.—For the experiment, we aimed to find the simplest possible version of these institutions: (i) implementations that are static and require only a single, simple decision from each participant and (ii) implementations in which the self-selection decision can be represented for subjects in a very similar fashion.

Betting Markets—Parimutuel Betting: We implemented a *parimutuel betting market*, a particularly simple betting institution. In it, bettors submit monetary bets on multiple securities, only one of which will turn out to be valuable. The total money bet is then redistributed to bettors on the winning security in proportion to the amount each of those agents bet. A canonical example for parimutuel betting markets is horse-race betting. However, there are also direct analogies to financial markets, where bettors 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 to real-world markets (e.g., Plott, Wit, and Yang 2003).

In our implementation, participants were informed that a cohort of nine 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 online Appendix Figure 7 for a sample 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 reweighting 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, $\bar{x}$, 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 1 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 part 1 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 ECUs and decides how many to bid using a slider, see online Appendix Figure 9. The five highest bidders win the auction and pay their own bid.[8] 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 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 $\bar{x}$, 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 can, again, compare this outcome of the auction to 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 online Appendix Figure 8.

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, $\pi_i^{Committee} = 100 \times \theta^{Committee}$. 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

[8] 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.

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.

D. *Measuring Confidence*

Throughout the paper, confidence is defined as the strength of belief in (the probability assigned to) the ex ante optimality (rationality) of one's decision. We implement simple binary notions of optimality, such that a given part 1 answer can unambiguously be classified as objectively correct or false.

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 make institutional decisions ("within-subjects design"). Second, one could elicit confidence in a "between-subjects design," in which those subjects who report their confidence never make any institutional decisions, and vice versa. The two potential designs each 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 make 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-à-vis the experimenter.

On the other hand, a between-subjects design introduces nontrivial measurement error. If we relate the institutional improvements observed in one sample of subjects with the confidence-performance correlations observed in another, the correlation will be attenuated in any finite sample because of the attempt to link behaviors from two different groups of people. Moreover, a between-subjects design does not allow us 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 of the resulting treatment design.

Between-Subjects Design: For this we run an additional treatment, Confidence, that follows the same outline as the institutions treatments discussed above, consisting 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 instructions introduce the idea of an "optimal decision" to subjects, which we define as "the decision that maximizes your earnings, on average."[9] The confidence question then asks: "How certain are

[9] 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

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 robustness treatments are reported throughout the paper as they become relevant.

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 percent and 100 percent, with no initialization for the slider; see online Appendix Figure 10. Following the classification of confidence types offered by Moore and Healy (2008), we note that our item-level confidence measures reflect both "estimation" (belief in one's ability) and "precision" (confidence in the accuracy of beliefs).

A main design objective for us is to make our insights about the predictability of institutional filtering based on confidence data portable and scalable to different experiments and surveys. Thus, we designed the confidence elicitation to be as simple as possible, which means that we deliberately do not financially incentivize it. To the degree that this produces noisier data than an incentivized elicitation would, our results provide a lower bound estimate of the role of the confidence-performance correlation for institutional filtering.

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 made an incentivized institutional decision.

E. *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 cognitive 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

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.

insight of Danz, Vesterlund, and Wilson (2022) that simple scoring rules are most effective in inducing truth telling, our overarching goal was to keep the incentive structure relatively simple and transparent. Online Appendix F 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 ECUs if their response was correct and nothing otherwise. In tasks of type (iii), we deployed simple linear scoring rules with maximum payoffs of 100 ECUs, 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 ECUs earned, while for the remaining 785 subjects it was \$10 per 100 ECUs. 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 preregister 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 ECUs earned.

F. *Optimality and Confidence*

Given the wide variety of cognitive biases that we study, many formal features of the underlying tasks (response scale, incentive structure, stochasticity of the environment) vary. We discuss here why this is irrelevant from the perspective of our objective of measuring the confidence-performance correlation.

First, we opted for a binary definition of optimality that allows us to use the same performance metric across tasks. 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. Nonetheless, the requirement that a response be exactly optimal is more demanding in tasks that have continuous response scales rather than a discrete (e.g., binary) response scale. This might affect optimality rates in a task. However, this is fine for our purposes because our interest is not variation in the *level* of performance (or confidence) across tasks but instead in the *confidence-performance correlation*. For example, it is likely the case that subjects' confidence is mechanically higher in tasks that have response scales with only a few potential response options, yet for the same reason optimality rates will also be higher.

Second, and relatedly, the steepness of the incentives varies with the cardinality of the response scales and other features of the tasks. Again, this may affect the level of performance, but our interest is only in the confidence-performance correlation.

Third, some tasks are deterministic, while others have stochastic environments. When a task has a stochastic state, we elicit confidence about the ex ante optimality of the decision. Thus, confidence always captures the *perceived proficiency of solving a task given the available information*, rather than imperfect information at the time of the decision. As a result, confidence always applies to the same notion of ex ante optimality. Notice that this implies that confidence is 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 exhibit strictly positive variance.

G. *Logistics*

All experiments were conducted on Prolific. We preregistered 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. All experimental data were collected in June 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 toward the number of preregistered completes. See online Appendix F for the comprehension check questions in all treatments.

We preregistered two aspects of our experiments.[10] First, we preregistered 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 preregistered that we would conduct three types of analyses: (i) the performance improvement that is caused by an institution, (ii) the relationship between confidence and institutional choices across tasks, and (iii) the extent to which the correlation between performance and confidence predicts for which tasks we observe larger institutional improvements.

II. Framework and Hypotheses

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

[10] https://aspredicted.org/hg4zi.pdf.

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 preinstitutional 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 $\theta^{pre} \equiv E[\Theta^{pre}] = \frac{1}{N}\sum_i p_i$. 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 $\theta^{post} \equiv E[\Theta^{post}]$ 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 depends on the distribution of self-selection in the population. The way this dependence works varies slightly across the institutions we consider. As discussed in Section IC, for Betting and Committee 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

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

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:

(i) The relationship, β, between confidence, c_i, and expected task performance, p_i.

