Persistent Overconfidence and Biased Memory: Evidence from Managers

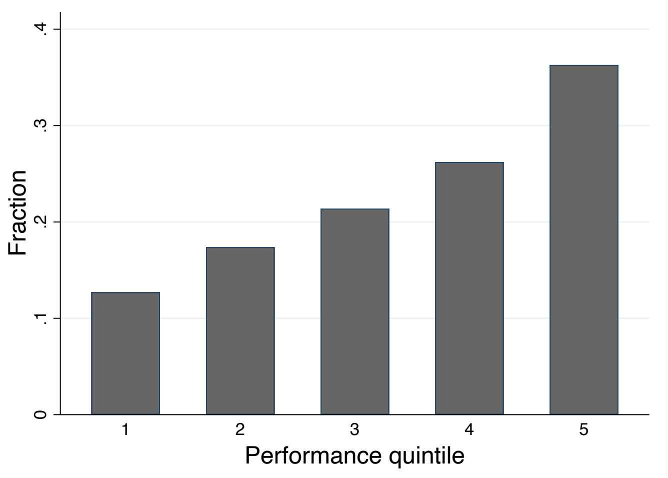
David Huffman Collin Raymond Julia Shvets

Online Appendix

A The shape of the incentive scheme

The figure illustrates that the incentive scheme is relatively high-powered and also convex. Managers can lose or gain a substantial amount of money, relative to the base quarterly salary, depending on what quintile they achieve in the performance ranking for the quarterly tournament.

Figure A1: Median bonus as a fraction of quarterly base salary by quintile of performance (5=best)



Notes: The figure uses the sample period Q1 of 2008 to Q4 of 2015.

B Details on creation of the historical performance dataset

The creation of the dataset involved addressing a few issues. First, in a few quarters two managers were assigned to the same store for a period of time. In such situations, the tournament outcome of the store was assigned to the manager who spent more of the quarter running the store. Second, in the first quarter that a store opens, the company does not include the store in the regular ranking for the tournament. Thus, the analysis excludes observations for the first quarter that a store opens. Third, newly hired managers are not part of the tournament in their first quarter so data on initial quarters of managers are not part of the analysis. Fourth, the scope of the tournament has changed in alternating quarters in recent years, so we 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. Prior 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.[1] Fifth, the rules of the tournaments change very slightly over the history of the firm, particularly the precise scores assigned to each band on a given dimension of performance and the number of bands. To achieve a consistent performance measure over time we used the Q4 of 2015 tournament rules and the raw performance data to construct the Final Bonus and overall rank of each manager in each quarter. The average correlation between the recorded overall rank and the rank according to Q4 of 2015 rules in a given quarter is 0.95 (Spearman; $p > 0.01$).

[1]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; $p < 0.001$).

C Example feedback table given to managers after quarterly tournaments

Figure C1: Example feedback table for the quarterly bonus tournament

National tournament

Quarter X, 20XX

	Overall rank	Store number	Store in charge	Store name	Area code	Store manager (name)	Regional manager (name)	Sales growth	Sales growth score	Profit relative to target	Profit score	Service	Service score	Regional manager evaluation	Regional manager evaluation score	Base Bonus	Top performance bonus	Extra bonus for all stores	Area bonus	Final bonus
Column ID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	1																			
	2																			
	3																			
	4																			
	5																			
	...																			

Notes: Column ID is added by the authors. The wording in several headings is changed slightly for clarity and to avoid using the firm's internal labels

A = rank(T); P=I*K*M*O; T=P*Q*R*S

D Instructions for prediction measure and measure of memory about rank

For easy reference this section provides instructions for the two key measures from the lab-in-the-field study. One is the measure of manager predictions about rank in the Q4 of 2015 tournament. The other is a measure of manager memories of rank in the Q2 of 2015 tournament. These can be found in Part 9 and Part 10, respectively, of the full instructions for the study, which are provided in Appendix U.

D.1 Prediction measure

How does it work?

