Persistent Overconfidence and Biased Memory: Evidence from Managers*

David Huffman Collin Raymond Julia Shvets

May 3, 2019

Abstract

A long-standing puzzle is how overconfidence can persist in settings characterized by repeated feedback. This paper studies managers who participate repeatedly in a high-powered tournament incentive system, learning relative performance each time. Using reduced form and structural methods we find that: (i) managers make overconfident predictions about future performance; (ii) managers have overly-positive memories of past performance; (iii) the two phenomena are linked at an individual level. Our results are consistent with models of motivated beliefs in which individuals preserve unrealistic expectations by distorting memories.

JEL classification:
Keywords: Overconfidence; memory; tournament; motivated beliefs.

*We thank Sandro Ambuehl, Ned Augenblick, Heski Bar-Isaac, Roland Bénabou, Doug Bernheim, Matias Cattaneo, Gary Charness, Steffano DellaVigna, Armin Falk, Yoram Halevy, Paul Heidhues, PJ Healy, Mitch Hoffman, Jun Ishii, Botond Kőszegi, Yiming Liu, George Loewenstein, Ulrike Malmendier, Pietro Ortoleva, Luis Santos-Pinto, Jean Tirole, Lise Vesterlund, Leat Yariv, Florian Zimmerman, and seminar and workshop participants at SITE; Santa Barbara Conference on Experimental and Behavioral Economics; Advances with Field Experiments; CESifo Behavioral Economics; Princeton; Carnegie Mellon; ESA, Tucson; Institute for Research on Behavior and Inequality, Bonn; University of Nottingham, University of Trento, University of Lund, Yonsei University, Hong Kong University, Middlesex University; Pennsylvania State University; University of Cambridge; and Vienna University of Economics and Business for helpful comments. We thank the Keynes Fund for financial support; Martin Oldfield for discussions, and help with logistics and payments; Sidney Ho for help processing payments. For excellent research assistance, we thank Mihir Bhatia, Marta Boczon, Lukas Hensel, Balint Lenkei, Myrto Oikonomou, and Lu Wang. We are grateful to the firm for providing us with the data and to Nick, Eliana, Stephen and Eric for all their help. Huffman: University of Pittsburgh; huffmand@pitt.edu. Raymond: Purdue University; collinbraymond@gmail.com. Shvets: Christ's College, University of Cambridge; js591@cam.ac.uk.

1 Introduction

Overconfidence has often been described as a fundamental bias in human decision making (e.g., Smith, 1776). A long-standing puzzle, however, is whether and how overconfidence can be more than an ephemeral phenomenon. In many of the field settings where economic theory posits a crucial role for beliefs about relative performance – the workplace, school, university, and competitive environments more generally – individuals receive repeated performance feedback. And yet, there is emerging evidence that overconfidence may exist in field settings. For example, the behaviors of investors, CEO's, gym members, and others (Barber and Odean, 2001; Malmendier and Tate, 2005 and 2015; DellaVigna and Malmendier, 2006; Oster et al., 2013, Cheng et al., 2014), and the beliefs of truckers and professional poker players (Hoffman and Burks, 2017; Park and Santos-Pinto, 2010), all show signs of overconfidence in their respective decision environments, even though these individuals are presumably observing signals that should challenge their beliefs.

Economists have considered different mechanisms that might generate persistent overconfidence, but one leading explanation is "motivated beliefs." Simply put, these models assume that individuals desire to be confident (Bénabou and Tirole, 2002; Brunnermeier and Parker, 2005; Kőszegi, 2006, 2006; Gottlieb, 2011; for surveys see Bénabou, 2015 and Bénabou and Tirole, 2016).[1] The precise reason for this "demand" for confidence differs across models, but they share the implication that individuals will actively try to subvert negative feedback, and can end up persistently overconfident. The models also make various assumptions about how individuals are able to "supply" confident beliefs for themselves. For example, some models assume that individuals can use a technology of biased memory to selectively distort memories of past feedback (Bénabou and Tirole, 2002; Compte and Postlewaite, 2004; Gottlieb, 2011); basing predictions on overly positive memories, future selves will be overconfident.

This paper seeks to establish: (1) whether there is persistent overconfidence about relative task performance in an important field setting – a workplace in which managers compete regularly for performance bonuses while receiving detailed feedback; (2) whether these managers have overly-positive memories about past workplace performance; and, (3) whether overly-positive memories are associated with making overconfident predictions.[2] To our knowledge this paper provides the first evidence from

[1] See also the literature on "motivated reasoning," e.g., Kunda (1990) and Sharot et al. (2011).

[2] The literature has used the term overconfidence in different ways. We follow previous studies on task performance – lab studies where students do tasks, or field studies on various types of actors (CEO's, poker players, etc.) – and use the term to describe biased beliefs about relative task performance. Task performance can depend on both internal factors, like own ability, and external factors, like the abilities of competitors or favorability of the environment; we do not distinguish whether the overconfidence comes from biased beliefs about internal or external factors or both. Our paper is about *relative* performance so

the field on whether persistent overconfidence about the future goes hand-in-hand with biased memories of the past. This is in line with explanations for overconfidence based on motivated beliefs. As we discuss in the literature review, the currently formalized models describe the effect of memories on beliefs as causal, but there are other interpretations of motivated beliefs that could generate the same relationship with different assumptions. Findings regarding (1)-(3) are supported through five main sets of results. In the discussion at the end of the paper we provide asixth, exploratory analysis on how manager overconfidence is related to future performance and management style.

As we discuss in more detail in Section 2, the study involved approximately 230 managers, each of whom runs a separate store. The managers compete repeatedly in a high-powered tournament incentive scheme, with detailed feedback, and many managers have observed a large number of tournament outcomes. One source of data is the historical records of the firm on each manager's tournament outcomes. The other is a lab-in-the-field study. This study elicited manager predictions about relative performance in the upcoming tournament at their job, for Q4 of 2015, and also elicited memories about performance in a previous tournament, in Q2 of 2015, for which results had been provided approximately two months earlier.

Our data allow us to address some of the challenges that studies face when trying to establish overconfidence. One key issue for assessing whether predictions are reasonable is the need to take into account what information individuals had access to when making predictions. For example, predictions may "appear" overconfident, but in fact be fully Bayesian if one takes into account the signals that informed these beliefs (see Benoit and Dubra, 2011; Benoit et al., 2015). Oftentimes, researchers do not have any information about past signals, but in our setting, a key type of signal, past tournament outcomes, is public and observable. We can thus check directly whether manager predictions are explainable by past public signals. We also assess in various ways whether the results could be explained by managers having access to additional, private signals.

In Section 3 we take a simple, reduced form approach to address points (1)-(3). Our first set of results is that managers are overconfident relative to a range of different reduced-form predictors one could form using past public signals. This overconfidence is similarly prevalent among managers with substantial experience, so overconfidence is persistent in the face of feedback. This latter result casts doubt on an explanation based on managers having private signals about their types; if such signals are informative, and managers are Bayesian, overconfidence should disappear with experience. Turning to our second set of results, we find evidence consistent with managers being motivated to have positive memories of past performance. Specifically, top-performing managers

in the terminology of Moore and Healy (2008) we study "overplacement", as opposed to overconfidence about absolute performance, i.e., "overestimation", or precision of knowledge, i.e., "overprecision."

are quite accurate in recalling their good performances. Managers with worse past performances, however, have substantial recall errors, and these are strongly skewed towards overly-positive memories. Our third and final set of results in Section 3 shows that it is those managers who have overly-positive memories of past signals who are particularly likely to make overconfident predictions about future performance.

The second portion of the analysis, in Section 4, approaches the same issues but takes a structural modeling approach. Adding structure allows formulating different potential mechanisms, and exploring whether these help the model to match the data. The analysis starts by estimating the parameters of a structural model of Bayesian belief formation, which generates predictions that managers would make if they were fully Bayesian. Our fourth set of results mirror the results in Section 3: Managers are overconfident, in this case compared to an explicitly formulated Bayesian benchmark model. This is also true in various robustness checks about model assumptions. The model also highlights that Bayesian managers should learn relatively quickly in our setting, making persistent overconfidence hard to explain. In the same structure it is possible to explicitly model the possibility of private signals received while on the job, and estimate the private signal structure that brings the model as close to the data as possible. It turns out that the resulting private signal structure has some implausible features, and the model is still far from the data, providing another indication that private signals do not drive the results. For our fifth set of results, the structural model is augmented with biased memory, calibrated using the data on manager recall. The model allows for heterogeneity in the extent to which memories are distorted, as well as the degree of self-awareness among those who do distort. We cannot reject that the model matches the data, at conventional significance levels, and the model is substantially closer to the data than other versions of our structural model.

Taken together, our findings are consistent with explanations for persistent overconfidence provided by models of motivated beliefs. Section 5 discusses implications of the results, and also provides some exploratory analysis on differences in job performance and management styles associated with being overconfident. The rest of the introduction reviews, in turn, the related empirical and theoretical literatures.

1.1 Empirical literature on overconfidence and biased memory

This paper is complementary to a literature on overconfidence in the laboratory. Many studies have measured apparently overconfident behavior (e.g., Camerer and Lovallo, 1990), and some have measured beliefs using designs that can rule out any Bayesian explanation (Merkle and Weber, 2011; Burks et al., 2013; Benoit et al., 2015). There is less lab evidence on motivated beliefs and overconfidence, but Eil and Rao (2011) find that individuals adjust beliefs more in response to good than bad information, immediately after it is received, and Schwardmann and van der Weel (2018) find that subjects become more overconfident when they have a strategic need to impress others.[3] On biased memory, Chew et al. (2018) show evidence that students can have falsely positive memories of performance on a cognitive ability test, and Zimmermann (2018) sheds light on the dynamics of biased memory, showing that memories are accurate immediately after feedback but are biased one month afterwards. The lab evidence has clear strengths in terms of control. Our study is complementary by providing evidence that the mechanisms of motivated beliefs and biased memory are empirically relevant in more naturalistic settings; managers have biased memories and overconfident beliefs about their real workplace performances, which are central to their livelihood, and on which they have extensive past feedback.

Another related literature documents behaviors and beliefs in the field that are consistent with overconfidence. Malmendier and Tate (2005 and 2015) identify a cluster of behaviors among CEO's that are consistent with a model of overconfident decision making. DellaVigna and Malmendier (2006) show that individuals choose gym contracts in ways that are consistent with naive hyperbolic discounting, which is a type of overconfidence about future patience. Oster et al. (2013) provide evidence that individuals deliberately avoid opportunities to take a test that is diagnostic of having a deadly disease, and subsequently engage in behaviors, and have beliefs, that are indicative of overconfidence about mortality chances. Our paper adds the first field evidence linking overconfident beliefs to biased memory, and provides evidence for a mechanism that can potentially sustain overconfidence even in settings where feedback is hard to avoid.

A smaller set of previous studies have measured individuals' beliefs about task performance in field settings. This includes Park and Santos-Pinto (2010), who find that chess players are optimistic in predicting their relative performance in a chess tournament even though they have a very informative measure of their own and of their opponents' past performance. Another study, Hoffman and Burks (2017), finds that truckers are persistently overconfident in predicting the absolute number of miles they will drive each week. Besides studying managers, and taking different approaches to address alternative explanations such as private information, our paper is distinct because it sheds light on a particular explanation for overconfidence based on motivated beliefs and biased memory.

[3]See also Mobius et al. (2011), Charness et al. (2013), and Hoffman (2016). There is also evidence for motivated beliefs in the domain of prosociality, with individuals desiring to believe that they are a prosocial person (see Haisley et al., 2010; Gneezy et al., 2015; Di Tella et al., 2015).

1.2 Theoretical literature on overconfidence and biased memory

The paper is also relevant for a theoretical literature on motivated beliefs, which assumes that individuals have a "demand" for being confident. The exact source of motivation differs across models (for a survey see Bénabou, 2015). In some, individuals get a direct utility benefit from more positive beliefs (e.g., Kőszegi, 2006, 2006; Brunnermeier and Parker, 2005, Sarver, 2018; Bracha and Brown, 2012). Others allow for confidence to have instrumental value (Compte and Postlewaite, 2004; Bénabou and Tirole, 2002). For example, in Bénabou and Tirole (2002) individuals have self-control problems, and can use inflated confidence to induce future selves to work harder when effort and ability are complements.[4] Overconfidence also has costs in these models, which may be psychological (e.g., Sarver, 2018) or financial (e.g., Bénabou and Tirole, 2002; Brunnermeier and Parker, 2005).

Models of motivated beliefs also share the feature that individuals have some way of "supplying" distorted beliefs, albeit typically subject to some "reality constraints" that limit how far individuals want to, or are able to, distort beliefs away from the truth (Bénabou and Tirole, 2002, discuss various reasons for such constraints). In some models individuals directly choose beliefs about themselves or future outcomes (e.g., Brunnermeier and Parker, 2005). Others suppose that individuals can indirectly preserve overconfidence, by taking steps to limit exposure to negative feedback (e.g., Carillo and Mariotti, 2000; Kőszegi, 2006).[5] Another strand of the literature assumes that individuals can subvert negative feedback even if exposure is unavoidable, using a technology of memory distortion; they can to some extent (and at a cost) distort memories of past signals in an overly-positive direction, thereby fostering overconfidence in future selves (Bénabou and Tirole, 2002; Compte and Postlewaite, 2004; Gottlieb, 2011).[6]

A signature prediction of models of motivated beliefs is that overconfidence about the future may be associated with biased memory of the past, although the precise mechanism and nature of causality can differ. In any model maintaining the assumption that individuals form predictions at least partly based on signals observed in the past, biased memory can have a causal impact on overconfidence. An example is Bénabou and Tirole (2002), in which individuals are able to use biased memory of signals to cause overconfidence, because future selves base predictions on memories of the past. If positive memories themselves have a consumption value, in the spirit of models in

[4]Another type of strategic motivation is social signaling; it may be easier to convince others that one has high ability, if one believes this as well (Bénabou and Tirole, 2002; Burks et al., 2013; Schwardmann and van der Weel, 2018).