(ii) 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:[11]

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

Rather than viewing equation (8) as a behavioral microfoundation 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. Throughout the paper, we refer to the relationship between confidence and performance as *confidence-performance correlation*. Average overconfidence (in this setting a metric which combines overestimation and overprecision), $d \equiv \bar{c} - \bar{p} = \alpha + (\beta - 1)\bar{p}$, is a function of both α and β.

The expression in equation (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 be partly governed by risk aversion or higher-order beliefs about others' confidence.

These relationships allow us to derive a preregistered prediction about the relationship between institutional filtering ($\mathbb{G}$) and a summary statistic of the distribution of confidence.[12]

[11] 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 $k_i \in (0, 1)$.

[12] All predictions that we state 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 0, people's average subjective probability that they are correct is strictly larger than 0.

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) *Institutional 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 online Appendix B.

Much of the prior literature focuses on average overconfidence (average overestimation and/or overprecision) in a task. In contrast, our prediction highlights that the degree to which self-selection produces institutional improvements depends on the *confidence-performance correlation* across agents (the slope between confidence and performance). The following prediction clarifies the role of average overconfidence (again, here, a statistic that combines overestimation and overprecision). In contrast to the first prediction above, this one was not preregistered, 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. To see the intuition for why average overconfidence should not exhibit a clear relationship with institutional filtering, consider the auction. Suppose that the confidence of all agents was exogenously increased by the same amount, such that average overconfidence increases. Then, the bid of all agents will also increase by the same amount. This, however, does not change the resulting allocation, which only depends on who bids more rather than on how much. In Betting and Committee, the intuition is a bit more involved because the parimutuel market price and the average vote share are not linear in people's bets and votes. As a result, a uniform confidence shift can under certain conditions lead to a performance decrease (if the confidence shift implies that the bets and votes of more error-prone people increase more strongly in percentage terms).

Variation across Institutions: It is conceivable that the importance of the confidence-performance correlation for filtering out errors varies across the different institutions that we study. 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. All of these mechanisms could lead ω to differ across institutions, such that the confidence-performance correlation matters more in some than in other institutions.

At the same time, there are also two reasons to expect that theoretical differences in the strategic environments across institutions do not translate into differences in how much confidence-performance correlation matters. First, we designed our experiments with the objective in mind to keep the institutions and self-selection decision as simple and similar as possible. For instance, in all three institutions, subjects' decision is essentially given by navigating a slider between 0 and 100 to indicate how intensively they would like to act in the market/auction/committee. It is plausible that this design choice attenuates or even eliminates differences across institutions.

Second, 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 generate variation in ω across institutions. Indeed, it is not obvious why, in practice, people would exhibit the cognitive sophistication to solve for the equilibrium of an institutional mechanism in part 2 if they don't have the cognitive sophistication to solve the part 1 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 confidence-performance correlation.

III. 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 one another. For the sake of brevity, we present here the results from the between-subjects treatments and always refer the reader to corresponding analyses for the within-subjects treatments in the online Appendix.

A. *Performance across Tasks and Subjects*

The average of optimal part 1 responses across all tasks in treatments Betting, Auction, Committee, and Confidence is 28 percent. 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 percent and 30 percent, the total range spans from less than 10 percent to more than 80 percent.[13] Online Appendix Figure 13 provides a complementary subject-level perspective by showing a cumulative distribution function (CDF) of the number of optimal part 1 decisions per subject.

B. *Which Errors Do Social Institutions Filter?*

Recall from Section II 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 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

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

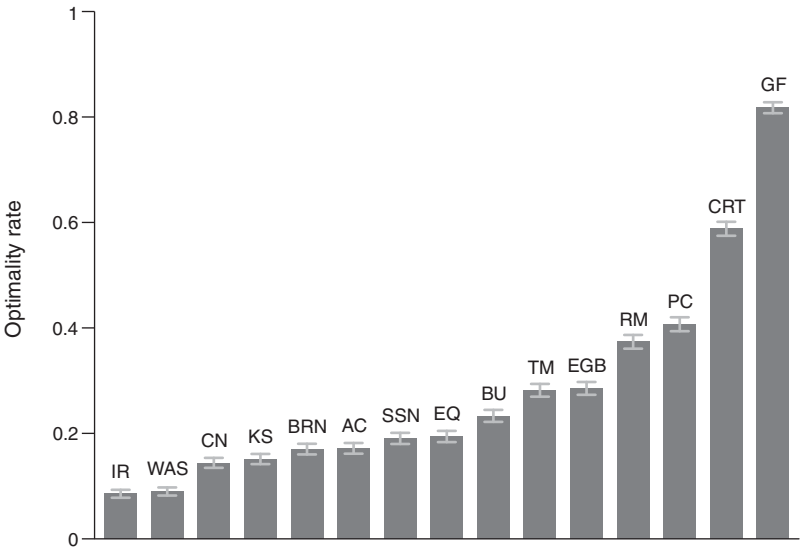

Figure 1. Fraction of Optimal Responses in a Cognitive Task (Part 1 Decision), Separately for Each Task in Treatments Betting, Auction, Committee, and Confidence

Notes: $N = 1{,}381$ participants completed each of the 15 tasks in individually randomized order. The tasks and the corresponding definitions of an optimal response are described in online Appendices A and F Whiskers indicate standard errors of the binomial mean. See Table 1 for task codes.