> Think about the company *bonus tournament* **in this quarter, Q4 of this year**.
>
> We will ask for your best guess about your shop's overall position (rank) in the company *bonus tournament* for Q4.Base Bonus
>
> *Company X* expects roughly 300 shops to take part in the Q4 *bonus tournament*.

! You do not have to guess your exact position, just a range!

How do you earn money?

> *Please remember that only one part of the study, chosen at random, will contribute to your earnings.*
>
> We will get information from the *bonus tournament*, and will pay you *$22* if your actual position falls within the range you guessed.

Make your decision:

Now, mark the range that you think is most likely for your overall position (rank).

Top 20%	Roughly, ranks 1 to 60	☐
Top Middle 20%	Roughly, ranks 61 to 120	☐
Middle 20%	Roughly, ranks 121 to 180	☐
Bottom Middle 20%	Roughly, ranks 181 to 240	☐
Bottom 20%	Roughly, ranks 241 to 300	☐

D.2 Measure of recalled Q2 of 2015 tournament rank

How does it work ?

> We ask you eight questions about the **company *bonus tournament* that has already happened** in Q2 of this year.

How do you earn money?

> You get $3 if you get all the parts of a question correct. If not we pay a proportion for the parts you got right.
>
> A hint:
>
> The company *bonus tournament* results table looked like this in Q2 (but longer):

<*Tournament table column titles and one row as an example here*>

In the Q2 company *bonus tournament*:

What was (a) your shop's **rank** (position); (b) your overall **Final Bonus?** (We count your answers as right if they are within plus/minus 10 of the actual).

<Tournament table column titles here with the required titles circled>

Rank	Final Bonus
	%

E Summary descriptions of additional control variables from the lab-in-the-field study

This section gives a high-level summary of the measures from the lab-in-the-field study that are used as control variables in the analysis. The full instructions for the lab-in-the field study can be found in Appendix U.

Incentivized measure of risk taking (Part 1 in instructions): The measure is based on Gneezy and Potter (1997): Managers were given an endowment, and could choose how much money to allocate to a safe asset, or to a risky asset. Allocating more money to the risky asset is an indication of willingness to take risks.

Incentivized addition task with piece rate (Part 2 in instructions): Managers had the opportunity to solve addition problems without the aid of a calculator. The time limit was 3 minutes. An addition problem consisted of adding 5 two-digit numbers. Managers were offered a piece rate of about $2 per correct answer.

Incentivized questions about knowledge of the incentive scheme (Q7 of Part 10 in instructions): Managers were asked for the maximum possible value, and the minimum possible value, for the scale that the firm uses to evaluate relative performance on each of the four dimensions of performance. Managers were allowed to give the values for the dimension of their choosing. We paid managers for giving the correct values. We construct an indicator for whether the manager knew the top and bottom values.

Incentivized question about understanding multiplicative nature of the firm's incentive scheme (Q8 of Part 10 in instructions): Managers were asked to imagine a Store A getting the same score Z for all four dimensions, and a Store B getting score values that were different across dimensions, but with a mean of Z across dimensions. Everything else relevant for the Final Bonus is the same for the two stores. The manager was asked which store would have a higher Final Bonus. The correct answer was Store A; managers were paid for the correct answer.

Incentivized measure of willingness to mis-report (Part 10 in instructions, final question): Managers were given a six-sided die and a cup. They were instructed to roll the die in the cup, and then write down the number that they rolled. No-one else could observe their die roll. Managers were offered financial incentives that increased in the die roll reported: zero for rolling 1 or 2; increasing amounts for reporting higher numbers, with the highest payoff for reporting 6. Aggregate data suggests mis-reporting: Roughly 10% report each of 1, 2, or 3, for a total mass of 30%; the remaining 70% reported numbers 4 or higher, with roughly equal proportions for each value. Reporting a higher number is a noisy measure of individual willingness to mis-report.

Self-assessment of willingness to take risks (Part 11 in instructions): Question asking: "Are you a person who is generally fully prepared to take risks, or do you try to avoid risks?" Response scale was from 0 (completely unwilling) to 10 (completely willing).

Self-assessment of willingness to compete (Part 11 in instructions): "Are you generally a person who is fully prepared to compete, or do you prefer to avoid competition?" Response scale was from 0 (completely unwilling) to 10 (completely willing).