[5] Zabojnik (2004) provide a rational model of overconfidence which features information avoidance.

[6]Other technologies considered in the literature include self-signaling, i.e., taking actions or stating beliefs that signal confidence to future selves (Quattrone and Tversky, 1984; Bénabou and Tirole, 2004; Mijovic-Prelec and Prelec, 2010; Bénabou and Tirole, 2011).

which beliefs enter utility directly (e.g., Kőszegi, 2006), this provides an independent reason to have biased memories, but there is still a causal impact on overconfidence as long as predictions are based on memories. Other interpretations of the idea of motivated beliefs, although not currently formalized, could potentially generate a similar correlation for different reasons. For example, it is conceivable that individuals first choose optimistic beliefs, and then distort memories of the past to match these beliefs (perhaps to minimize "cognitive dissonance," Akerlof and Dickens, 1982). Although causality would be reversed, the idea is quite similar to existing models, as memories are biased in order to justify overconfidence. If individuals have the ability to freely choose beliefs, about the future and about the past, and do so because of the same underlying demand for positive beliefs, the relationship of biased memory and overconfidence could be correlational rather than causal.

In summary, a link between biased memory and overconfidence is consistent with a range of existing models of motivated beliefs, or natural extensions thereof. We do not directly assess causality in our data, but we do show that a structural model in which biased memory causes overconfidence via Bayesian updating can various features of the data quite well. Regardless of the causality between memory and overconfidence, a motivated beliefs explanation for overconfidence has important implications, which we discuss further at the end of the paper.

There are other explanations for persistent overconfidence, which do not involve motivated beliefs, and the findings in this paper do not rule out that such mechanisms may also contribute to manager overconfidence. For example, in environments where performance depends on bundles of skills, individuals might have different beliefs about the relative importance of these skills, and base predictions on whichever is their best skill (Santos-Pinto and Sobel, 2005; see also Van den Steen, 2004).[7] Another possibility is some type of cognitive mistake in updating, for example, if individuals have mistaken views about the noisiness of signals (e.g., Hoffman and Burks, 2017). Another example of an explanation based on misperception is the Dunning-Kreuger effect discussed in social psychology, in which low ability people fail to understand their incompetence (Kreuger and Dunning, 1999). As the focus of these types of models is on explaining how beliefs can be mistaken despite knowledge of true signals at the time of forming predictions, they are not designed to speak to biased memory about signals. Thus, while evidence of a link between overconfidence and biased memory will not rule out the mechanisms featured in these models, it will suggests that they are not the full story.

[7]Another explanation based on priors is that individuals have priors that put zero probability on having low ability; this leads to persistent overconfidence, since no amount of signals can cause a Bayesian to update a prior of zero to a positive probability (Heidhues et al., 2018; see also Hestermann and Yaouanq, 2019.

2 Work setting and datasets

2.1 Nature of the work setting and the incentive scheme

The subjects of the study are managers working for a chain of food and beverage stores in a developed country. Each manager is in charge of a separate store, and makes a range of important decisions: the number of workers to employ, task allocation, and how many and which types of products to sell. A typical store has roughly fourteen employees including one or more assistant managers. The manager receives a base salary, but can also earn substantial performance bonuses, awarded through a tournament conducted each quarter.

To determine the quarterly performance bonus the firm first ranks the managers on each of several dimensions of performance: (1) store profits relative to a target determined by store characteristics, (2) sales growth, (3) a customer service rating by an undercover "mystery shopper," and (4) a review of the store manager by a regional manager against centrally set criteria.[8] A manager's rank on a given dimension puts him or her into one of several bands, with each band being assigned a score. The scores increase approximately linearly going from the worst to best band.

The firm then creates an overall relative performance indicator by multiplying the scores from the different dimensions together, and ranking managers according to their overall score. The amount of the bonus is calculated by multiplying the overall scores by 30% of the base salary. It thus rises continuously with rank and all ranks receive a prize. There are also some additional prizes, awarded to roughly the top 25 performers, in the form of increased scores; this leads to higher bonuses and implies some convexity at the top of the scheme. For most of these the score is increased by 50%, but the top one or two performers may receive an increase of 100% up to 200%.[9]

Figure A1 in the appendix shows the shape of the incentive scheme. Managers get a substantial portion of earnings from the scheme: The median bonus is equal to about 22% of the base quarterly salary. The strength of incentives, in terms of the prize spread, is also substantial. The median bonus for the top quintile of performance (5 is best) is about 36% of the base quarterly salary, compared to only about 13% for the bottom quintile. A more "local" measure of the strength of incentives is the reduction in earnings

[8]The firm measures store profits relative to targets that are constructed as predicted values from regressions of historical store profits on store characteristics such as region, store age, etc.. Measuring profit relative to targets is intended to purge that particular performance measure of effects that might come from store characteristics rather than manager performance. Customer service ratings are assigned by an undercover evaluator posing as a customer. The review by a senior manager evaluates adherence to, e.g., health and safety rules.

[9]Managers scores are also weighted by the overall performance of their area, with managers in better performing areas getting higher scores. Including or excluding these does not appreciably change the shape of the scheme.

from dropping by 1 quintile. This is 8% of the base quarterly salary going from quintile 5 to 4, and about 4% for each of the quintiles 1 to 4.

Managers receive detailed feedback about their performance every quarter, in the form of learning the complete performance distribution across stores in the tournament and their place in the distribution. After the final tournament results are determined, each manager receives information about their bonus and also a table listing the rankings of managers as well as their absolute performances on the individual dimensions. Managers receive the information and discuss it with senior managers in regularly scheduled meetings after each quarter. Thus, managers do receive the feedback at one point, although they can potentially forget this information later on.

The scope of the tournament has changed in alternating quarters in recent years, but we can construct a comparable performance measure over time. Specifically, in every other quarter since Q2 of 2012 there has been a national tournament that included all of the stores across the country. For the rest of the quarters, the company divided the country into a few large regions, and conducted the tournament separately within each region. Previous to 2012, the tournaments were always regional. We construct a directly comparable (nationwide) performance measure over time, even for quarters with regional tournaments, by using the absolute performance measures for the managers, and ranking these according to the rules of a nationwide tournament. Managers themselves can also make a good inference about national rank in such quarters, even if they do not perform any calculations, by using regional rank as a proxy; the average correlation between regional rank and constructed national rank is 0.70 (Spearman; $p < 0.001$). Our study asks managers to predict the outcome of a national tournament; in case they neglect or discount outcomes from quarters with regional tournaments when making predictions, the analysis checks robustness to including or excluding quarters with regional tournaments.

2.2 Historical performance data

The company has shared its historical data on manager performance going back to Q1 of 2008. The data include overall performance, performance on each of the dimensions that underly the aggregate performance measure, and a few pieces of additional information such as how many assistant managers the manager chooses to hire. The analysis in the paper uses 8.5 years of this performance data, from Q1 of 2008 through Q2 of 2016. Appendix B discusses additional details about the creation of the dataset.

The historical performance data yield some descriptive statistics about the work environment and the managers. The average number of stores active in any given quarter over the sample period is about 230, but the company has grown over time, reaching about 300 stores by the end of the sample period. Managers sometimes switch stores during their tenure. For example, among managers working in Q4 of 2015, which is the quarter in which we elicited manager predictions, roughly 48% have switched stores at least once during their tenure. Median tenure in the current store is 5 quarters, and median total tenure is 10 quarters. An analysis of managers who switch stores indicates that there is a role for manager characteristics in influencing performance, although store characteristics clearly matter as well.[10] Over the sample period the fraction of managers leaving the managerial job is around 6% per quarter, with no significant time trends. The analysis investigates whether manager switching of stores, or manager turnover (attrition), are important for the results.

2.3 Data on manager predictions, memories, and traits

Measurements of manager predictions about future performance, and memories of past performance, were obtained in a lab-in-the-field study conducted with managers in early Q4 of 2015. The researchers attended some of the regularly scheduled meetings in which groups of roughly 8 to 10 store managers meet with a more senior manager. These meetings took place in private rooms in various locations, e.g., in store break rooms.

The study followed a standardized protocol across sessions (meetings). Managers were seated at a table, with dividers to ensure that decisions were made individually, and were not allowed to speak to one another. The study materials were provided in written form, but there was also a verbal summary of the instructions for each part by the attending researcher, following a script, to ensure understanding. Piloting with a few managers before the study made clear that the instructions needed to be very simple and clear, as the managers were not used to participating in such exercises.

To address manager concerns about confidentiality, the researchers conducting the sessions gave their academic affiliations, explained that they were not employed by the firm, and guaranteed that the managers' individual responses would be kept completely confidential from the firm and co-workers. It was also emphasized that funds came from an academic grant and that checks would be mailed directly to the managers' home addresses, by the researchers, early in Q1 of 2016. Thus, no-one in the company would ever learn the managers' individual earnings in the study.

A total of 239 managers participated in the study. About 56% were female, me-

[10]A one standard deviation increase in the mean of a manager's performance at his or her past stores is associated with an increase of about 0.17 standard deviation in performance at a manager's current store. While performance in past stores cannot be caused by characteristics of the current store, there could be a correlation in the types of stores to which a manager is assigned over time. For this reason our regression includes controls for current store characteristics (results available upon request).

dian age was 36, and median tenure at the company was 2.5 years. Managers received a participation payment of $20, and on average earned another $20 in incentive payments from the study. There were 32 sessions, with the earliest taking place on October 22nd, 2015 and the latest taking place on December 7th, 2015. Of the 32 sessions, 22 took place in October. The timing of sessions varied how long ago managers had seen the tournament results they were asked to remember, and how far in the future were the tournament results they were asked to predict. The analysis therefore investigates whether the timing of sessions is related to the accuracy of manager memories and quality of manager predictions.

2.3.1 Measuring manager's predictions of future performance

One part of the study elicited managers' predictions for how they would rank in the upcoming (nationwide) tournament for Q4 of 2015. Managers were asked to guess whether they would be in the top 20%, the second 20%, the third 20%, the fourth 20%, or the lowest 20% of the tournament ranking. In other words, managers were asked for their modal quintile, based on their beliefs about the probabilities of different quintiles. The study provided an incentive to guess correctly: about $22 for getting it right. The managers knew that researchers would check the outcomes of the tournament, once they were available, and then mail payments in Q1 of 2016.

Several features of the design were intended to help minimize measurement error. One potential source of error could be inattention: due to cognitive costs, managers might state a prediction that is different from what they would come up with if they thought more carefully. The incentives provided in the study, however, were intended to encourage managers to think carefully. Furthermore, managers arguably already had substantial incentives to be thinking carefully about the tournament, due to the large amount of money tied up in the performance bonus scheme. Another source of measurement error would be if managers stated overconfident predictions to impress others, rather than stating their true beliefs. The confidentiality protocol, however, should have minimized motives related to impressing superiors or co-workers. Furthermore, it is not clear that making overconfident predictions is impressive; it could be taken as a sign of incompetence. Note that it is unlikely that managers would have developed a habit of routinely boasting about performance in the workplace, since tournament outcomes are public and verifiable by co-workers and superiors. This leaves the possibility that managers wanted to impress the researchers with false statements, but providing incentives for correct guesses is the standard remedy in experimental economics for mitigating such motives.[11] Furthermore, it is not clear that managers would expect

[11]The level of incentives provided was substantial relative to those that are typical for one-shot decision

researchers to be impressed if they state overconfident beliefs that are subsequently checked and verified to be wrong.

A different type of measurement error could arise because the study did not elicit complete probability distributions from managers, i.e., the likelihoods that they attached to ending up in each of the five quintiles. This was dictated by the need to keep the elicitation as simple and naturalistic as possible. Piloting suggested that more complex approaches, and the relatively complex rules needed to make responses incentive compatible, would not be well understood. The key benefit of this approach is we are confident that the managers understood what they were being asked. One potential downside of this elicitation approach is that risk averse individuals might want insure themselves against poor performance on the job, by placing their bet on a low quintile. Any such hedging (i.e. insurance) motives, however, would work against finding overconfidence. Thus, if managers engage in hedging, this source of measurement error makes findings of overconfidence a lower bound. We also check whether predictions are related to a measure of manager risk aversion, and find no statistically significant relationship, which casts doubt on an insurance motive.

2.3.2 Measuring manager's memories of Q2 tournament rank

Another part of the study asked managers to recall how they performed in the most recent (nationwide) tournament, which was Q2 of 2015. Managers had learned the results of this tournament roughly two months earlier. Specifically, managers were asked to recall their rank in the tournament, and offered a payment of \$1.50 for being within +/- 10 ranks of their true past rank. The incentives provided for recall were smaller than in the prediction task, because recalling a number that they had learned is arguably easier than predicting the future. The instructions provided the header row of the tournament outcome table from Q2, and circled the relevant column header, to maximize clarity about what was being asked. Managers had to answer the question on the spot, and could not talk to each other or use their phones to look it up, so the question was a test of their memory.