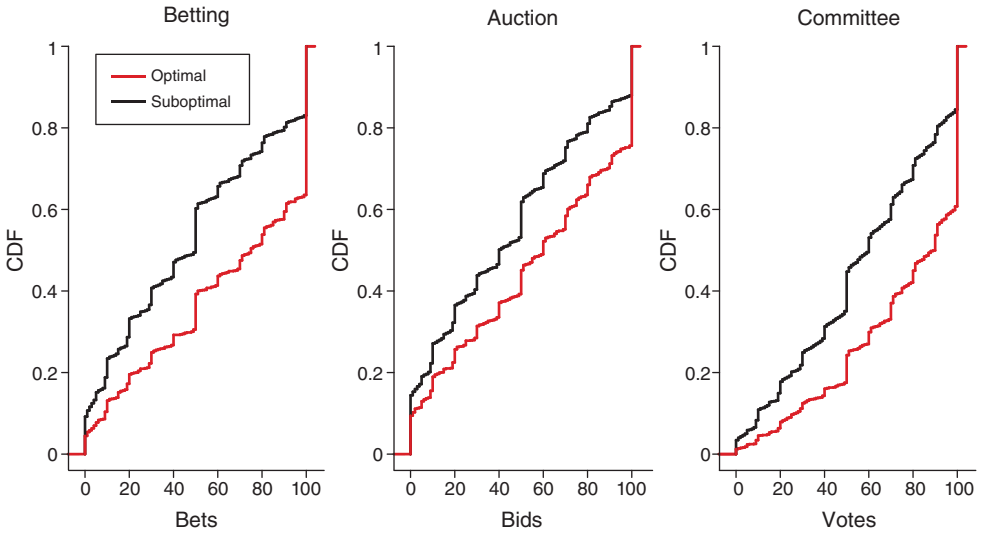

Figure 2. Institutional Choices (Part 2 Decisions), Split by Whether the Response to the Corresponding Cognitive Task (Part 1 Decision) Was Optimal

Notes: For each institution, empirical cumulative distribution functions (CDFs) are displayed. The tasks and the corresponding definitions of an optimal response are described in online Appendices A and F. 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. The fraction of optimal part 1 decisions is 29 percent in Betting, 28 percent in Auction and 28 percent in Committee.

in Betting (64.8 average bet for optimal part 1 decisions and 47.4 for suboptimal, a difference of 37 percent) than in Committee (75 average votes for optimal versus 57.9 for suboptimal, 29 percent), or Auction (56.4 average bid for optimal versus 43.6 for suboptimal, difference of 29 percent).[14]

These patterns immediately imply that, on average across tasks, self-selection is positive: 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 percent. Further suppose that, in the committee institution, those 5 subjects that got the task right each submit 100 votes, that 1 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 (base rate neglect), or CN (correlation neglect).[15]

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, selection is positive, meaning that it is mostly people who make suboptimal decisions that select out. In other tasks, however, optimal and suboptimal decision makers make roughly the same Part 2 decisions, such that selection can even be negative. Indeed, the within-task correlation between bids/bets/votes and optimality ranges from $r = -0.12$ in AC bets to $r = 0.49$ in EGB bets; see online Appendix Figure 16.

Although there is some variation in which tasks are most and least improved across institutions, there is for the most part 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.

Efficiency of Institutions: Our main measure of institutional improvement is a measure of *absolute* improvement. An alternative is to consider the efficiency of institutions in reducing biases: *what fraction of the theoretically possible improvement*

[14] Online Appendix Figure 26 provides an analogous analysis for the within-subjects treatments.
[15] Online Appendix Figure 27 provides an analogous analysis for the within-subjects treatments.

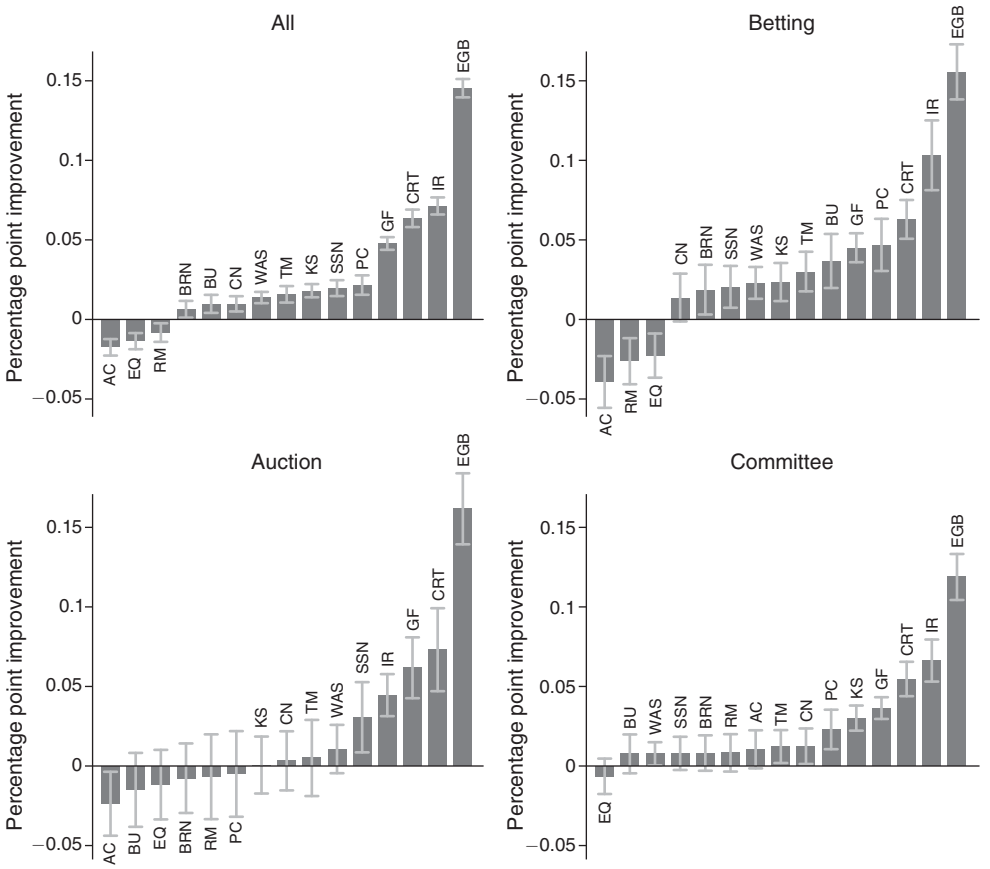

FIGURE 3. PERFORMANCE IMPROVEMENT THROUGH INSTITUTIONS ACROSS TASKS

Notes: Percentage point improvement is computed as the aggregate performance of the institutional summary statistic minus the raw fraction of optimal responses in a cognitive task. The institutional summary statistics are given by the parimutuel market price in Betting, the average rate of bias among the set of winners in the Auction, and the vote share for the optimal decision in Committee. The aggregate performance is based on 10,000 randomly constructed 10-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). See Table 1 for task codes.