Self-assessment of confidence (Part 11 in instructions): "In general, are you a person who is confident that you can do better than others, or are you not that confident?" Response scale was from 0 (not at all) to 10 (very).

Self-assessment of patience (Part 11 in instructions): "How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?" Response scale was from 0 (completely unwilling) to 10 (completely willing).

F Model selection, statistical tests, and robustness checks for manager predictions vs. multinomial logit

F.1 Model selection

We used cross validation, a simple machine learning technique, to select the model with the best predictive power out of a set of candidate models. Cross validation involves randomly dividing the data into k subsets, using $k-1$ subsets to estimate a given model, predicting out-of-sample in the remaining subset, and doing this k times. We did this for $k = 5$, using all data before Q4 of 2015. We considered models using from 1 up to 8 lags, and we considered two different ways of specifying past performance within a given quarter: Percentile of performance, a relatively continuous classification, but restricted to enter linearly; and separate dummy variables for quintile of performance, a coarser classification, but without the restriction of linearity. The model with 8 lags and percentile of performance as the independent variable yielded the smallest average error for predicting out of sample, where we measured the error as the sum of Euclidean distances from the predicted values and the actual quintile outcomes for a given quarter. Using an alternative distance metric that scales the Euclidean difference by the number of quintiles between the model prediction and actual quintile yielded the same results. The model with 8 lags also performed best using traditional within-sample measures that penalize overfitting, such as the AIC.

F.2 Statistical tests for manager predictions vs. baseline multinomial logit

To test whether predictions of the multinomial logit model are significantly different from manager predictions we bootstrap the multinomial logit model. Specifically, 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.[2] We then add up all of these distances across the managers, to get the total Euclidean distance between the bootstrapped distribution of bets and the distribution based on the original sample.

Panel (a) of Figure F1 shows the cumulative distribution of Euclidean distances. The vertical line shows the Euclidean distance of actual manager predictions from the predictions based on the original sample. This is far in the tail, so we can reject at the 1% level that the difference between the model predictions and the manager predictions lies within the bounds of the noise in the model predictions.[3] Results also 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.[4]

To test whether the noise in model predictions can account for the asymmetry in overconfident versus underconfident predictions, we calculate for each bootstrap, the fraction of managers who are overconfident relative to predictions based on the original sample, and the fraction who are underconfident. Panel (b) of Figure F1 shows the

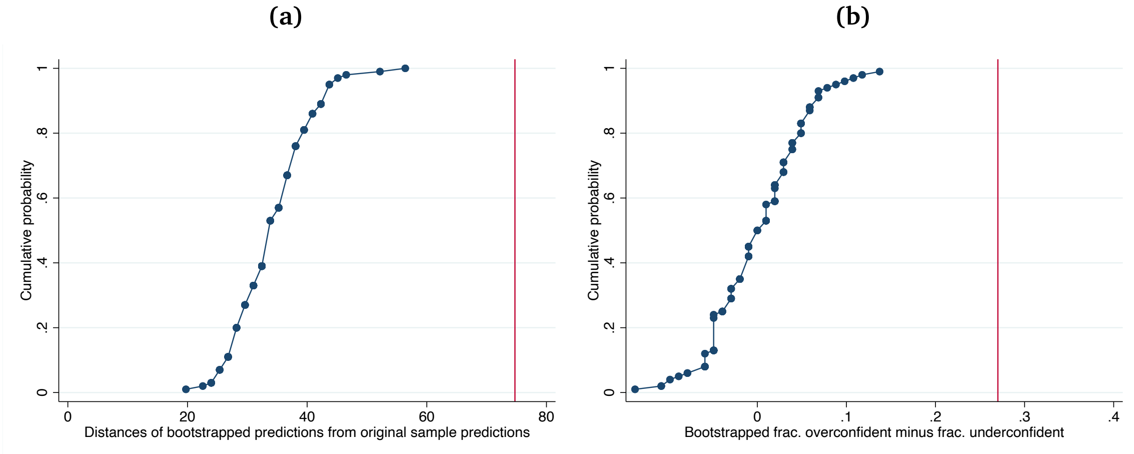
[2]Euclidean distance is the straight line distance between two points given by the Pythagorean formula. For each manager, we are comparing two vectors that describe betting behavior, one from the bootstrap and one from the predictions based on the original sample. These vectors have 5 elements that take on a value of 1 if the manager bets on that quintile and 0 otherwise. If two vectors differ, there is a difference of 1 for two different entries, and the Euclidean distance is $\sqrt{(1-0)^2+(1-0)^2}=\sqrt{2}$.