The design was intended to mitigate possible measurement error regarding manager reports of memory. To be clear, inaccurate memory is not measurement error, but rather the object of study; measurement error would be if managers stated something different from what they actually remembered. One source of such error could be inattention, such that managers do not think carefully about what they actually remember, and state something different. This should have been mitigated by the provision of incentives for correct recall, however, and also arguably by the fact that workplace in-

in lab experiments, although managers do have higher incomes than typical student subjects.

centives are tied to the performance indicator they were asked to recall. Another source of measurement error would have been if managers thought that stating biased recollections would impress others, but confidentiality and financial incentives for correct recall worked against motives to impress co-workers or the researchers. Also, it is unclear that being inaccurate in recall is something that is viewed as impressive. Unlike for the prediction measure, there was no hedging motive for the memory measure, as it was retrospective rather than prospective.

2.3.3 Measures of other manager traits

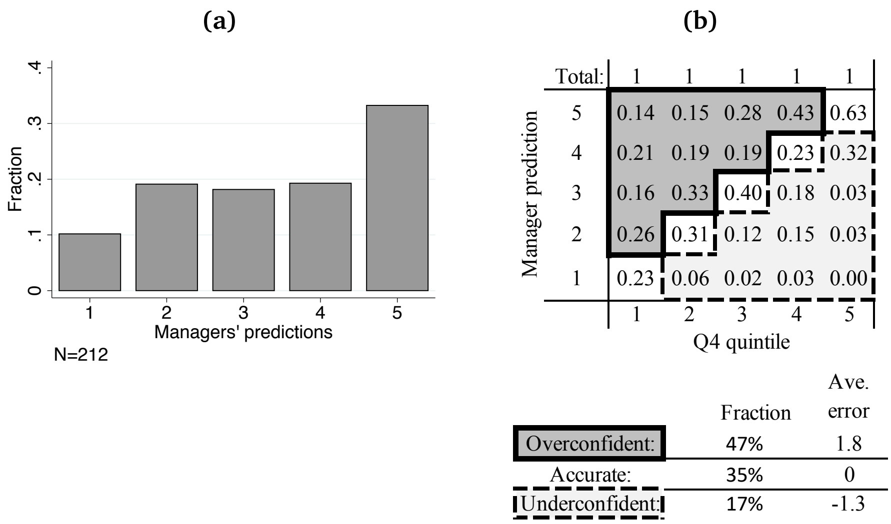
The study also measured some other manager traits in case these might be related to overconfidence: Gender, as previous studies have found gender differences in overconfidence for some types of tasks (e.g., Niederle and Vesterlund, 2007); experience at the company (tenure), to allow investigating whether greater exposure to feedback might be related to accuracy of predictions; manager age, in case greater life experience is related to reduced overconfidence. These traits are featured in the main analysis as control variables. The study also included measures of other types traits and attitudes: incentivized measures of willingness to mis-report information, mathematical ability, and risk aversion; and self-assessments of willingness to take risks, willingness to compete, relative confidence, and patience (more details on the measures are provided in Appendix C). We show in robustness checks that controlling for these does not change our results. The study also included an experiment designed to measure one aspect of an overconfident management style; this is used as an outcome variable in the exploratory analysis on management style discussed at the end of the paper.

3 Reduced form analysis

3.1 Descriptives on manager predictions and empirical strategy for identifying overconfidence

As a first look at the data, Panel (a) of Figure 1 shows the distribution of manager predictions. The most salient feature is the skew towards predicting higher quintiles (throughout the paper we order quintiles such that 5 is the best). Only about 10 percent predict achieving the worst quintile, roughly 20 percent predict each one of the intermediate quintiles, and 33 percent predict achieving the top quintile. A comparison to realized outcomes in Q4, shown in Panel (b), reveals that managers do have insights into predicting future performance: Achieved outcome and prediction are significantly positively correlated, 0.47 (Spearman; $p < 0.001$). On the other hand, Panel (b) shows that many managers make prediction errors, and these errors are highly asymmetric: 47 percent of managers bet on a higher quintile than their realized quintile, versus less than half as many, 17 percent, betting on a worse quintile. This bias in manager forecasts is statistically significant ($p < 0.001$).[12] In terms of magnitudes, prediction errors are substantial, and are larger in the overconfident direction: 1.8 quintiles for errors in the overconfident direction, versus 1.3 quintiles in the underconfident direction.

Figure 1: Distribution of manager predictions and manager predictions compared to Q4 outcomes



These results are suggestive of overconfidence, but there are some important limitations of using Q4 outcomes as a benchmark for assessing overconfidence bias. Simply because a manager's guess deviates from the outcome does not mean it was incorrect guess — it could have been correct given the information the manager had ex ante, and if managers base guesses on noisy signals, errors are to be expected. Moreover, there is no reason to believe such prediction errors must be symmetric for a given quarter.

Indeed, a recent paper by Benoit and Dubra (2011) points out in detail how asymmetric prediction errors in the direction of overconfidence can be Bayesian. Specifically, if individuals are learning about their type (type meaning some fixed, performance relevant characteristic) by observing signals and updating in a Bayesian fashion, then certain classes of information structures can lead to asymmetric prediction errors, although the errors go to zero as more signals are drawn. Intuitively, the necessary information structures involve relatively rare, highly informative negative signals, and relatively common, weakly positive signals. The prevalence of weakly positive signals can lead most people to think, at least initially, they are slightly more likely to be a very good type, than any worse type. Benoit and Dubra (2011) do not make claims

[12]This result is from an OLS regression of the prediction error on a constant term.

that such information structures are generally plausible, but they point out the importance of taking into account the past signals that individuals have seen, for identifying overconfidence bias.

To address these issues, we turn to an empirical strategy that uses a more satisfactory benchmark for assessing overconfidence: what a manager "should" have predicted given their histories of past tournament outcomes. This exploits a key advantage of the setting, that a set of key signals observed by managers – past tournament outcomes – are publicly observable and recorded in the historical performance data.[13]

A pre-requisite for this to be a viable strategy is that tournament outcomes be informative. Table 1 shows the frequencies of managers ending up in different tournament quintiles in quarter t conditional on quintile in $t-1$. We denote this transition matrix $\hat{Z}$. The transition probabilities show that the quintile outcome in any given quarter $t-1$ is in fact predictive of the quintile outcome in quarter t: The modal outcome is for that same quintile to occur in the next quarter. Another take-away is that quintiles 1 and 5 are particularly informative; the probability of repeating a 1 or a 5 is roughly 40%, compared to a probability of 25-30% for an intermediate signal.[14]

Table 1: Quintile-to-quintile transition matrix $\hat{Z}$

	Fractions of managers				
	Quintile in t				
Quintile in $t-1$:	1	2	3	4	5
5	0.05	0.11	0.15	0.26	0.43
4	0.11	0.17	0.21	0.27	0.24
3	0.17	0.23	0.27	0.20	0.13
2	0.24	0.26	0.23	0.17	0.10
1	0.43	0.23	0.20	0.09	0.05
N:	961	1,018	1,034	1,007	962

Notes: The rows show the average proportions of managers achieving different quintile outcomes in the national tournament ranking for quarter t conditional on a given quintile outcome in quarter $t-1$, using all quarters from Q1 of 2008 to Q4 of 2015. The number of observations differs across quintiles in $t-1$ due to attrition and opening of new stores.

[13] Burks et al. (2013) propose a test that can reject a Bayesian explanation for overconfidence, even without any data on past signals, if the observed overconfidence relative to realized outcomes is sufficiently extreme. Conducting this test on our data, we cannot statistically reject the Bayesian model at conventional significance levels. Our data also fail the bounds proposed by Benoit and Dubra (2011). Thus, we leverage the fact that we do observe data on past signals.

[14] This feature of the information structure could be consistent with a normal-shaped distribution of underlying manager ability: The mass of managers in the middle would have relatively similar abilities, choose similar effort levels, and thus have tournament outcomes that are largely random; managers in the tails would be quite different from everyone else in terms of ability and thus consistently have the worst or best outcomes.

Our next step is to investigate whether the managers' apparently overconfident predictions might in fact be consistent with them employing some form of prediction model that uses past tournament outcomes. A challenge is that there are different ways managers might form predictions. Our approach is to consider a wide range of alternatives. We consider different classes that vary in terms of complexity: simple rules of thumb, more complicated, regression-based prediction models, and in Section 4, a structural model. Within each class we also consider alternative assumptions about: (1) the types of initial priors that managers might combine with tournament outcomes to form beliefs; (2) differential weights that managers might apply to recent signals versus signals from further in the past; (3) differential weights on signals from particular periods, due to potentially imperfect manager knowledge about these signals. While the selection of predictors is partly guided by our intuition, we also explore a more data-driven approach to identifying good predictors, using standard model-selection techniques.

3.2 Testing for overconfidence

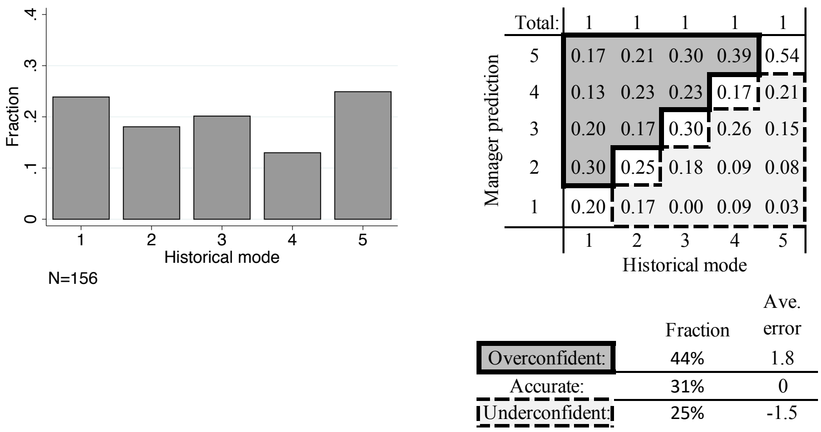
We begin by comparing manager predictions about their modal quintile to an arbitrary but also seemingly natural rule of thumb that a manager could have used: A manager's most frequent quintile in the past. Calculating each manager's modal quintile over past quarters, and dropping managers who do not have a unique mode, yields the distribution of historical modes shown in Panel (a) of Figure 2. The distribution is slightly u-shaped, with the highest masses for modes of 1 and 5. This is understandable given that extreme outcomes are more persistent (Table 1).[15] Panel (b) shows that many managers make predictions that differ from the mode, and the prediction errors tend to be in the direction of overconfidence: 43% of managers predict a higher quintile for Q4 of 2015 than their historical modal quintile, compared to 25% predicting a lower quintile. The average of prediction errors in the overconfident direction is also relatively larger, 1.8 quintiles versus 1.5 quintiles.

Conclusions are similar exploring rules of thumb based on a range of alternative assumptions; Table D1 in the appendix summarizes the results. For example, suppose one assumes that managers start the job with (rationally) overconfident priors. This could lead them to be overconfident initially relative to predictors based on tournament outcomes. If managers are Bayesian, however, this overconfidence should dissipate as

[15]Given infinite signals, the distribution should converge to a uniform, but in a finite sample, there can be a u-shape. To see this, suppose there are 5 types of managers. The worst and best types are quite likely to have an outcome of 1 or 5, respectively, and never get an outcome of 3. The remaining types have a more uniform probability of getting outcomes 2, 3, and 4, but also non-zero probabilities of getting 1 and 5. In a finite sample, due to chance, some intermediate types could have modes of 1 and 5, but no high or low types will have modes of 3. Intermediate types are also more likely to have a non-unique mode, and be excluded, which also contributes to fewer modes for middle quintiles.

Figure 2: Distribution of historical modes, and manager predictions compared to historical modes

(a) **(b)**



Notes: Predictions are in terms of quintiles of Q4 performance, with 5 being the best. Prediction errors are also in terms of quintiles.

more feedback arrives. We investigate whether overconfidence relative to the mode predictor goes away when focusing on managers who are relatively experienced, i.e., with more than 2 years of tenure. It turns out that the prevalence of overconfidence relative to the mode is essentially unchanged when we focus on this sub-sample (see Table D1). The 2 year cut-off is arbitrary, but later in the analysis we confirm that various measures of overconfidence are not significantly related to a continuous measure of experience.

An assumption entailed in the historical mode predictor is an equal weighting of tournament outcomes from all periods, but this could be incorrect if recent signals are more informative due to some type of non-stationarity within managers. For example, if managers learn by doing initially, then signals from early in their tenure might be less informative. Likewise, if store characteristics matter for outcomes, then signals from past stores might be less informative about outcomes in the current store. We find that managers are similarly overconfident, however, compared to the mode calculated without signals from early in a manager's career, or from previous stores.