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 selected the most competent players, the postinstitution performance would not be 100 percent 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 postinstitutional performance to be 100 percent, 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, which happens relatively rarely. To account for this, we compute an efficiency metric of improvement. This measure is given by the fraction of the theoretically possible improvement (given the distribution of performance among subjects) that is actually achieved by an institution.

Online Appendix Figure 14 shows that the efficiency of the institutions in reducing bias also strongly varies across tasks. For example, in the Auction treatment, the institution's efficiency ranges from -15 percent for AC to 70 percent for iterated reasoning. We conclude from this exercise that the efficiency of canonical economic institutions strongly depends on the particular cognitive bias.

IV. The Role of the Confidence-Performance Correlation

A. *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 percent, while it is 64 percent in the pool of suboptimal decisions; see Online Appendix Figure 15 As in previous work, we find that individual-level heterogeneity in confidence is correlated with demographics; see online Appendix Table 4: (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 not in the rate of overconfidence in the population. Rather, based on our theoretical framework, our main interest is in the variation of the performance-confidence correlation. Figure 4 shows the within-task correlation between part 1 optimality and part 2 confidence across our 15 tasks. We see large variation across tasks. 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, suboptimal 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. Online Appendix Figure 28 presents an analogous analysis for the within-subjects treatments.

Measurement Error: As in any experiment with a finite sample, some of the across-task variation in the confidence-performance correlation that is displayed in Figure 4 will reflect measurement error. The figure clarifies this by displaying standard errors of the estimated correlation coefficients between confidence and optimality.

Because we view the documentation of across-task differences in the confidence-performance correlation (and its predictive power for institutional filtering) as one of the main contributions of this paper, we investigate through simulations how likely it is that the entirety of the variation across tasks could be spurious and driven by random noise. To this effect, we implement the following procedure. In treatment Confidence, we take as given the empirical marginal distributions of both confidence and optimality in each task. We then randomly scramble these two variables such that, in expectation, in each task, the confidence-optimality correlation is identical and equal to zero (what matters is not that the expected confidence-performance correlation is zero in all tasks, only that it is identical across

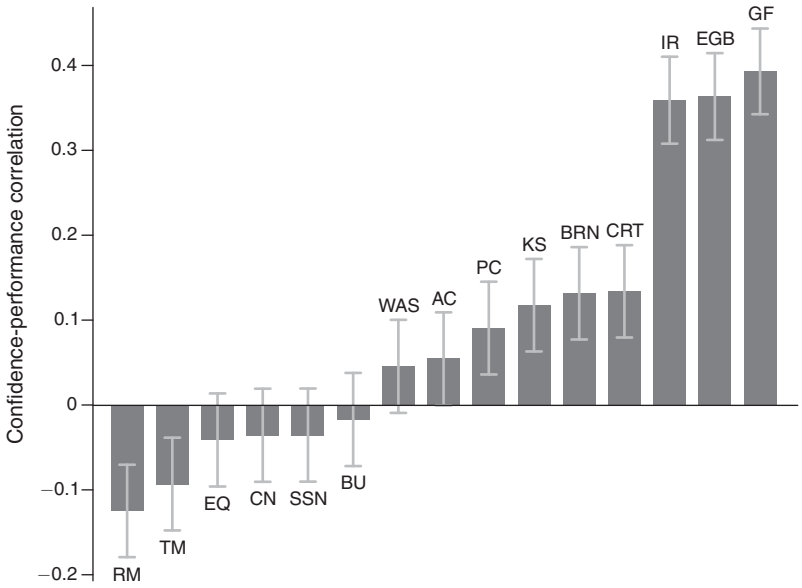

FIGURE 4. WITHIN-TASK CORRELATION BETWEEN THE OPTIMALITY OF A RESPONSE TO A COGNITIVE TASK (PART 1 DECISION) AND STATED CONFIDENCE, SEPARATELY FOR EACH TASK IN TREATMENT CONFIDENCE

Notes: Displayed are Pearson correlation coefficients, based on $N = 334$ participants. The definitions of an optimal response in each task are provided in online Appendix A Whiskers indicate standard errors of the estimated correlation coefficients. See Table 1 for task codes.

tasks). Because in any given simulation the actual confidence-optimality correlation in any given task will not be zero (due to the finite sample of 334 subjects), these simulations tell us what degree of across-task variation in the confidence-optimality correlation will purely result from noise in finite samples. We implement this procedure 10,000 times and analyze how the resulting distribution of across-task variations compares to our actually observed across-task variation. Specifically, we consider how the observed across-task standard deviation (and range) of the confidence-optimality correlation compares with the simulated distribution. Online Appendix Figures 17 and 18 show the details. We find that the empirically observed standard deviation and range are substantially larger than in every single simulation. This suggests that the probability that random noise generates the entire across-task variation in Figure 4 is essentially zero.