[3]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.

[4]Results are stronger in this case 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 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.

cumulative distribution of these differences. The vertical line indicates the fraction of managers who are overconfident relative to the original sample predictions, minus the fraction underconfident. The latter is far in the tail, so we can reject at the 1% level that the noise in the model can generate as large of an asymmetry between overconfident and underconfident predictions as is observed for managers.

Figure F1: Statistical Test of Manager Predictions vs. Multinomial Logit Predictions



Notes: The connected (blue) dots in Panel (a) show the cumulative distribution of Euclidean distances between the bootstrapped multinomial logit predictions and predictions based on the original sample. See Section 3.2 in the text for more details on the bootstrapping. The vertical (red) line in Panel (a) shows the Euclidean distance of manager predictions from the predictions of the model using the original sample. The connected (blue) dots in Panel (b) show the cumulative distribution of the differences, for all of the bootstrapped predictions, of the fraction overconfident relative to the predictions based on the original sample minus the fraction underconfident. The vertical (red) line in Panel (b) shows the fraction of managers overconfident relative to the predictions using the original sample minus the fraction of managers underconfident.

F.3 Robustness checks for manager predictions vs. multinomial logit model predictors

This section explores the robustness of the result that manager predictions are overconfident relative to our baseline reduced form multinomial prediction model. The question is whether manager predictions might be explainable by some other plausible prediction model. Table F1 summarizes the results from considering a range of different estimation samples, or specifications that differ in terms of number of lags. Tables F2 and F3 provide the coefficient estimates underlying the summarized results. See Section 3.2 in the text for more discussion on the rationales for the different robustness checks, and table notes for further details on the estimations. All of these regressions maintain the parametric assumption that past performance in a given quarter enters linearly. Table F4 summarizes the results of running the same robustness checks, but with a less parametric specification for past performance: performance in a given quarter is captured by separate dummy variables for quintiles 1 to 4, with 5 being the omitted category. See the table notes for more details. The coefficient estimates underlying the results in Table F4 are available upon request. Manager predictions are consistently overconfident, regardless of which prediction model is used.

Table F1: Summary of robustness checks on manager predictions vs. multinomial logit predictors

	Manager vs. multinomial logit predictors					
	Fraction of managers:			P-values		
	overconfident	accurate	underconfident	different	frac. overconf. - frac. underconf.	N
Overconfident priors:						
8 lag	0.48	0.31	0.21	p<0.01	p<0.01	1744
3 lag	0.43	0.33	0.24	p<0.01	p<0.01	3568
Manager non-stationarity:						
8 lag, drop early	0.46	0.34	0.20	p<0.01	p<0.01	1272
3 lag, current store only	0.39	0.34	0.27	p<0.01	p<0.03	891
Environment non-stationarity:						
3 lag, recent tournaments	0.43	0.32	0.25	p<0.01	p<0.02	667
Imperfect knowledge:						
Excluding Q3 tournament	0.43	0.33	0.24	p<0.01	p<0.01	3391
Nationwide tournaments	0.43	0.32	0.25	p<0.01	p<0.01	1042