Another reason why recent signals might be more informative is due to non-stationarity in the environment, e.g., if patterns of turnover and hiring lead to a changing distribution of manager abilities at the company. We find that results are similar, however, calculating the mode using only recent outcomes, Q3, Q2, and Q1 of 2015. This suggests that non-stationarity does not drive the results. Furthermore if the environment were non-stationary, one would also expect that correlation structure of quintile outcomes from one quarter to the next would be changing over time. For example, if turnover leads to a more homogeneous ability distribution over time, then there is a larger random component to winning the tournament, and tournament outcomes would become less informative over time. Instead, as shown in Appendix E, if one calculates a transition matrix Z_t for each pair of quarters, there is no evidence that the elements of Z_t are changing over time. The data are thus consistent with the company being in a steady state, such that the manager types leaving the company are replaced by managers with similar types.[16]

Other robustness checks investigate predictors that allow for the possibility that managers might not perfectly observe certain signals. In particular, managers who were interviewed early in Q4 of 2015 might not have known the outcome of the Q3 tournament when making their predictions, since the results became available partway into Q4. Managers might also not have a full understanding of how to construct national rank in quarters with regional tournaments. Calculating the mode without Q3, however, or using only signals from quarters with national tournaments, yields similar results. Finally, Table D1 also shows that the results are robust to different approaches to including managers with non-unique modes.

Rules of thumb might be an overly simplistic description of how managers form predictions, and we have selected these arbitrarily, so we turn to more complex regression models, and use a data-driven process to select the best performing model. The focus is on panel regression models that predict future performance based on lagged measures of past performance. For a given model we use multinomial logit estimation to generate predicted probabilities of each quintile ranking in Q4 of 2015 for each manager, and select the quintile with highest probability as the prediction. Within this class of models, two specific questions we consider are: What is the optimal number of lagged performance outcomes for maximizing predictive power; should performance outcomes in a given past quarter be measured linearly in terms of percentile of performance, or non-parametrically with separate indicators for each quintile of performance?

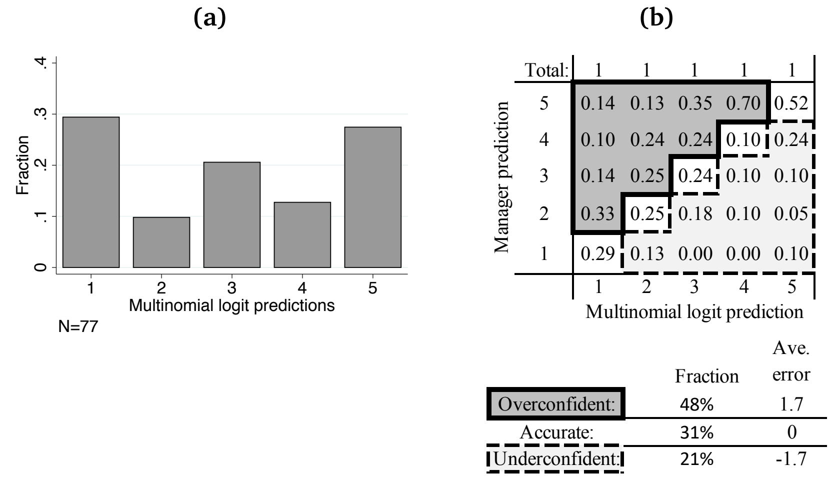
It turns out that using a substantial number of lags (8 lags), and using the linear specification with percentile of performance for each past quarter, delivers the best predictive power in our model selection exercise. The exercise was based on cross-validation, a simple machine learning technique that tests predictive power using randomly selected "hold-out" samples (for details see Appendix F). The resulting prediction model can be written:

$$q_{i,t} = \alpha + \sum_{j=t-1}^{t-9} \beta_j y_{i,j} + \epsilon_{i,t} \tag{1}$$

[16]Suppose the company loses 3 bad managers and 2 good managers each quarter, reflecting greater turnover among bad as opposed to good managers. The environment will be stationary if the company attracts 5 new managers, and 3 of these are bad and 2 are good. Note that in equilibrium, if all companies try to get rid of worse employees, the pool of unemployed will be skewed towards worse types.

Where the dependent variable $q_{i,t}$, is performance quintile for manager i in quarter t, and independent variables are performance outcomes in earlier quarters, $y_{i,j}$, $j \in (t-1, ..., t-9)$. It is not surprising that the model does best when it includes a large number of lags, as this entails estimation on a sample of (relatively experienced) managers, for whom we have a large number of signals and thus better precision in assessing individual manager types.[17] The robustness checks include estimating models with fewer lags, and also using less parametric specifications.

Figure 3: Distribution of multinomial logit predictions, and manager predictions compared to model predictions



Notes: Predictions are in terms of quintiles of Q4 performance, with 5 being the best. Prediction errors are also in terms of quintiles.

Panel (a) of Figure 3 shows that the distribution of predicted quintiles from the regression model is u-shaped, like the distribution of the historical mode predictor. Panel (b) shows that many managers made predictions that are substantially different from the predictions of the model, and the prediction errors tend to be in the overconfident direction: 48% of managers bet on a higher quintile than the model says was most likely for them, versus 21% betting on a worse quintile. In this case the average size of the prediction errors is the same for overconfident and underconfident directions, 1.7 quintiles. The estimated coefficients of the baseline model are reported in Columns (1) to (4) of Table F2 in the appendix.[18]

The model predictions are of course subject to error, because we have to estimate the model parameters from a sample of tournament outcomes, which have random

[17]Better performance of the linear specification can be due to the fact that it provides a finer grained measure of performance, compared to the less-parametric but also coarser specification using quintile dummies.

[18]Most of the individual lag coefficients are not statistically significant individually, but this reflects correlated performance over time for managers. The coefficients are highly significant in a joint test (χ^2; $p < 0.001$) and fit is improved by including all of the lags.

component. This raises the question whether manager predictions might differ from our model predictions not due to sub-optimal predictions on their part, but due to estimation error on our part. For example, suppose that managers use the same model that we use, and are fully informed about the true parameters of the model; our model prediction could differ from the manager prediction because our model is imperfect and suffers from estimation error. We use bootstrapping to assess the extent of the error in our model, and check whether the difference between manager and model predictions lies within the bounds of this error.[19] We generate 100 new samples, by drawing with replacement from the original sample (recall that the unit of observation in the sample is a manager-quarter pair). The resulting bootstrap samples have different realizations of tournament outcomes, which respect the empirical frequencies in the original sample. We re-estimate the model using each bootstrapped sample, and generate predictions of the modal quintile for each manager using the re-estimated coefficients. For a given bootstrap, we calculate the distance of each manager's bootstrapped prediction from the prediction of the model based on the original sample, using the Euclidean distance metric.[20] Summing up these distances across all managers gives a total (Euclidian) distance between a given bootstrapped distribution of bets and the distribution of bets from the original model.

Given the 100 bootstraps, we have a distribution of 100 distances, which gives bounds on the sensitivity of our model predictions to sampling error. We also calculate the distance of observed manager predictions from the original model predictions. This latter distance lies far in the tail of the bootstrapped distances (beyond the 99th percentile); see Figure F1 in the appendix. Thus, we can reject at the 1-percent level that the difference between the model and manager predictions lies within the bounds of the error in our model.[21] Results go through using non-Euclidean distance metrics or other types of statistical tests; for example, weighting the Euclidian distances by the magnitudes of the prediction errors yields even stronger results.[22]

[19]In contrast to this scenario, if the managers have to estimate the parameters as we do, using the same data, then their predictions should accord with our predictions precisely, and we do not need confidence intervals to reject that the model and manager predictions are the same.

[20]The total Euclidean distance is just a monotonic transformation of the "failure rate:" The fraction of managers who differ from the model. We focus on Euclidean distance because it naturally generalizes to situations where we are computing the distance between vectors that do not all have 0 or 1 entires, something that arises later in our analysis when we simulate some versions of our structural model.

[21]An alternative interpretation could be that we are assessing whether the difference between model and manager predictions is explainable by managers being Bayesian but estimating the model with slightly different data, e.g., due to noisy memory of past signals.

[22]Results are stronger because the bootstraps not only generate fewer prediction errors than managers, but the magnitudes of the errors are smaller. An alternative test of whether the distribution of manager and model predictions are significantly different is a χ-squared test ($p < 0.001$). We prefer the bootstrapping approach to the χ-square test (or alternatives like Kolmogorov-Smirnov) because the latter assume that one of the distributions being tested is independent of the data, but the model predictions obviously

It is also possible to reject at the 1-percent level that the degree of asymmetry in prediction errors towards overconfidence that we observe for the managers lies within the bounds of the asymmetry that could be generated by noise in our model. For each of the 100 bootstraps, we calculate the fraction of bootstrapped predictions that are overconfident relative to the original model minus the fraction of bootstrapped predictions that are underconfident. This yields a distribution of 100 differences. The corresponding difference comparing actual manager predictions to the predictions of the original model is beyond the 99th percentile of the distribution (see Figure F1 in the appendix).

We also conduct robustness checks within the multinomial logit framework. Most of these have a similar logic to the robustness checks for the rule of thumb predictors, e.g., testing whether results are similar using only tournament outcomes from the current store to estimate the model, or using only outcomes from recent quarters. Results are summarized in Table F1 in the appendix, and coefficient estimates are reported in Tables F2 and F3. Other robustness checks, summarized in Table F4, examine the effect of non-parametric specifications. In all cases, roughly 40% or more of managers are overconfident relative to the model predictions, compared to roughly 25% underconfident, and the differences with respect to model predictions are statistically significant.

In summary, manager predictions are systematically overconfident relative to a range of different reduced-form predictors one can construct based on their past tournament outcomes. This provides stronger evidence of overconfidence bias than the initial naive approach of simply comparing predictions to future outcomes.

One remaining concern about the reduced form analysis could be that it assumes past tournament outcomes are the key signal for predicting manager performance, but managers might have access to additional, private signals; to some extent, however, this concern is already addressed by our finding that overconfidence is largely unchanged with experience. One possibility would be if managers observe private signals previous to being hired about their quality as a manager, from the type of signal structure discussed in Benoit and Dubra (2011). This could lead them to enter the firm with *rationally* overconfident priors about their quality, and could explain initial overconfidence relative to reduced form predictors. As discussed above, however, the finding that the prevalence of overconfidence does not dissipate with experience casts doubt on an explanation based on overconfident priors.[23] Another possibility is that managers

depend on the data, and manager predictions are presumably informed by tournament outcomes as well. Also, our bootstrapping procedure provides a test at the individual rather than aggregate level.

[23] This is not to say that managers do not start the job with overconfident priors; indeed, we find that managers with relatively little experience (less than one year) have predictions that are skewed towards higher quintiles, similar to what we observe for the sample as a whole (see Figure G1 in the appendix). This initial overconfidence may not be Bayesian, however, as there appears to be a mechanism that is working against learning.

might draw private signals about their quality while on the job. If such signals are informative, however, Bayesian managers should still learn, and again one would expect overconfidence to dissipate with experience. Notably, it is not just the prevalence of overconfidence that stays constant with experience; additional analysis shows that the magnitudes of the prediction errors, relative to predictors based on tournament outcomes, are also not decreasing with experience (see Figure G2 in the appendix). Note that we would miss some manager learning, if those who do learn their types tend to leave the firm, but this does not alter the fact that those managers who remain should be more accurate than inexperienced managers, if they are Bayesian. Furthermore, as discussed at the end of the paper, we that manager overconfidence is not significantly related to the probability of remaining at the firm. Thus, there does not seem to be much scope for differential attrition on the basis of overconfidence to play a role in explaining our results.

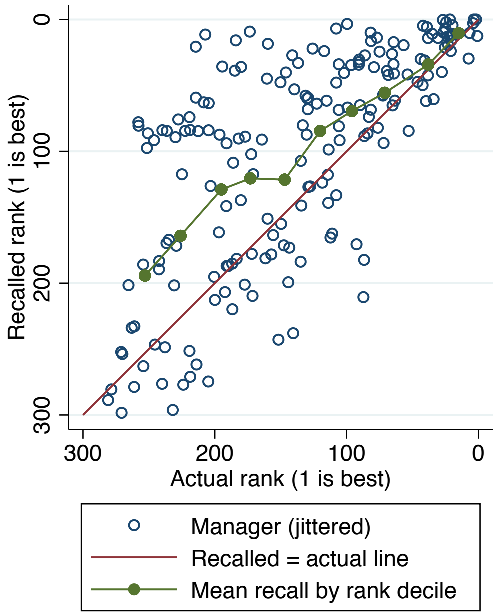
There are other forms of private signals, however, which could generate rationally overconfident predictions even for experienced managers. Specifically, signals about transitory, manager-specific shocks. For example, suppose some managers receive private signals in early Q4, right before making predictions, that their store will do well in the Q4 tournament. This could lead them to rationally predict a higher outcome than a prediction model that only uses past tournament outcomes. Because the signal is about a transitory shock, it is not something a manager could have learned about based on past signals. In the subsequent analysis we therefore address the issue of private signals in several other ways. One way is our investigation of whether there is biased memory, and whether this is linked to overconfident predictions; such findings are not explainable by any of the explanations based on private signals discussed above. Another way is the development of a structural model, in Section 4. In the context of that model, we can verify how quickly a Bayesian manager should learn, and whether two years of experience should be sufficient to cause a substantial reduction in overconfidence. We can also explicitly model the possibility of private signals, with the signals being observed right before managers make predictions, and test whether the best-fitting private signal structure can bring a structural model of Bayesian updating close to the data.

3.3 Testing for biased memory

The evidence so far suggests that managers are persistently overconfident, despite repeated feedback. This section investigates an explanation suggested by models of motivated beliefs, that managers want to be overconfident, and use a technology of imperfect memory to subvert negative feedback. Specifically, managers may tend to have overly-positive memories of past performances, potentially subject to some "reality constraints" that place some limits on how much they deviate from the truth.