B. *The Confidence-Performance Correlation and Institutional Improvement*

Our hypothesis is that the sign and magnitude of the optimality-confidence correlation illustrated in Figure 4 are predictive of institutional improvement. As discussed in Section ID, 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

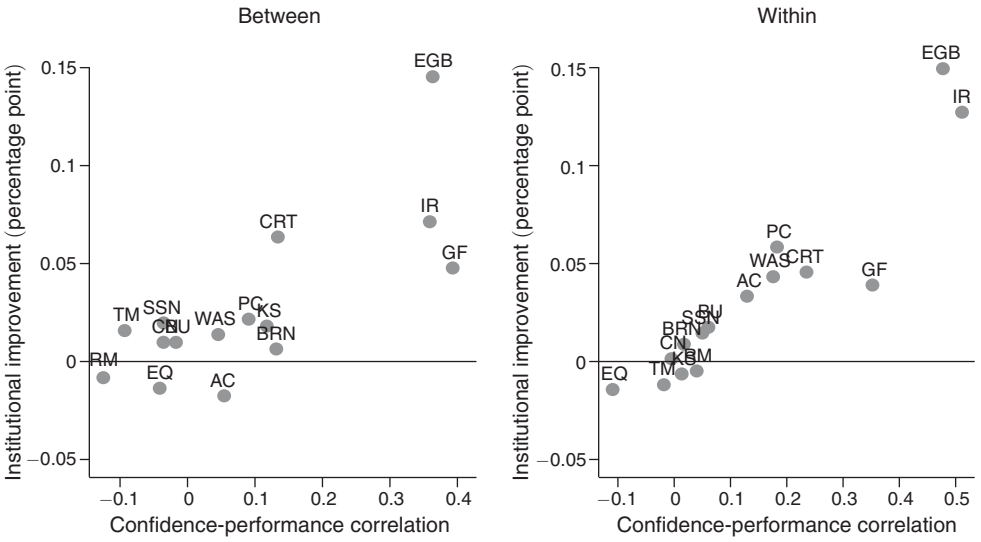

FIGURE 5. CONFIDENCE-PERFORMANCE CORRELATION AND INSTITUTIONAL IMPROVEMENT

Notes: 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. Percentage point improvement is computed as the aggregate performance of the institutional summary statistic minus the raw fraction of optimal responses in a cognitive task, averaged across institutions. The institutional summary statistics are given by the parimutuel market price in Betting, the average rate of bias among the set of winners in the Auction and the vote share for the optimal decision in Committee. The aggregate performance is based on 10,000 randomly constructed ten-subject cohorts for each institution, taking the mean over all samples. See Table 1 for task codes.

predict the institutional improvement observed in one sample of subjects with the confidence-performance correlation 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 predict here the institutional improvement observed in one sample of subjects with the confidence-performance correlation of those same subjects.

We make two main observations. First, the figures visually confirm our hypothesis. In tasks with a strong confidence-performance correlation, the institutional improvement is large. This is the case for tasks such as EGB, IR, and GF. Opposite patterns hold for attribution (understanding regression to the mean), TM, CN, and EQ. 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.

Note that because we generally estimate both the confidence-performance correlation and the institutional improvement with random noise, the patterns displayed in Figure 5 are likely to be attenuated relative to their true magnitudes by standard measurement error arguments. Indeed, as discussed in Section ID, through this lens it is also unsurprising to see that the visual link between confidence-performance

correlation and institutional improvement is stronger in the within-subjects treatments because in any finite sample the between-subjects approach necessarily introduces additional measurement error relative to the within-subjects treatments because institutional improvement and confidence-performance correlation are observed in different samples of people. In any case, as we hypothesized, the relationship between institutional improvement and confidence-performance correlation is always strong.

Variation across Institutions: An immediate question is whether the predictability of institutional improvement through the confidence-performance correlation 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 confidence-performance correlation are $r^{auction} = 0.69$, $r^{betting} = 0.73$, $r^{committee} = 0.77$, $r^{auction,within} = 0.9$, $r^{betting,within} = 0.9$, and $r^{committee,within} = 0.91$; see online Appendix Figures 20 and 30. There are two different interpretations of this similarity across institutions, both of which we embrace. A first is that this result is surprising because, from a theoretical perspective, it is conceivable that confidence matters to a different quantitative degree in some institutions than in others. A second interpretation, however, is that this similarity is unsurprising because we specifically designed the institutions to be as simple as possible, including that the self-selection decision is very similar implementation-wise across institutions (a slider between 0 and 100).

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 more confident subjects are more likely to behave aggressively in the institution: that they bet less, bid lower amounts, and submit fewer votes. Through the lens of our conceptual framework in Section II, this amounts to saying that $\omega > 0$. In our within-subjects treatments, we can directly test this assumption. Online Appendix Figure 29 shows binned scatterplots of institutional actions against stated confidence separately for each institution. We find that the correlations between confidence and bids, bets, and votes are $r = 0.79$, $r = 0.85$, and $r = 0.89$, respectively.[16] In our between-subjects treatments, subjects never both report their confidence and make an institutional decision, such that it is impossible to report the correlation.

Predictability of Efficiency of Institutions: Our main analysis considers absolute institutional improvement. In Section III, we additionally introduced a measure that gauges the efficiency of institutions in reducing bias, relative to how much it could reduce bias given the distribution of performance in the population. A natural question is whether the confidence-performance correlation is also predictive of institutional efficiency. Online Appendix Figure 21 shows that the confidence-performance

[16] Online 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.

correlation turns out to be an even stronger predictor of efficiency, with r rising to 0.87 in the between-subjects data and to $r = 0.94$ in the within-subjects data.

Robustness—Sensitivity to Outliers: To examine whether our finding 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 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.

C. *Which Types of Errors Have Strong Confidence-Performance Correlation?*

Our results raise the question of *what characteristics* of tasks make decision makers' beliefs more predictive of their performance? The broader question of how people's self-awareness of their own errors varies across different cognitive tasks has received a fair amount of attention in the literature, such as in the hard-easy effect or the "bias blind spot" literature in psychology. However, this body of work focuses on how average overconfidence (or average overplacement) vary across tasks. For our purposes, the relevant object of interest is, instead, the performance-confidence correlation.