Notes: The estimations use historical data from Q3 of 2015 back to Q1 of 2008 unless otherwise noted. P-values test whether manager predictions are different from the model predictions, and whether they are more skewed towards overconfidence. See text for details on bootstrapping. The 8 lag model was selected over models with fewer lags in cross validation; it entails using a sample of relatively experienced managers, those with at least 8 consecutive tournament outcomes. The 3 lag model uses a larger sample that includes all managers with at least 3 tournament outcomes. The 8 lag model dropping early tournaments is estimated on the sample of managers with at least 16 tournament outcomes, dropping the first 8 tournaments for the purpose of estimating the model. The model for current store only uses outcomes from the store that a manager had as of Q4 of 2015 to estimate the model, restricted to managers who have three or more consecutive outcomes from that store. The model for recent quarters is a 3 lag model estimated using only Q3, Q2, and Q1 of 2015. The model excluding Q3 is estimated with 3 lags using all tournament outcomes except for Q3 of 2015. The model using nationwide tournaments is a 3 lag model that excludes outcomes from quarters with regional tournaments.

Table F2: Multinomial logit coefficient estimates I

	Overconfident priors								Manager non-stationarity							
	Tenure ≥ 8 quarters Performance quintile in t				Tenure ≥ 3 quarters Performance quintile in t				Tenure ≥ 8, drop early signals Performance quintile in t				Tenure ≥ 3, current store only Performance quintile in t			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Performance percentile in t-1	-0.10***	-0.05***	0.06***	0.09***	-0.11***	-0.04***	0.05***	0.10***	-0.11***	-0.05***	0.08***	0.09***	-0.10***	-0.04***	0.04***	0.12***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Performance percentile in t-2	-0.05***	-0.02*	-0.00	0.07***	-0.04***	-0.02***	0.01	0.06***	-0.05***	-0.01	-0.00	0.06***	-0.04***	-0.02	0.00	0.06***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Performance percentile in t-3	-0.01	0.00	0.01	-0.01	-0.02**	0.00	0.01	0.00	-0.01	0.01	0.01	-0.01	-0.03*	-0.01	0.02	-0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)
Performance percentile in t-4	-0.02	0.01	-0.01	-0.00					-0.01	0.00	-0.01	0.01				
	(0.01)	(0.01)	(0.01)	(0.01)					(0.01)	(0.01)	(0.01)	(0.01)				
Performance percentile in t-5	-0.00	0.00	-0.01	0.00					-0.00	0.00	-0.02	0.00				
	(0.01)	(0.01)	(0.01)	(0.01)					(0.01)	(0.01)	(0.01)	(0.01)				
Performance percentile in t-6	-0.00	-0.01	0.00	0.01					0.01	-0.01	0.01	-0.00				
	(0.01)	(0.01)	(0.01)	(0.01)					(0.01)	(0.01)	(0.01)	(0.01)				
Performance percentile in t-7	-0.00	-0.01	0.01	-0.00					-0.02	0.01	-0.00	-0.00				
	(0.01)	(0.01)	(0.01)	(0.01)					(0.01)	(0.01)	(0.01)	(0.01)				
Performance percentile in t-8	-0.01	-0.00	0.00	0.02**					-0.00	-0.01	0.01	0.02**				
	(0.01)	(0.01)	(0.01)	(0.01)					(0.01)	(0.01)	(0.01)	(0.01)				
Observations	1744				3568				1272				891			
Pseudo R^2	0.106				0.087				0.105				0.096			

Notes: The estimations use historical data from Q3 of 2015 back to Q1 of 2008 unless otherwise noted. The table reports marginal effects from multinomial logit regressions. Independent variables are standardized so the coefficients show the change in the probability of achieving a given quintile in period t associated with a 1 s.d. increase in percentile of performance. The base category is quintile 3. Columns (1) to (4) report results for the 8 lag model that was selected over models with fewer lags in cross validation; it entails using a sample of relatively experienced managers, those with at least 8 consecutive tournament outcomes. Columns (5) to (8) report results of a three 3 lag model that uses a larger sample that includes all managers with at least 3 tournament outcomes. Columns (9) to (12) report a model estimated on the sample of managers with at least 16 tournament outcomes, dropping the first 8 tournaments for the purpose of estimating the model. are based on (experienced) managers who have at least 16 outcomes, but dropping the first 8. Columns (13) to (16) reports results from a model that only uses outcomes from the store that a manager had as of Q4 of 2015 to estimate the model, restricted to managers who have three or more consecutive outcomes from that store. Robust standard errors are in parentheses, clustering on manager.