Figure 4 displays the raw data from our elicitation of manager memories, about rank in Q2 of 2015, with values jittered slightly to preserve manager confidentiality.[24] The x-axis measures individual managers' actual ranks in Q2, with 1 being the best, and the y-axis shows managers' recalled ranks. Whereas in the rest of the paper higher numbers indicate better performance, in this figure we use smaller numbers for better performance, since the question used to collect the recall data asked about rank.

Figure 4: Recalled performance for 2015Q2, by actual performance



Since models of motivated beliefs posit that managers want to have positive memories of past performances, a first implication is that managers with the very best ranks should have relatively accurate memories. They cannot remember much better than actual rank for mechanical reasons, and remembering worse than actual rank would be counterproductive. Indeed, this is what we observe in Figure 4. The best performing managers do not have substantial recall errors in either direction. Note that this result shows that the managers are capable of having accurate memories.

A second implication of the desire to have positive memories is that inaccuracy in memory should emerge for managers who are below the top of the rank distribution, and these errors should be in the direction of remembering better than actual rank.

[24] Jittering involves adding a small random mean zero perturbation to the values. Without jittering, the firm could in principle use its knowledge of the Q2 ranking to infer individual managers' reported memories from Figure 4.

These predictions are both supported by Figure 4. Going below the best ranks, there is a clear increase in the frequency of managers with inaccurate memories, and the correlation between an indicator for being inaccurate, and rank in Q2 of 2015, is statistically significant (Spearman; $\rho = -0.27$; $p < 0.001$). Furthermore, as shown in the figure, the average recalled rank, by decile of actual rank, is always above average actual rank. Overall, 70 precent of recall errors are "flattering", compared to 30 percent being "unflattering", and recalled rank minus actual rank is significantly different from zero in the overly-positive direction (t-test; $p < 0.001$). Notably, mechanical forces such as regression to the mean, or censoring at the best and worst possible ranks, would imply equally large errors at both the best and worse ranks (in opposite directions).

If manager memories are subject to some "reality constraints", as is typical in models of motivated beliefs, one might expect that recalled rank would be correlated with the truth.[25] Figure 4 shows that, indeed, the average recalled rank (by decile of actual rank) declines with actual rank. The correlation of actual and recalled rank is substantial, 0.71, and statistically significant (Spearman; $p < 0.01$). Thus, manager beliefs are not completely self-serving, but rather are tethered to actual past performance. We also observe variation in the extent of memory distortion for a given actual rank, which could be another indication of a limit on the ability of managers to distort memory, e.g., due to some randomness inherent in the memory distortion technology.[26] Another contributing factor could be individual heterogeneity in manager costs or benefits of memory distortions, such that some managers are more motivated than others to distort the memory of a given rank.

The conclusions from Figure 4 also hold up in regression analysis, which allows addressing some potential omitted variable issues. For the regressions we reverse rank, returning to our practice of having higher numbers indicate better performance. Independent variables are standardized to have a mean of zero and a standard deviation of 1. Coefficients thus show the change in the dependent variable associated with a 1 s.d. increase in an given independent variable.

Column (1) of Table 2 presents results of a Probit regression where the dependent variable equals 1 if a manager has an inaccurate memory and 0 otherwise; the results show that a 1 s.d. increase in Q2 performance is associated with a 0.12 lower probability

[25]As discussed in Bénabou and Tirole (2002), reality constraints could take various forms. One source could be the nature of the memory distortion technology itself. E.g., one way of constructing falsely positive memories might be to focus attention selectively on favorable information, and "rehearse" this information. To the extent that signals are correlated in our context, having a poor tournament outcome could be associated with a manager having less availability of favorable information in general, putting a limit on how positive of a memory can be constructed.

[26]For example, there might be idiosyncratic shocks to the arrival of the types of information that can be used to construct positive memories, leading to variation in memory distortion across managers in a given quarter (and across quarters for a given manager).

Table 2: Inaccurate memory and recall errors as a function of actual Q2 performance

	Inaccurate mem.		Flattering mem.		Unflattering mem.		Recalled - actual perf.	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Performance percentile in Q2 of 2015	-0.12***	-0.12***	-0.07*	-0.11**	-0.06*	-0.01		
	(0.03)	(0.04)	(0.04)	(0.05)	(0.03)	(0.04)		
Inaccurate memory							38.64***	39.09***
							(6.07)	(7.30)
Performance percentile in Q3 of 2015		0.02		0.07		-0.06		0.04
		(0.04)		(0.05)		(0.04)		(5.24)
Mean performance percentile pre- Q2 of 2015		-0.02		0.00		-0.03		-2.43
		(0.03)		(0.04)		(0.03)		(4.45)
Female		0.04		0.12		-0.09		11.64
		(0.06)		(0.08)		(0.07)		(11.00)
Age		0.02		0.01		0.01		-0.24
		(0.03)		(0.05)		(0.04)		(6.89)
Experience		-0.08**		-0.09**		0.01		-5.10
		(0.04)		(0.05)		(0.04)		(7.27)
Constant							0.11	-5.33
							(0.86)	(6.09)
Observations	172	149	172	149	172	149	176	151
Estimation method	Probit	Probit	Probit	Probit	Probit	Probit	OLS	OLS
Adjusted R^2							0.050	0.046
Pseudo R^2	0.092	0.123	0.013	0.061	0.015	0.045		

Notes: Columns (1) to (6) report marginal effects from probit regressions. Columns (7) and (8) report OLS estimates. The dependent variable for Columns (1) and (2) is an indicator for a manager's recalled performance for Q2 of 2015 being different from their actual performance by +/- 10 ranks (the elicitation gave an incentive to be accurate within this range). The dependent variables for Columns (3) to (4), and (5) to (6), are indicators for remembering a better than actual performance in Q2 by more than 10 ranks, or worse than actual by more than 10 ranks, respectively. The dependent variable for Columns (7) and (8) is constructed by taking the difference between recalled rank and actual rank, and multiplying by -1, so that positive numbers indicate recalling a better than actual performance. Independent variables are standardized, so coefficients give the change in the dependent variable (level or probability) associated with a 1 s.d. increase in the independent variable. Performance percentile independent variables are constructed as (recalled) rank expressed as a fraction of the worst rank in the corresponding quarter, and then reversed so that higher numbers reflect better performance. The independent variable inaccurate memory is an indicator for a manager's recalled performance for Q2 of 2015 being different from their actual performance by +/- 10 ranks. Robust standard errors are in parentheses.

of having inaccurate memory. Such a relationship could, however, be endogenous due to omitted variables. For example, some manager trait, e.g., lower cognitive ability, might foster both worse performance and inaccurate memory.[27] This suggests a benefit of controlling for manager ability.[28] Note, however, that such an explanation would not explain why, conditional on being inaccurate, errors are skewed towards being overly positive.

Column (2) of Table 4 shows that being inaccurate is still significantly related to performance in Q2 of 2015, controlling for manager ability by using performance in Q3 of 2015 as well as the mean performance across all pre-Q2 quarters. Thus, it is not good performance in general that is associated with accurate memory of Q2, but rather something special about a good performance in Q2. Controlling for some manager characteristics that could conceivably affect both performance and memory – gender, age, and experience – leaves the results unchanged. Interestingly, managers with more experience have a lower probability of having inaccurate memory, but the relationship is arguably relatively weak, as it takes about 3.7 years (1 s.d.) of additional experience for the probability to drop by 0.09.[29]

Turning to the direction of the recall errors, Columns (3) and (4) show that a worse Q2 performance is associated with a significantly higher probability of remembering better than actual performance, and this is true with controls as well. Columns (5) and (6) show, by contrast, that remembering worse than actual performance has a relatively weaker relationship to Q2 performance, and the relationship is not statistically significant once controls are included.[30] Thus, having a worse Q2 performance is associated with having errors in a particular direction, i.e., remembering better than actual performance. In Columns (7) and (8) the dependent variable is the difference between recalled and actual performance. The regressions include an indicator variable for inaccurate memory. The coefficient estimate for this variable shows that the average recall error is positive and significantly different from zero, and this is also true with the inclusion of the controls.

Robustness checks, reported in Appendix H, explore adding controls for additional factors that might conceivably affect the probability of mis-remembering, or the par-

[27]This would be akin to the Dunning-Kruger effect, in which low ability people make worse predictions about relative performance (see Kruger and Dunning, 1999), but for memory rather than predictions.

[28]A different explanation could be related to the convexity of the incentive scheme; managers who are typically in the worse quintiles of performance might perceive a relatively lower incentive to remember correctly. This would also suggest controlling for manager ability, as a potential difference in perceived incentives to remember would be present for managers who know that they are reliably in lower versus higher quintiles, as opposed to just having a bad quintile in Q2.

[29]The fraction of managers who have inaccurate memories is 0.83 for managers with 2 years or less experience, compared to 0.78 for managers with more than two years of experience.

[30]The imprecision in the estimates means that the difference in coefficients across Columns (4) and (6), for Q2 performance, is not statistically significant ($p < 0.32$).

ticular performance that is remembered. These include the degree to which Q2 performance deviates from a manager's typical (mean or median) pre-Q2 performance, in case this affects memorability, and the variance of manager past performance, in case managers with more variable performances are less likely to remember a given quarter's performance. We also control for the time elapsed between end of Q2 and when memory was elicited, various manager traits, and other summary statistics of past performance. Performance in Q2 remains the key explanatory variable, while these additional factors are by and large not significantly related to manager memories.[31] Taken together, these findings are consistent with an explanation for recall errors about Q2 being based on an underlying motivation to have positive memories about Q2 performance.

3.4 Testing for a Link Between Biased Memory and Overconfident Predictions

We have seen that, in the aggregate, managers have imperfect memory, and memory distortions are in the direction of being overly positive. This suggests that biased memory might be contributing to managers' overly-optimistic predictions about future performance. A sharper test, however, is to check whether there is a link at the individual level between having a more self-flattering memory of the past, and making overconfident predictions about the future.

Table 3 presents regressions that investigate the hypothesized link between biased memory and manager predictions. In Column (1) the dependent variable is manager predictions about the most likely quintile in Q4 of 2015. The estimation method is interval regression, which models the conditional mean of manager predictions while accounting for the fact that the dependent variable is measured in intervals (right and left censored).[32] Independent variables are standardized. Column (1) shows a significant positive relationship between recalled performance from Q2 and predictions about Q4, controlling for actual Q2 performance. Column (2) adds more controls for past performance, and manager traits. The coefficient on recalled performance remains significant, and implies that a 1 s.d. increase in recalled Q2 performance is associated with predicting about 0.5 quintiles higher performance in Q4. The coefficient on actual Q2

[31]One exception is manager experience: Managers who are quite experienced have a tendency to recall worse performances, leading to a modest decrease in the proportion with overly positive memories, and increases in both the proportion with accurate memories, and the proportion with overly negative memories. We discuss a possible interpretation of this pattern in the appendix.

[32]We prefer interval regression over multinomial logit in this case as the goal is to model the conditional mean rather than produce predictions of modal quintiles. Results are similar, however, using multinomial logit: more self-flattering memories of Q2 are associated with a lower probability of predicting low quintiles.

performance is half the size, and not statistically significant, consistent with managers basing predictions mainly on remembered rather than actual past performance.

Table 3: Manager predictions and overconfidence as a function of recalled Q2 performance

	Manager prediction		Overconfident (rel. to historical mode)		Overconfident (rel. to mult. logit)	
	(1)	(2)	(3)	(4)	(5)	(6)
Recalled performance percentile for Q2 of 2015	0.47**	0.43**				
	(0.20)	(0.18)				
Flattering memory about Q2 of 2015			0.18**	0.15*	0.20**	0.20**
			(0.08)	(0.08)	(0.10)	(0.10)
Performance percentile in Q2 of 2015	0.49***	0.22	-0.07	0.01	-0.14***	-0.14**
	(0.18)	(0.16)	(0.04)	(0.05)	(0.05)	(0.06)
Performance percentile in Q3 of 2015		0.62***		0.06		0.00
		(0.15)		(0.04)		(0.06)
Mean performance percentile pre-Q2 of 2015		0.07		-0.19***		-0.11*
		(0.10)		(0.03)		(0.06)
Female		-0.09		-0.04		-0.14
		(0.23)		(0.08)		(0.11)
Age		-0.03		-0.07		-0.00
		(0.12)		(0.05)		(0.07)
Experience		-0.04		-0.01		-0.06
		(0.14)		(0.05)		(0.08)
Constant	2.65***	2.69***				
	(0.23)	(0.26)				
Observations	176	152	128	120	75	75
Estimation method	Int. reg.	Int. reg.	Probit	Probit	Probit	Probit
Pseudo R^2	0.087	0.153	0.044	0.187	0.115	0.152

Notes: Columns (1) and (2) report marginal effects from interval regressions, which correct for the interval nature of the dependent variable (right and left censoring for each interval); the dependent variable is the manager's prediction about Q4 performance quintile. Columns (3) to (6) report marginal effects from probit regressions. The dependent variable for Columns (3) and (4) is an indicator for whether a manager predicted a higher quintile than their historical modal quintile. The dependent variable for Columns (5) and (6) is an indicator for whether a manager predicted a higher quintile than the quintile predicted by the baseline (8 lag) multinomial logit model. Independent variables are standardized, so coefficients give the change in the dependent variable (level or probability) associated with a 1 s.d. increase in the independent variable. Performance percentile independent variables are constructed as (recalled) rank expressed as a fraction of the worst rank in the corresponding quarter, and then reversed so that higher numbers reflect better performance. Robust standard errors are in parentheses.