Given that we are looking at a moderately sized sample of tasks, an analysis of this question is naturally very 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. Still, 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 backward induction, constrained optimization in the knapsack problem (KS), or the AC 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-performance correlation 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, Turner, and Pennycook 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.[17] 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 cognitive reflection test (CRT) or CN, 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

[17] Relatedly, there is some evidence from the psychology literature that the confidence-performance correlation is high in those tasks that produce correct answers and low in tasks that tend to produce incorrect answers (the "consensuality" principle). See, e.g., Koriat (2008, 2012).

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.

Online Appendix Figure 22 provides some tentative evidence that tasks in which wrong responses are strongly peaked ("intuition problems") see somewhat smaller confidence-performance correlations, though the results are a bit mixed. For example, in CN and BU, about 50 percent of all responses are concentrated on a single answer, and the optimality-confidence correlation in these tasks is roughly zero. In IR, EGB, and KS, on the other hand, the fraction of concentrated responses is between 10 percent and 30 percent, 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 three reasons. First, it is based on only nine tasks. Second, it ignores that the response scales across these nine tasks differ widely. Third, the results using the peakedness measure are relatively noisy. Future research is needed to shed more light on the determinants of the quality of the confidence-performance correlation.

D. *The Role of Average Overconfidence*

Our conceptual framework in Section II 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.

Online Appendix Figure 23 plots average confidence in treatment Confidence against the optimization rate in each task.[18] We observe two main patterns: First, there is average overconfidence in all of the 15 tasks. Second, average confidence and the optimization rate are strongly positively correlated across tasks, $r = 0.75$. 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 (e.g., Erev, Wallsten, and Budescu 1994; Moore and Healy 2008).

How does such average overconfidence relate to institutional behavior and resulting performance improvements? As discussed in Section II, 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. Online Appendix Figure 24 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

[18] Online Appendix Figure 31 shows the results for the within-subjects treatments.

from 0 at conventional levels).[19] These results are in line with Prediction 2 from our framework in Section II, and highlight that what matters for institutional filtering is indeed mostly the confidence-optimality correlation, rather than average overconfidence.

V. 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.[20] 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 in November 2021. Among those who indicated their professional level, 57 percent are faculty at all levels, 10 percent are postdoctoral researchers, and 33 percent are graduate students. Over 85 percent 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.[21] 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-performance correlations. Screenshots of the elicitation screens are reproduced in online Appendix Figures 11 and 12.

Panel A of Figure 6 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 calibration of the experts. We make four observations. First, expert forecasts track the distribution of tasks reasonably well: across tasks, higher average forecasts of institutional improvement and confidence differences tend to be associated with higher actual improvement and confidence differences. Second, average expert forecasts of improvement and confidence differences tend to be internally consistent with our framework: if one believes that confidence differences are larger than they actually are, then one should also believe that the institutional improvement will be larger than it actually is. This good qualitative calibration of forecasts resonates with previous findings on the accuracy of expert forecasts (see, e.g., DellaVigna, Pope, and Vivalt 2019), but is accompanied by quantitative miscalibration in our data. Specifically, third, experts generally overpredict both the magnitude of institutional improvement and the magnitude of confidence differences between

[19] 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.

[20] 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.

[21] 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.

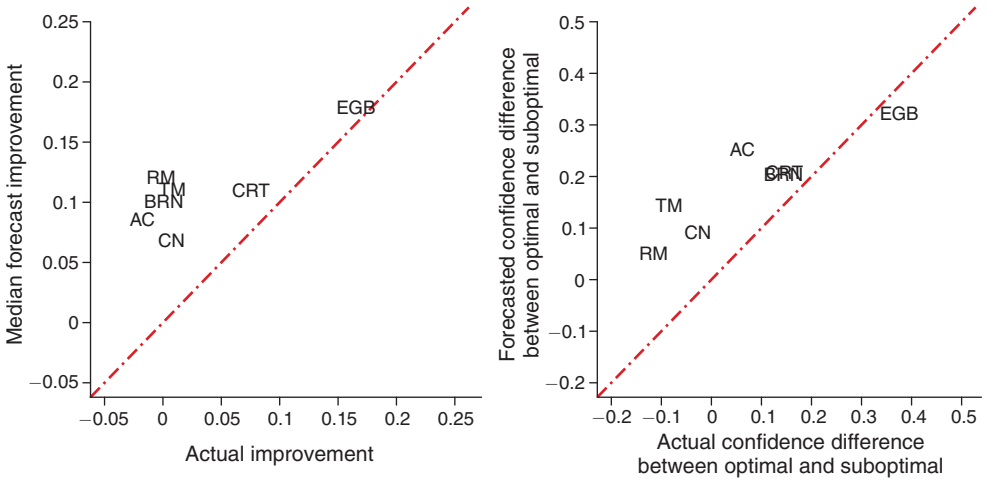

FIGURE 6. EXPERT FORECASTS AND EMPIRICAL ANALOGUES

Notes: The left panel plots predicted institutional improvements against actual ones, separately for each task. The right panel plots the predicted difference in confidence between optimal and suboptimal decisions against the true difference. The diagonal line indicates perfect calibration. See Table 1 for task codes.

optimal and nonoptimal subjects. Fourth, panels A and B of Figure 6 show that the expert forecasts are excessively compressed relative to the truth: experts predict that the degree of confidence differences and institutional improvement are more similar across tasks than is actually the case.

VI. 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 understate the influence biases are likely to have for the aggregate outcomes produced by real-world institutions.

In this paper, we take some steps toward 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. 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 the predictiveness of subjects' beliefs about their own decision quality: 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 here summarize implications and limitations of our work in order to help inform future research on the topic.

Limitations: We deliberately studied only the simplest variants of institutions, and designed the self-selection decision to be operationally extremely similar across institutions. This probably minimizes any latent differences in self-selection across institutional contexts. This comes at the cost that our design is not well suited to intensively investigate how strongly institutions can differ in their scope for self-selection. We think this encourages follow-up work that implements dynamic, feedback-rich variations of markets and organizations that allow subjects to make more sophisticated, experienced choices, allowing any latent differences across these institutions to show themselves.