Table F3: Multinomial logit coefficient estimates II

	Environment non-stationarity				Imperfect knowledge							
	Recent quarters Performance quintile in t				Excluding Q3 of 2015 Performance quintile in t				National tournaments only Performance quintile in t			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Performance percentile in t-1	-0.10***	-0.03**	0.06***	0.10***	-0.11***	-0.04***	0.05***	0.10***	-0.09***	-0.03**	0.04***	0.11***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Performance percentile in t-2	-0.05***	-0.01	0.01	0.05***	-0.04***	-0.02***	0.01	0.06***	-0.08***	-0.02	0.02	0.06***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Performance percentile in t-3	-0.04**	-0.02	0.01	0.01	-0.02**	0.01	0.01	0.00	-0.03**	0.00	0.01	-0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	667				3391				1042			
Pseudo R^2	0.09				0.11				0.11			

Notes: The estimations use historical data from Q3 of 2015 back to Q1 of 2008 unless otherwise noted. The table reports marginal effects from multinomial logit regressions. Independent variables are standardized so the coefficients show the change in the probability of achieving a given quintile in period t associated with a 1 s.d. increase in percentile of performance. The base category is quintile 3. Columns (1) to (4) report results for a 3 lag model estimated using only Q3, Q2, and Q1 of 2015. Columns (5) to (8) is estimated with 3 lags using all tournament outcomes except for Q3 of 2015. Columns (9) to (12) report results of a 3 lag model that excludes outcomes from quarters with regional tournaments. Robust standard errors are in parentheses, clustering on manager.

Table F4: Summary of robustness checks on manager predictions vs. multinomial logit predictors with non-parametric specifications

	Manager vs. multinomial logit predictors					
	Fraction of managers:			P-values		
	overconfident	accurate	underconfident	different	frac. overconf. - frac. underconf.	N
Overconfident priors:						
8 lag	0.43	0.35	0.22	p<0.01	p<0.01	1744
3 lag	0.43	0.31	0.26	p<0.01	p<0.01	3568
Manager non-stationarity:						
8 lag, drop early	0.43	0.36	0.21	p<0.01	p<0.01	1272
3 lag, current store only	0.40	0.32	0.28	p<0.01	p<0.01	891
Environment non-stationarity:						
3 lag, recent tournaments	0.44	0.29	0.27	p<0.01	p<0.01	667
Imperfect knowledge:						
Excluding Q3 tournament	0.43	0.33	0.24	p<0.01	p<0.01	3391
Nationwide tournaments	0.41	0.30	0.29	p<0.01	p<0.01	1042

Notes: The specifications include separate dummy variables for quintiles 1 to 4 in a given quarter, rather than the more parametric linear specification used in the main set of multinomial logit estimations. The estimations use historical data from Q3 of 2015 back to Q1 of 2008 unless otherwise noted. P-values test whether manager predictions are different from the model predictions, and whether they are more skewed towards overconfidence. See text for details on bootstrapping. The 8 lag model entails using a sample of relatively experienced managers, those with at least 8 consecutive tournament outcomes. The 3 lag model uses a larger sample that includes all managers with at least 3 tournament outcomes. The 8 lag model dropping early tournaments is estimated on the sample of managers with at least 16 tournament outcomes, dropping the first 8 tournaments for the purpose of estimating the model. The model for current store only uses outcomes from the store that a manager had as of Q4 of 2015 to estimate the model, restricted to managers who have three or more consecutive outcomes from that store. The model for recent quarters is a 3 lag model estimated using only Q3, Q2, and Q1 of 2015. The model excluding Q3 is estimated with 3 lags using all tournament outcomes except for Q3 of 2015. The model using nationwide tournaments is a 3 lag model that excludes outcomes from quarters with regional tournaments.