Columns (3) to (6) of Table 3 use two different indicators for overconfidence as the dependent variables, to check whether having overly positive memories is associated with overconfidence about the future. There is a significantly higher probability of being overconfident, according to both indicators, if a manager has a flattering (overly positive) memory of Q2. The coefficient on manager experience is small and not statistically significant, so the likelihood of overconfidence does not diminish with experience. In case tenure at the firm is endogenous to overconfidence, we checked robustness to excluding experience from the regression, but other coefficients are qualitatively unchanged (at the end of the paper we discuss empirical evidence that suggests tenure is not in fact affected by overconfidence). Interestingly, there is also no evidence of a gender difference in confidence, in line with previous literature showing that such gender differences are found only for some types of tasks.

Appendix I presents robustness checks on whether these results extend to using other types of indicators for overconfidence, to using indicators of underconfidence, and using non-binary measures of manager prediction errors and recall errors. Across a wide range of different models, overconfident predictions are associated with overly positive memories.[33] Other robustness checks, reported in Appendix I, show that results are similar if we add additional controls. These include other moments of the distribution of past performance (median, mode, max, min), in case manager memories of Q2 are correlated with these other summary statistics of past performance. The regressions also include days between eliciting manager predictions and the end of Q4 of 2015, and controls for other manager traits. These controls are not significantly related to managers' overconfident predictions, with the exception of the survey question measuring self-assessed confidence. Finally, Table I8 in the appendix shows robustness checks on whether the relationship of overly positive memories to overconfidence remains similarly strong as manager experience increases; in our various specifications, interaction terms between the indicator for overly-positive memories and manager experience are not statistically significant, but the point estimates suggests that if anything the relationship is getting stronger with experience. In summary, our reduced form analysis finds support for a signature prediction of the motivated beliefs explanation for overconfidence, that persistent overconfidence about the future goes hand-in-hand with overly-positive memories of past feedback.

4 Structural analysis

This section provides results from estimating the parameters of different variants of a structural model. The framework allows formulating different types of mechanisms, to see which of these might help bring the model closest to the data.

4.1 Baseline Structural Model of Bayesian Prediction

The first version of the structural model, which we call the Baseline Structural Model, assumes that managers are Bayesian. It explores whether manager predictions might be rationalizable by a mechanism in which managers learn about an underlying "type" based on individual histories of tournament outcomes. The model assumes that there are a finite number of periods $t = 1, 2, ...T$ corresponding to quarters. Each manager

[33]Focusing on the 42 regression specifications that include the full set of controls, 41 have a coefficient for the measure of manager recall that is of the expected sign, and 30 are statistically significant.

k has a type a_k that takes on a value between 1-5 and is time invariant.[34] Type depends on immutable characteristics of the manager such as managerial ability and time-invariant characteristics of the manager's store, denoted θ_k (which we call quality), and so $a_k = \Gamma(\theta_k)$ (alternative interpretations of mapping our formal model to observables are discussed in Appendix K).

Every period a public signal $s_{k,t}$ is generated for each manager, taking on an integer value between 1 and 5. This is manager k's quintile in the quarterly tournament in period t.[35] Suppose that a manager's signal is a stochastic function of the manager's type $a_{k,t}$, i.e., $s_{k,t}$ depends partly on type but partly on luck. Denote by $p_t(s|a)$ the probability of a given signal s, conditional on a particular type a, in time period t. All information about the probabilities of signals associated with different types can then be summarized in a 5 by 5 "type-to-signal" matrix denoted P_t. Each row of the matrix corresponds to a type, and moving across the columns the $p_t(s|a)$'s give the probabilities of observing different signals for that type.

At any given time a manager will have a belief distribution f that captures the probabilities that the manager assigns to being each of the possible types, with $f_{k,t}(a)$ denoting the belief that individual k is of type a in time period t. Beliefs about types also give rise to beliefs about what signal will be generated at the end of period t. Manager posterior beliefs about signal probabilities are denoted g, with $g_{k,t}(s) = \sum_a f_k(a)p_t(s|a)$. For example, if a manager thinks there is a 50/50 chance of being type 5 or type 4, then $g_{k,t}(s)$ is constructed by combining the probability distributions for rows 5 and 4 of P with equal weights. We assume that the manager bets on whatever is the most likely signal according to $g_{k,t}(s)$, i.e. the modal signal.

The goal is to establish, in the context of the model, what individual managers should have believed about their type, and thus their probabilities of observing different signals, captured by $g_{k,t}(s)$. Given $g_{k,t}(s)$, it is possible to specify on which signal an individuals should have bet. As researchers we do not, however, observe manager type $a_{k,t}$, P, $f_{k,t}$, nor $g_{k,t}(s)$. Thus, these need to be estimated.

Estimation of the model relies on various identifying assumptions, including ones

[34]The assumption of five types is arbitrary, but has a natural interpretation in terms of reflecting individuals' quintiles in terms of ability. Adding more types makes it less likely that the data could be rationalized in a Bayesian way. As Benoit and Dubra (2011) point out, rational overconfidence requires weak beliefs, but adding more types allows for stronger beliefs (as an extreme example, note that with a single type, individuals always must believe that each quintile is equally likely, regardless of history). Adding additional types would require adding additional data, which in our setting means looking at the probability of a signal conditional on two periods of history. However, most of these entries are sparsely populated making for difficult identification.

[35]In reality individuals observe their rank precisely. Thus, our assumption that they only observe the quintile implies that we model individuals receiving a coarser signal than they actually do. As is well known, supposing individuals receive a coarser signal means that their posterior beliefs will be less extreme, making it easier to rationalize behavior with private information, as in Benoit and Dubra (2011).

that ensure time invariance of a manager's type, as well as time invariance of the environment, encapsulated by P. Intuitively, these assumptions ensure that variation in signals over time is just due to (mean zero) noise, so that it is possible for managers (and the researchers) to update initial priors about a manager's type using tournament outcomes as a sequence of noisy signals.

In the baseline model, manager type is time invariant because it depends only on θ_k. Managers are initially uncertain about θ_k and learn over time. Specifically, they begin with uniform common prior beliefs $f^0(a) = 0.2$ for all a. They observe a series of public signals about their type, and subsequently update their beliefs about time invariant θ_k (and $a_{k,t}$) using P and Bayes' rule. Based on their beliefs, they make a best guess of what signal they will see in Q4 of 2015. In the robustness checks for the baseline model, we consider a version of the model in which type is potentially non-stationary, because it depends on an endogenous variable, manager effort. We discuss an alternative identification strategy in this case.

This, and subsequent, versions of the structural model assume that managers start with uniform priors, even though in reality managers might start with overconfident priors. This reflects data limitations; while we can observe the distribution of predictions about modal quintile among inexperienced managers, we have no empirically disciplined way to calibrate the strength of their priors. While this might reduce the fit of the model to the data, it does not interfere with comparing the relative goodness of fit of different versions of the model, as long as we maintain the same assumption of uniform priors for all models.

Estimation of the model is done in three steps; we summarize these briefly here but give further technical details in Appendix J.1. *Step 1* is to estimate P. The structure of the model implies a mapping between this unobserved matrix, and a $5x5$ "signal-to-signal" matrix denoted Z. The rows of Z give the probabilities of getting each of 5 possible signals in quarter t conditional on a given signal in $t-1$. In the data, each period is a quarter, and the signals $s_{k,t}$ are observed. Thus, for each period t we have a (noisy) observation on Z, denoted Z_t. Based on the number of quarters in the historical performance data we have 33 Z_ts. Note that the average of these Z_t's is the transition matrix $\hat{Z}$ discussed above in Section 3. It is possible to estimate the P that best fits these data, subject to several constraints, as described in Appendix J.1. The resulting estimate is denoted by $\hat{P}$ (we report $\hat{P}$ in the appendix). *Step 2* is to start with uniform priors about each manager's type, and then use $\hat{P}$, each manager's history of tournament outcomes, and Bayes' rule to calculate a posterior distribution across types for each manager. For example, for a manager with a sequence of 1's and 2's, the posterior puts most of the weight on the manager being the worst type, but also non-zero weight on better types. *Step 3* is to use the posterior distribution across types to construct the posterior distribution of the probabilities of different signals, our estimate of $g_{k,t}(s)$, and identify each manager's modal quintile signal for Q4 of 2015. This is denoted the "Bayesian prediction" for a manager. For example, for a manager who is likely to be the worst type, the modal signal will be 1. As in the reduced form model, we suppose that managers bet on the signal they believe is most likely to occur.

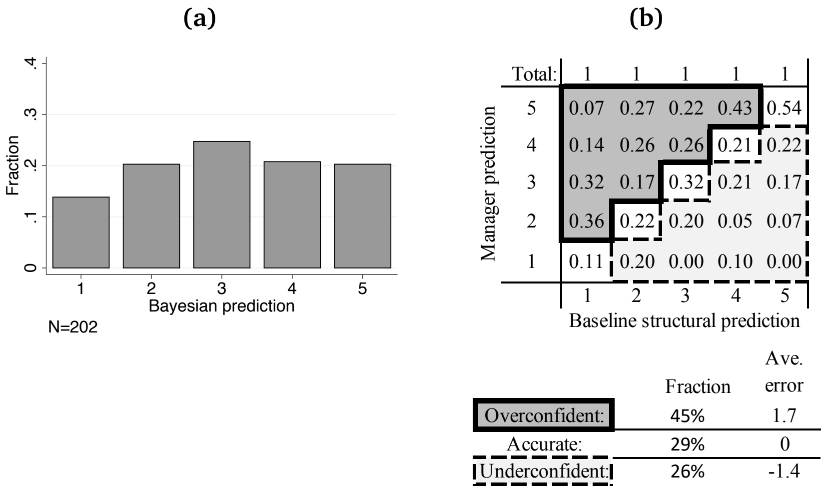
Panel (a) of Figure 5 shows the distribution of Bayesian predictions. In contrast to the reduced form predictors, the distribution of predictions from the structural model is more hump-shaped. This reflects the assumed structure of manager types, and the fact that errors in identifying a manager's type, and most likely signal, can only be upwards for the worst type, and downwards for the best type.[36] Rule of thumb predictors, and multinomial logit predictors, did not have this extra step of inference about a manager's (assumed) underlying type.

Turning to a comparison with manager predictions, a first observation is that predictions from the baseline structural model are less skewed towards predicting high quintiles than manager predictions (recall Figure 1). Furthermore, Panel (b) of Figure 5 shows that manager predictions are overconfident relative to predictions from the baseline structural model. The plurality of managers, 45%, predict a higher quintile for Q4 than the structural model identified as their most likely quintile, whereas only 26% predict a lower quintile. The magnitude of the average prediction error is also larger in the overconfident compared to underconfident direction, 1.7 quintiles versus 1.4 quintiles, respectively. Thus, manager predictions are overconfident, compared to what one would expect if they form predictors in a purely Bayesian way as specified by the model.

The model assumes that managers use the true P to form predictions, but we must estimate P; this raises the question whether the difference between manager predictions and the model predictions could be explained by the estimation error in our model. We take into account noise in the model predictions by re-estimating P 100 times using a moving block bootstrapping design (for details see Appendix J.1), denoting each matrix $\tilde{P}_n$, and generating a new distribution of predictions each time. This yields a distribution of total Euclidean distances from the predictions formed using the bootstrapped $\tilde{P}$'s and the predictions based on the original sample and $\hat{P}$. The Euclidean distance of manager

[36]To see this, suppose that the model says that the worst type is very likely for a manager, with a corresponding modal signal of 1. There will still be some positive probability placed on better types, and their modal signals, however, in the posterior distribution across signals. The modal signal from that posterior distribution could be, e.g., a signal of 2. Likewise, for managers who are likely to be the best type, with a signal of 5, the posterior distribution across signals could put their modal signal at 4. For managers who are likely to be the middle type, the posterior puts weight on both better and worse types, and errors can be more symmetric.

Figure 5: Distribution of Bayesian Predictions and Manager Predictions Compared to Bayesian



Notes: Predictions are in terms of quintiles of Q4 performance, with 5 being the best. Prediction errors are also in terms of quintiles.

predictions from the predictions based on $\hat{P}$ lies far in the tail of the bootstrapped distances.[37] A similar procedure allows rejecting at the 5-percent level that the error in the model can generate the degree of overconfidence in manager predictions, as measured by the fraction of overconfident predictions minus the fraction of underconfident predictions (see Figure J1 in the appendix). Using other distance metrics, for example one that weights Euclidean distance by the size of the deviation in quintiles, delivers the same result: the observed distance is well outside the test-statistic distances.

4.1.1 Robustness checks

The model makes various assumptions as part of the identification strategy. These have analogues in the reduced form analysis, so robustness checks on the structural model follow a similar logic to those for the reduced form analysis. The results from robustness checks on the structural model are summarized in Table K1 in the appendix. These show that results are similar across a variety of different approaches.