A second limitation of our paper is that we do not consider the welfare effects of self-selection. Rather, we only focus on the efficiency of the aggregate quantity that an institution produces. Yet, in many contexts, it may be that self-selecting out of a market has real intrinsic costs (for example, when a person should purchase insurance but selects out of the market due to confusion). In such cases, the welfare benefits of self-selection (debiased aggregate quantities) are counteracted by the welfare costs of nonparticipation. Future research may helpfully study when and why which of these two effects dominates.

Methodological Takeaway: Future researchers can tentatively gauge 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 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.

Importance of the Confidence-Performance Correlation: One reason why we encourage that future research on cognitive biases report the performance-confidence correlation is that, thus far, most research on confidence tends to focus on average overconfidence or overplacement, which—as we highlight here—is less relevant as a driver of self-selection than the performance-confidence correlation. We, hence, conjecture that more research energies might profitably be spent measuring and understanding this object and its sensitivity to features of the choice environment. There are at 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 the confidence-performance correlation. (ii) Although 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 the confidence-performance correlation. For example, the effect of policy interventions on the confidence-performance link may be every bit as important for social science outcomes as their effect on biases themselves. After all, casual introspection suggests that public policy might affect not only whether

people make mistakes but also which people are aware of their mistakes. (iii) For future theorizing and practical predictions, it would be very useful to understand *why* it is that in some tasks people's confidence and performance are reasonably tightly linked, but not in others. Why do sometimes the "right" and other times the "wrong" people believe that they are getting things wrong?

REFERENCES

Barberis, Nicholas, and Richard Thaler. 2003. "A Survey of Behavioral Finance." In *Handbook of the Economics of Finance*, Vol. 1B, edited by Nicholas Barberis and Richard Thaler, 1053–1128. Amsterdam: Elsevier.

Bar-Hillel, Maya. 1979. "The Role of Sample Size in Sample Evaluation." *Organizational Behavior and Human Performance* 24 (2): 245–57.

Bosch-Rosa, Ciril, and Thomas Meissner. 2020. "The One Player Guessing Game: A Diagnosis on the Relationship between Equilibrium Play, Beliefs, and Best Responses." *Experimental Economics* 23 (4): 1129–47.

Cain, Daylian M., Don A. Moore, and Uriel Haran. 2015. "Making Sense of Overconfidence in Market Entry." *Strategic Management Journal* 36 (1): 1–18.

Camerer, Colin F. 1987. "Do Biases in Probability Judgment Matter in Markets? Experimental Evidence." *American Economic Review* 77 (5): 981–97.

Camerer, Colin, and Dan Lovallo. 1999. "Overconfidence and Excess Entry: An Experimental Approach." *American Economic Review* 89 (1): 306–18.

Carrera, Mariana, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky. 2022. "Who Chooses Commitment? Evidence and Welfare Implications." *Review of Economic Studies* 89 (3): 1205–44.

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer. 2018. "Econographics." NBER Working Paper 24931.

Charness, Gary, and Dan Levin. 2009. "The Origin of the Winner's Curse: A Laboratory Study." *American Economic Journal: Microeconomics* 1 (1): 207–36.

Charness, Gary, and Matthias Sutter. 2012. "Groups Make Better Self-Interested Decisions." *Journal of Economic Perspectives* 26 (3): 157–76.

Dal Bó, Ernesto, Pedro Dal Bó, and Erik Eyster. 2018. "The Demand for Bad Policy when Voters Underappreciate Equilibrium Effects." *Review of Economic Studies* 85 (2): 964–98.

Danz, David, Lise Vesterlund, and Alistair J. Wilson. 2022. "Belief Elicitation and Behavioral Incentive Compatibility." 112 (9): 2851–83.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt. 2019. "Predict Science to Improve Science." *Science* 366 (6464): 428–29.

Dohmen, Thomas, Armin Falk, David Huffman, Felix Marklein, and Uwe Sunde. 2009. "Biased Probability Judgment: Evidence of Incidence and Relationship to Economic Outcomes from a Representative Sample." *Journal of Economic Behavior and Organization* 72 (3): 903–15.

Enke, Benjamin, and Thomas Graeber. 2021a. "Cognitive Uncertainty." NBER Working Paper 26518.

Enke, Benjamin, and Thomas Graeber. 2021b. "Noisy Cognition and Intertemporal Choice." Unpublished.

Enke, Benjamin, Thomas Graebr, and Ryan Oprea. 2023. "Replication Data for: Confidence, Self-Selection, and Bias in the Aggregate." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E185741V1.

Enke, Benjamin, and Florian Zimmermann. 2019. "Correlation Neglect in Belief Formation." *Review of Economic Studies* 86 (1): 313–32.

Erev, Ido, Thomas S. Wallsten, and David V. Budescu. 1994. "Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processes." *Psychological Review* 101 (3): 519.

Fehr, Ernst, and Jean-Robert Tyran. 2005. "Individual Irrationality and Aggregate Outcomes." *Journal of Economic Perspectives* 19 (4): 43–66.

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

Friedman, Daniel. 2010. "Laboratory Financial Markets." In *Behavioural and Experimental Economics*, edited by Steven N. Durlauf and Lawrence E. Blume, 178–85. New York: Springer.

Hollard, Guillaume, and Fabien Perez. 2021. "Self-Selection Filters Irrationality in One-Shot Games." Unpublished.

John, Anett. 2020. "When Commitment Fails: Evidence from a Field Experiment." *Management Science* 66 (2): 503–29.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux.

Kahneman, Daniel, and Amos Tversky. 1972. "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3 (3): 430–54.

Kahneman, Daniel, and Amos Tversky. 1973. "On the Psychology of Prediction." *Psychological Review* 80 (4): 237–51.