One concern is that even if the aggregate distribution of types may be stationary, an individual manager's placement within this distribution may be non-stationary, at least early in a manager's career; this can be captured by modifying our model so that manager type is partly endogenous, depending on effort. Denoting effort by $e_{k,t}$ we have with $a_{k,t} = \Gamma(\theta_k, e_{k,t})$. As managers learn about the immutable component of type, θ_k, they will be adjusting optimal effort over time. With experience, however, managers

[37] If we run a χ-squared test of the difference between model and manager predictions we obtain $p < 0.01$. We prefer the bootstrapping procedure, however, for reasons discussed in Section 3.2.

should learn θ_k with certainty, and converge to a constant optimal effort and stationary type. This convergence argument suggests that the model with endogenous type can be identified by looking at managers who are experienced, and excluding signals from early in their tenures when forming predictions (see Appendix K for more details on this version of the model).

Checking whether results are similar for experienced managers is also a way to address the possibility that managers start the job with rationally overconfident priors; while rationally overconfident priors could explain why managers are more optimistic than the structural predictions initially, this cannot explain overconfidence that persists despite substantial feedback. Of course, this raises the question how much feedback is substantial enough for a Bayesian manager to learn, if they start with overconfident priors. Based on our estimated P, however, we calculate that mangers should learn relatively quickly even if they start with very strong priors. For example, if a manager initially places a 90% probability on being the best type, and 2% on each of the other types, but is actually the worst type, on average he or she will be expected to converge to a correct belief about their most likely type after only two quarters. As another example, suppose the manager starts with the same priors but is actually the middle type. Then after 4 quarters the expected modal belief has converged to the truth.

It turns out empirically that overconfidence is similarly prevalent estimating the structural model using only experienced managers and recent signals. This shows that results are not driven by initial within-manager non-stationarity as managers are learning their types, and it is also not consistent with a Bayesian explanation based on rationally overconfident priors. We also try to address concerns about environment non-stationarity, such that the aggregate distribution of types may change over time, by estimating P, and basing predictions, using only signals from recent quarters. Other robustness checks show that results are similar dropping signals from previous stores; if predictions are based on excluding Q3 of 2015; or if we use only signals from quarters with national tournaments.

The structural model also assumes individuals always bet on the most likely signal. This is a strong assumption, and so it is reasonable to wonder whether allowing for choice errors may accommodate the behavior we observe. The model can be augmented to allow choice errors; individuals might make random choice errors at the stage of selecting on which signal to bet, in the sense of a standard discrete choice model (e.g. McFadden, 1974). However, such an extension does not prove to be much help in matching the data. Although large enough choice errors can help rationalize the total observed degree of deviation from the average model generated predictions, it cannot match the asymmetry of deviations, i.e. the extreme degree of overconfidence relative to underconfidence. Moreover, the degree of choice errors required for this seems extreme. Details are provided in Appendix L.[38] A different type of bounded rationality would be if individuals misperceive the informativeness of signals. Misperception of past signals can be closely linked to memory distortions, so we discuss this robustness check in the appendix on distorted memory, Appendix N.

4.1.2 Model augmented with private information

So far the structural model has assumed that managers only use public signals to form predictions, but it is possible to build on the model framework to allow for managers receiving private signals as well. We explore whether there is a signal structure for private signals that can bring the model close to matching the observed data on manager predictions (details are in Appendix M).

Suppose that after observing all public signals and having a posterior belief vector over types, managers receive a private signal. Since for Bayesians, signals are exchangeable in order, we can suppose that the private signal occurred in the last period, i.e., Q3 of 2015, without loss of generality. These signals can either be interpreted as summarizing a sequence of signals drawn over time about an underlying quality, or as a one-time set of signals, received right before the manager makes the prediction, that give information about a shock that will affect the manager's quality starting in the next period. There are 5 potential private signals, 1 to 5, and the probability distribution over these signals may vary by manager type.[39] This private information can be summarized in a 5 by 5 type-to-signal matrix, which we denote Q, with the same interpretation as P in the baseline model, i.e., each row corresponds to a type, and the entries give the probabilities of that type receiving the different possible signals. It is possible to estimate the elements of Q that minimize the distance of the model predictions from the observed manager predictions. Since realizations of private signals are not observed, and cannot be fed into the model to generate predictions, the model predictions are based on calculating the expectation across different possible private signals conditional on type as well as the different possible manager types. The estimated Q thus minimizes the

[38] An alternative hypothesis is that individuals bet incorrectly due to a distortion of probabilities unrelated to motivated beliefs. For example, individuals may distort beliefs in a way consistent with cumulative prospect theory or rank-dependent utility. We test such a hypothesis, supposing that individuals follow the baseline model of Bayesian updating up until they generate their probabilities of signals g. We then assume they distort the probabilities in a way consistent with rank-dependent weighting functions, using the functional form and parameter estimates from Bruhin, Epper and Fehr-Duda (2010). Such an approach still leads to rejecting the model at the 1% level, even if we allow for individual heterogeneity and assign individuals to either use probability weighting, or act as true Bayesians, in a way that helps the model to best match the data. Results are available upon request.

[39] The proof of Theorem 4 of Benoit and Dubra (2011) shows that, when considering quintiles, that considering at most 5 signals is sufficient to achieve the maximal distortion of beliefs towards overconfidence with private information.

distance between what managers actually bet and what the model predicts managers should bet on average (across many draws from the private signal distributions).

The estimated Q (see Appendix Table M1) has a structure that does not seem particularly plausible as a real information structure accessible to our managers. For example, the modal private signal for the worst type is 4, whereas the modal private signal for the best type is 2. This implies a negative correlation between public and private signals conditional on type, i.e., people who do poorly in tournaments tend to systematically observe more positive private signals than people who do well in tournaments.

We want to assess whether the difference between the model predictions and manager predictions could be explained by noise in the model. To do this we simulate the model 100 times. We start with the 100 bootstrapped $\tilde{P}_n$'s derived when testing the baseline model, so that we incorporate noise coming from estimating the model using tournament outcomes. Each of these implies posteriors over manager types based on public signals, and through the estimated Q, an associated probability distribution over the possible private signals. For each of these 100 cases we then incorporate noise coming from the randomness in private signals. We draw from the distribution of private signals each time and generate a distribution of bets using the updated beliefs. Averaging across all 100 simulations gives the average betting behavior.[40] For each simulation we calculate the Euclidean distance of the simulated bets from the average betting behavior. This yields a distribution of 100 distances. It turns out that the distance of manager predictions from the average betting behavior lies far in the tail of the simulated distances and we can reject that the manager predictions lie within the bounds of the randomness in the model at the 1-percent level. Also, the degree of asymmetry in prediction errors is significantly different from the observed asymmetry at the 1-percent level.

4.2 Model augmented with biased memory

In this section we augment the structural model to take into account the data on biased memory. This may be expected to help the model better match the data on manager predictions, for two reasons. First, if managers form predictions based on overly-positive memories of past signals, this could help generate predictions that are more confident than the baseline structural model. Second, incorporating individual-level variation in memory distortion may help explain heterogeneity in manager overconfidence. Indeed, we have seen that having overly positive memories is predictive of overconfidence relative to reduced form predictors (results are also similar measuring overconfidence

[40]In the limit this average is the expected betting behavior derived from the model, which was used to estimate Q.

relative to the baseline structural model, see Table N1). Our model incorporates individual heterogeneity in memory distortion in several ways, which are disciplined by the memory data and suggested by theory. In estimating the model we do not distinguish between different possible motivations for distorting memory, but merely seek to establish whether incorporating memory distortions can help better match observed manager predictions. More details on the model are provided in Appendix N.

We allow for the possibility that some managers are "motivated", in the sense that they want to distort memories, and we incorporate a technology for memory distortion by adding a "memory matrix", denoted M. Each row corresponds to having received one of the public signals 1 to 5. Each column gives probabilities that the manager remembers signals 1 to 5 (i.e., quintile ranking). The data on manager recall provide a way to calibrate M. For any given quintile of actual performance, the matrix uses the empirical frequencies of remembering different ranks that fall in quintiles 1 to 5 (we report M in Appendix N). The observed frequencies have several notable features. Memory distortions will most frequently occur in the overly-positive direction; memories will be correlated with actual signals; there will be variation in the extent of memory distortion conditional on a given signal. The latter feature introduces a first source of heterogeneity, within-managers over time, which can be thought of as reflecting an inherent randomness in the memory technology.[41]

We assume that all managers have access to the same M, but we allow for a second source of heterogeneity across managers, such that some managers may be motivated to distort memories but others may be "unmotivated." The latter managers will choose not to use M. If a manager does not use M then memories are always accurate. Managers are assumed to update beliefs each time they receive a new signal but using the remembered signal rather than the actual signal (the remembered signal could be the same as the actual signal).

In formulating how the motivated managers update beliefs based on remembered signals, it is necessary to make assumptions about self-awareness. One possibility is to assume that such managers are sophisticated, as in Bénabou and Tirole (2002). In this case, they do not have access to the actual past signal, but take into account the motives of past selves, and the technology for memory distortion encapsulated in M, when updating beliefs. Even with full sophistication, the individual can still make overconfident predictions, although as Compte and Postlewaite (2004) note, eventually sophisticates should learn their true type given sufficient signals. At the other extreme, one could assume motivated managers are completely naïve, treating remembered signals as the actual signals. This would lead to more extreme overconfidence, which can

[41]For more on this randomness, see discussion in footnotes for Section 3.3.

persist indefinitely. Perhaps more realistically, there could be heterogeneity in the degree of self-awareness among motivated managers. We also allow for this third type of heterogeneity in the model.

Since little is known about the prevalence of motivated versus unmotivated individuals, or about different levels of self-awareness about memory distortion, we let the data inform us about the most appropriate assumption for each manager – sophisticated, naïve, or "unmotivated" (always remember signals accurately). To do this we run 100 simulations for each manager, under each of the three different assumptions. Specifically, we start with the 100 bootstrapped $\tilde{P}_n$'s from the baseline estimation. For each of these we simulate the model, which involves drawing from the relevant probability distribution in M for each signal observed by a manager (if they are naïve or sophisticated) to establish the remembered signal, and having the manager update beliefs based on the sequence of remembered signals, $\tilde{P}_n$, and the assumption about the manager's type. This yields a distribution of 100 bets for a manager, for each assumption. Taking the average betting behavior for each assumption, we assign the manager to the type that has the smallest difference between the average betting behavior and the manager's actual bet. We arrive at 40% naifs, 31% sophisticates and 29% unmotivated.[42]

Our next step is to assess the ability of the model to generate overconfidence, and to fit the data on actual manager predictions. We do 100 simulations of the model with each individual fixed to their assigned type. Specifically, we start with the 100 bootstrapped $\tilde{P}_n$'s from the baseline estimation, we draw from the appropriate distributions in M to establish remembered signals (for managers who are motivated), and we assume that managers update beliefs using the relevant $\tilde{P}_n$ and the rule for their type. This yields a bet for each manager for each simulation. Taking the average across the 100 simulations gives expected betting behavior for each manager.

We find that the model generates average betting behavior that is substantially overconfident relative to the baseline structural model: 33% of managers are overconfident and 17% are underconfident. Recall that manager predictions entail 44% overconfident and 25% underconfident relative to the baseline model. Thus, the model with biased memory generates a similar difference of overconfident versus underconfident managers as the data on manager predictions, although the prevalence of overconfidence is still somewhat smaller.

To assess whether the model is significantly different from manager predictions, we

[42]One caveat is that turnover in the manager population could cause these sample estimates to be biased relative to the fractions present in the worker population as a whole. Suppose that managers who are sophisticated, or who are unmotivated to distort memory, are more likely to leave the firm over time, because those with low ability recognize this and leave the firm. In this case the sample that we use, which requires managers to be present long enough to have an estimated type, is missing some of the sophisticates and umotivated managers who are present in the population as a whole.

use the fact that the 100 simulations yield a distribution of 100 Euclidean distances from average betting behavior. The Euclidean distance of manager predictions from average betting behavior (conditional on our memory augmented model being true) lies at the 90th percentile of the simulated distances. The difference in the fractions of overconfident versus underconfident managers lies at the 87th percentile of the simulated distribution of differences. Thus, unlike for previous versions, we cannot reject that this version of the structural model matches manager predictions at conventional significance levels. Furthermore, the distance between manager predictions and the model, at 135, is substantially smaller than for other models, e.g., the distances for the baseline model and the model with private information are both greater than 200.

To check how sensitive the model performance is to assumptions about the extent and nature of heterogeneity, we explored several alternative restrictions: (1) optimally assigning managers to be either sophisticated or naïve, using the procedure described above, but without the possibility of unmotivated managers; (2) randomly assigning the three types (results turn out to be similar across various assignment probabilities); (3) assuming all managers are motivated and naïve; (4) assuming all managers are motivated and sophisticated. Each of these reduces the fit of the model to varying degrees, compared to our approach above, but the distances between the model and manager predictions are still smaller than for other versions of the structural model. Appendix N provides more details. While this version of the structural model is clearly still imperfect in terms of capturing the nuances of manager predictions, we conclude that it moves in the right direction in terms of helping to explain the data.

5 Discussion and Implications

The findings in this paper are consistent with managers being overconfident about their future relative performance in the workplace, despite substantial feedback. The evidence of overly-positive memories of past feedback, and a link between these and overconfident predictions, points to an explanation based on motivated beliefs. This is not to say that motivated beliefs are the entire explanation for the observed overconfidence; there could be other factors at work as well, both rational and psychological.

Evidence of motivated beliefs and biased memory in the field has important implications for economic theory. It implies that overconfidence can be a persistent phenomenon in field settings with feedback, in contrast to standard models of belief formation. It also changes the ways that individuals respond to feedback, relative to standard theories of information provision and optimal feedback, and it implies that variables that should not matter for behavior in standard models may influence decisions. For example, presenting feedback in ways that are less “ego-threatening” might matter for belief updating. There are also implications for theories of optimal incentive design if agents are persistently overconfident.

A motivated beliefs explanation for overconfidence also has different implications for welfare and policy, compared to if overconfidence is a cognitive mistake. In particular, it becomes less obvious that one should implement policies to minimize overconfidence. For individuals, welfare losses that arise because of making choices based on biased beliefs could be offset by an intrinsic utility benefit of positive beliefs, or by benefits in terms of counteracting other biases. From the perspective of a principal, biased beliefs might lead managers to make mistakes on the job, but there could also potentially be offsetting benefits, e.g., if it greater confidence counteracts self-control problems.[43] On the extensive margin, overconfidence might make managers overestimate the value of employment relative to the outside option, with the benefit to the principal of relaxing the participation constraint.

Although opening the black box of managerial performance is not the focus of this paper, our data can shed some light on whether and how manager beliefs feed into the ways that managers perform and make decisions. One caveat is that the sample of managers is relatively small, to study determinants of managerial performance, and there are limited outcomes on decision making that we can study. Another caveat is actually a methodological implication of our evidence that beliefs are motivated. Once overconfidence is motivated it is endogenous, which may complicate efforts to understand the impact of overconfidence on outcomes such as performance. For example, suppose that some individuals have self-control problems in the form of present-biased preferences, as in Bénabou and Tirole (2002). Those with self-control problems may anticipate poor performance in the future, and thus implement overconfident beliefs. If the overconfidence does not completely counteract self-control problems, there could actually be a negative correlation between greater confidence and performance, but this would conceal a positive effect, because the counterfactual would have been even worse performance. This methodological implication is potentially important for interpreting past and future empirical research on overconfidence.

With these caveats in mind, we regressed different aspects of future manager performance (from Q1 and Q2 of 2016) on manager predictions about Q4 of 2015, as well as various binary indicators for overconfidence. It turns out that managers who are overconfident about Q4 of 2015 do not do any worse, or better, in terms of overall future performance compared to other managers. Digging deeper into the underlying

[43] See, e.g., Hvide (2002), Bénabou and Tirole (2003), Fang and Moscarini (2005), Gervais and Goldstein (2007), Santos-Pinto (2008, 2010), De la Rosa (2011), and Foschi and Santos-Pinto (2017) for discussions of implications of biased beliefs for contract form, performance, and welfare.

dimensions of performance, however, there are differences. Overconfident managers have higher profits, but they also have worse customer service scores (these results are generally statistically significant but not in all specifications; Appendix O provides details). The findings are intriguing, as they suggest the possibility that overconfidence might be associated with strengths and weakness on different aspects of the job. Overconfidence might also be related to a manager's tendency to stay at the firm, if it causes managers to value the job more relative to outside options. Using different indicators for overconfidence, point estimates suggest lower hazard rates of leaving the firm for overconfident managers, but these differences are relatively small and not statistically significant (see Figure P1). One explanation for the weak relationship could be that the managers' overconfidence is not entirely job-specific, and inflates estimates of their outside options as well.

To explore how manager beliefs are related to managerial decision-making, we related some indicators of management style to our measures of manager overconfidence. One finding is that overconfident managers tended to hire fewer assistant managers than recommended by store-specific guidelines provided by the firm (results are less precisely estimated for some of the binary indicators; details are in Appendix Q). This suggests a type of overconfidence in terms of being able to manage the store without additional help. It could also potentially contribute to higher profits, because hiring fewer assistant managers reduces the wage bill, but it could seemingly also have downsides, e.g., possibly harming customer service. Managers with overconfident predictions also exhibited a type of overconfidence in management style in a measure collected in the lab in the field study. Specifically, overconfident managers were more likely to be willing to pay a cost, to be able to choose for a worker which of two brain teaser problems to try to solve, rather than allowing the worker to choose which one to solve (empirically, the two questions were equally difficult). Manager payoffs depended on the worker getting the answer correct; for more details see Appendix Q. This suggests that overconfident managers could tend to think that they are better able to assess task difficulty for a worker than the worker himself, even in situations where this is unlikely to be the case. Although correlational and exploratory, these findings provide a starting point for future research on how overconfidence shapes managerial style.

A final point is that our analysis, like much of the literature, focuses on overconfidence. There is a smaller literature, however, that discusses underconfidence (e.g.the original Kruger and Dunning, 1999, analysis finds that very competent individuals are underconfident). Our results demonstrate that while overconfidence is more prevalent in our field setting, some individuals exhibit underconfidence. This heterogeneity, and the reasons for it, are an interesting area for future research.

References

Akerlof, George A and William T Dickens, “The economic consequences of cognitive dissonance,” *The American economic review*, 1982, *72* (3), 307–319.

Barber, Brad M and Terrance Odean, “Boys will be boys: Gender, overconfidence, and common stock investment,” *The Quarterly Journal of Economics*, 2001, *116* (1), 261–292.

Bénabou, Roland, “The economics of motivated beliefs,” *Revue d’économie politique*, 2015, *125* (5), 665–685.

_ **and Jean Tirole**, “Self-confidence and personal motivation,” *The Quarterly Journal of Economics*, 2002, *117* (3), 871–915.

_ **and** _ , “Intrinsic and extrinsic motivation,” *The Review of Economic Studies*, 2003, *70* (3), 489–520.

_ **and** _ , “Willpower and personal rules,” *Journal of Political Economy*, 2004, *112* (4), 848–886.

_ **and** _ , “Identity, morals, and taboos: Beliefs as assets,” *The Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

_ **and** _ , “Mindful economics: The production, consumption, and value of beliefs,” *Journal of Economic Perspectives*, 2016, *30* (3), 141–64.

Benoît, Jean-Pierre and Juan Dubra, “Apparent overconfidence,” *Econometrica*, 2011, *79* (5), 1591–1625.

_ **,** _ **, and Don A Moore**, “Does the better-than-average effect show that people are overconfident?: Two experiments,” *Journal of the European Economic Association*, 2015, *13* (2), 293–329.

Bracha, Anat and Donald J Brown, “Affective decision making: A theory of optimism bias,” *Games and Economic Behavior*, 2012, *75* (1), 67–80.

Bruhin, Adrian, Helga Fehr-Duda, and Thomas Epper, “Risk and rationality: Uncovering heterogeneity in probability distortion,” *Econometrica*, 2010, *78* (4), 1375–1412.

Brunnermeier, Markus K and Jonathan A Parker, “Optimal expectations,” *The American Economic Review*, 2005, *95* (4), 1092–1118.

Burks, Stephen V, Jeffrey P Carpenter, Lorenz Goette, and Aldo Rustichini, "Overconfidence and social signalling," *Review of Economic Studies*, 2013, *80* (3), 949–983.

Camerer, Colin and Dan Lovallo, "Overconfidence and excess entry: An experimental approach," *American Economic Review*, 1999, *89* (1), 306–318.

Carrillo, Juan D and Thomas Mariotti, "Strategic ignorance as a self-disciplining device," *The Review of Economic Studies*, 2000, *67* (3), 529–544.

Charness, Gary, Aldo Rustichini, and Jeroen Van de Ven, "Self-confidence and strategic behavior," *Experimental Economics*, 2013, pp. 1–27.

Cheng, Ing-Haw, Sahil Raina, and Wei Xiong, "Wall Street and the housing bubble," *American Economic Review*, 2014, *104* (9), 2797–2829.

Chew, Soo Hong, Wei Huang, and Xiaojian Zhao, "Motivated False Memory," *National University of Singapore Working Paper*, 2018.

Compte, Olivier and Andrew Postlewaite, "Confidence-enhanced performance," *American Economic Review*, 2004, *94* (5), 1536–1557.

DellaVigna, S. and U. Malmendier, "Paying not to go to the gym," *American Economic Review*, 2006, *96* (3), 694–719.

den Steen, Eric Van, "Rational overoptimism (and other biases)," *American Economic Review*, 2004, *94* (4), 1141–1151.

Di Tella, Rafael and Perez-Truglia, Ricardo and Babino, Andres and Sigman, Mariano, "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism," *American Economic Review*, 2015, *105* (11), 3416–42.

Eil, David and Justin M Rao, "The good news-bad news effect: asymmetric processing of objective information about yourself," *American Economic Journal: Microeconomics*, 2011, *3* (2), 114–138.

Fang, Hanming and Giuseppe Moscarini, "Morale hazard," *Journal of Monetary Economics*, 2005, *52* (4), 749–777.

Foschi, Matteo and Luís Santos-Pinto, "Subjective performance evaluation of employees with biased beliefs," *European University Institute Working Paper*, 2017.

Gervais, Simon and Itay Goldstein, "The positive effects of biased self-perceptions in firms," *Review of Finance*, 2007, *11* (3), 453–496.

Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen, "Motivated Self-Deception and Unethical Behavior," *University of California, San Diego Working Paper*, 2015.

Gottlieb, Daniel, "Will you never learn? Self deception and biases in information processing," *Unpublished Manuscript, Princeton University*, 2011.

Haisley, Emily C and Roberto A Weber, "Self-serving interpretations of ambiguity in other-regarding behavior," *Games and Economic Behavior*, 2010, *68* (2), 614–625.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack, "Unrealistic expectations and misguided learning," *Econometrica*, 2018, *86* (4), 1159–1214.

Hestermann, Nina and Yves Le Yaouanq, "It's not my Fault! Self-Confidence and Experimentation," 2019.

Hoffman, Mitchell, "How is Information Valued? Evidence from Framed Field Experiments," *The Economic Journal*, 2016, *126* (595), 1884–1911.

_ **and Stephen V Burks**, "Worker Overconfidence: Field Evidence and Implications for Employee Turnover and Returns from Training," Technical Report, National Bureau of Economic Research 2017.

Hvide, Hans K, "Pragmatic beliefs and overconfidence," *Journal of Economic Behavior & Organization*, 2002, *48* (1), 15–28.

Köszegi, Botond, "Ego utility, overconfidence, and task choice," *Journal of the European Economic Association*, 2006, *4* (4), 673–707.

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

Kunda, Ziva, "The case for motivated reasoning.," *Psychological bulletin*, 1990, *108* (3), 480.

la Rosa, Leonidas Enrique De, "Overconfidence and moral hazard," *Games and Economic Behavior*, 2011, *73* (2), 429–451.

Malmendier, Ulrike and Geoffrey Tate, "CEO overconfidence and corporate investment," *The Journal of Finance*, 2005, *60* (6), 2661–2700.

_ **and** _ , "Behavioral CEOs: The role of managerial overconfidence," *The Journal of Economic Perspectives*, 2015, *29* (4), 37–60.

Merkle, Christoph and Martin Weber, "True overconfidence: The inability of rational information processing to account for apparent overconfidence," *Organizational Behavior and Human Decision Processes*, 2011, *116* (2), 262–271.

Mijović-Prelec, Danica and Drazen Prelec, "Self-deception as self-signalling: a model and experimental evidence," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 2010, *365* (1538), 227–240.

Mobius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat, "Managing self-confidence: Theory and experimental evidence," Technical Report, National Bureau of Economic Research 2011.

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

Niederle, Muriel and Lise Vesterlund, "Do women shy away from competition? Do men compete too much?," *The Quarterly Journal of Economics*, 2007, *122* (3), 1067–1101.

Oster, Emily, Ira Shoulson, and E Dorsey, "Optimal expectations and limited medical testing: evidence from Huntington disease," *The American Economic Review*, 2013, *103* (2), 804–830.

Park, Young Joon and Luís Santos-Pinto, "Overconfidence in tournaments: evidence from the field," *Theory and Decision*, 2010, *69* (1), 143–166.

Quattrone, George A and Amos Tversky, "Causal versus diagnostic contingencies: On self-deception and on the voter's illusion.," *Journal of Personality and Social Psychology*, 1984, *46* (2), 237.

Santos-Pinto, Luís, "Positive self-image and incentives in organisations," *The Economic Journal*, 2008, *118* (531), 1315–1332.

_ , "Positive Self-Image in Tournaments," *International Economic Review*, 2010, *51* (2), 475–496.

_ **and Joel Sobel**, "A model of positive self-image in subjective assessments," *The American Economic Review*, 2005, *95* (5), 1386–1402.

Sarver, Todd, "Dynamic Mixture-Averse Preferences," *Econometrica*, 2018, *86* (4), 1347–1382.

Schwardmann, Peter and Joël J Van der Weele, "Deception and self-deception," 2018.

Sharot, Tali, Christoph W Korn, and Raymond J Dolan, "How unrealistic optimism is maintained in the face of reality," *Nature neuroscience*, 2011, *14* (11), 1475.

Smith, Adam, "An inquiry into the wealth of nations," *Strahan and Cadell, London*, 1776.

Vigna, Stefano Della and Ulrike Malmendier, "Paying not to go to the gym," *The American Economic Review*, 2006, *96* (3), 694–719.

Zabojnik, Jan, "A model of rational bias in self-assessments," *Economic Theory*, 2004, *23* (2), 259–282.

Zimmermann, Florian, "The Dynamics of Motivated Beliefs," *Institute on Behavior and Inequality Working Paper*, 2018.