Kendall, Chad, and Ryan Oprea. 2018. "Are Biased Beliefs Fit to Survive? An Experimental Test of the Market Selection Hypothesis." *Journal of Economic Theory* 176: 342–71.

Koriat, Asher. 2008. "Subjective Confidence in One's Answers: The Consensuality Principle." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34 (4): 945–59.

Koriat, Asher. 2012. "When Are Two Heads Better than One and Why?" *Science* 336 (6079): 360–62.

Koriat, Asher, Sarah Lichtenstein, and Baruch Fischhoff. 1980. "Reasons for Confidence." *Journal of Experimental Psychology: Human Learning and Memory* 6 (2): 107–18.

Krishna, Vijay. 2009. *Auction Theory.* Amsterdam: Academic Press.

Kruger, Justin, and David Dunning. 1999. "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments." *Journal of Personality and Social Psychology* 77 (6): 1121–34.

Levy, Matthew, and Joshua Tasoff. 2016. "Exponential-Growth Bias and Lifecycle Consumption." *Journal of the European Economic Association* 14 (3): 545–83.

List, John A. 2003. "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics* 118 (1): 41–71.

Moore, Don A., and Daylian M. Cain. 2007. "Overconfidence and Underconfidence: When and Why People Underestimate (and Overestimate) the Competition." *Organizational Behavior and Human Decision Processes* 103 (2): 197–213.

Moore, Don A., and Paul J. Healy. 2008. "The Trouble with Overconfidence." *Psychological Review* 115 (2): 502–17.

Murawski, Carsten, and Peter Bossaerts. 2016. "How Humans Solve Complex Problems: The Case of the Knapsack Problem." *Scientific Reports* 6 (1): 1–10.

Murphy, Allan H. 1973. "A New Vector Partition of the Probability Score." *Journal of Applied Meteorology and Climatology* 12 (4): 595–600.

Nelson, Thomas O. 1984. "A Comparison of Current Measures of the Accuracy of Feeling-of-Knowing Predictions." *Psychological Bulletin* 95 (1): 109–33.

Odean, Terrance. 1998. "Volume, Volatility, Price, and Profit when All Traders are Above Average." *Journal of Finance* 53 (6): 1887–1934.

Odean, Terrance. 1999. "Do Investors Trade Too Much?" *American Economic Review* 89 (5): 1279–98.

O'Donoghue, Ted, and Matthew Rabin. 1999. "Doing It Now or Later." *American Economic Review* 89 (1): 103–24.

Plott, Charles, Jorgen Wit, and Winston Yang. 2003. "Parimutuel Betting Markets as Information Aggregation Devices: Experimental Results." *Economic Theory* 22 (2): 311–51.

Pronin, Emily. 2007. "Perception and Misperception of Bias in Human Judgment." *Trends in Cognitive Sciences* 11 (1): 37–43.

Pronin, Emily, Thomas Gilovich, and Lee Ross. 2004. "Objectivity in the Eye of the Beholder: Divergent Perceptions of Bias in Self versus Others." *Psychological Review* 111 (3): 781–99.

Pronin, Emily, Daniel Y. Lin, and Lee Ross. 2002. "The Bias Blind Spot: Perceptions of Bias in Self versus Others." *Personality and Social Psychology Bulletin* 28 (3): 369–381.

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

Russell, Thomas, and Richard Thaler. 1985. "The Relevance of Quasi Rationality in Competitive Markets." *American Economic Review* 75 (5): 1071–82.

Scopelliti, Irene, Carey K. Morewedge, Erin McCormick, H. Lauren Min, Sophie Lebrecht, and Karim S. Kassam. 2015. "Bias Blind Spot: Structure, Measurement, and Consequences." *Management Science* 61 (10): 2468–86.

Silver, Ike, Barbara A. Mellers, and Philip E. Tetlock. 2021. "Wise Teamwork: Collective Confidence Calibration Predicts the Effectiveness of Group Discussion." *Journal of Experimental Social Psychology* 96: 104157.

Sniezek, Janet A., and Lyn M. van Swol. 2001. "Trust, Confidence, and Expertise in a Judge-Advisor System." *Organizational Behavior and Human Decision Processes* 84 (2): 288–307.

Sonnemann, Ulrich, Colin F. Camerer, Craig R. Fox, and Thomas Langer. 2013. "How Psychological Framing Affects Economic Market Prices in the Lab and Field." *Proceedings of the National Academy of Sciences* 110 (29): 11779–84.

Stango, Victor, and Jonathan Zinman. 2020. "Behavioral Biases Are Temporally Stable." NBER Working Paper 27860.

Stewart, Sharla A. 2005. "Can Behavioral Economics Save Us from Ourselves." *University of Chicago Magazine* 97 (3): 36–42.

Thompson, Valerie A. 2009. "Dual-Process Theories: A Metacognitive Perspective." In *In Two Minds: Dual Processes and Beyond*, edited by Jonathan Evans and Keith Frankish, 171–95. Oxford: Oxford University Press.

Thompson, Valerie A., Jamie A. Prowse Turner, and Gordon Pennycook. 2011. "Intuition, Reason, and Metacognition." *Cognitive Psychology* 63 (3): 107–40.

Tversky, Amos, and Daniel Kahneman. 1982. "Evidential Impact of Base Rates." In *Judgment under Uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic, and Amos Tversky, 153–60. Cambridge, UK: Cambridge University Press.

Wason, Peter C. 1968. "Reasoning about a Rule." *Quarterly Journal of Experimental Psychology* 20 (3): 273–81.

Yaniv, Ilan, J. Frank Yates, and J. Keith Smith. 1991. "Measures of Discrimination Skill in Probabilistic Judgment." *Psychological Bulletin* 110 (3): 611.

Yates, J. Frank. 1982. "External Correspondence: Decompositions of the Mean Probability Score." *Organizational Behavior and Human Performance* 30 (1): 132–56.