G Stationarity of transition matrices Z_t

This section sheds light on whether the informativeness of tournament outcomes, and the decision environment facing managers, has been stable over time. If the environment were non-stationary, for example because turnover led to changes in the composition of types of managers over time, this would be reflected in changes in the informativeness of tournament outcomes, captured in the transition matrixes between and two quarters, Z_t. To see this, suppose that the pool of managers becomes more homogeneous over time in terms of ability. This would lead to greater randomness in tournament outcomes, and declining correlations of tournament outcomes from one quarter to the next. Other sources of non-stationarity could be some sort of unobserved changes in policies or production function of the firm. The table below shows that there is little evidence that the correlation structure of tournament outcomes across quarters is changing over time. See the table notes for more details.

Table G1: Stationarity of environment: Test of time trends in Z_t

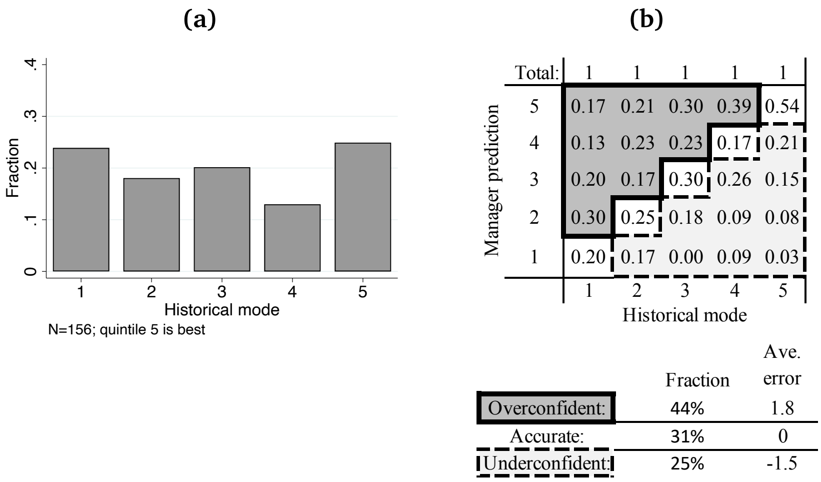
			Signal			
	1	2	3	4	5	Total
Type 1	0.657	0.248	0.044	0.044	0.008	1
Type 2	0.260	0.304	0.297	0.118	0.020	1
Type 3	0.055	0.293	0.380	0.180	0.093	1
Type 4	0.021	0.105	0.160	0.479	0.236	1
Type 5	0.008	0.050	0.120	0.180	0.642	1
Total	1	1	1	1	1	

Notes: The table shows results of regressing each element of Z_t on a constant and t. It displays only the coefficient on t. Some of the individual coefficients are significant when considered individually. However, we are testing 25 hypotheses at the same time. Using the standard Boneferroni correction, one rejects a null hypothesis at level α only if the p-value is less than $\frac{\alpha}{\zeta}$ where $\zeta = 25$ is the number of hypotheses that are being tested. With the correction, none of the test statistics are significant at $\alpha = .01$ and only one, $Z_{1,1}$, at $\alpha = .05$. *** indicates significance at .01, ** significance at .05, * at .1 for each test individually. Standard errors in parentheses.

H Robustness checks for manager predictions vs. rule of thumb predictors

This section reports results on whether managers are overconfident relative to rule of thumb predictors. Panel (a) of Figure H1 shows the distribution of historical modes for managers, and Panel (b) compares manager predictions about the modal quintile compared to predictions they should have made if they had bet on their historical modal quintile. We see that 44 percent predict a higher quintile for Q4 of 2015 than their most frequent quintile in the past, whereas only 25 percent predict a lower quintile. Thus, we see a pattern of manager overconfidence relative to this rule of thumb.

Figure H1: Distribution of historical modes and comparison to manager predictions



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

Table H1 summarizes results from a range of different rules of thumb that involve different assumptions about manager priors, or the optimal way to combine past outcomes, or the knowledge of managers about outcomes. See Section 3.2 for more information on the rationales for the different rules of thumb. Details on the construction of the predictors are provided in the table notes. The results show that managers are consistently overconfident, regardless of which rule of thumb is used.

I Predictions and prediction errors of managers as a function of experience

This section provides information about how predictions and prediction errors of managers vary with experience. To shed some light on the nature of manager priors when

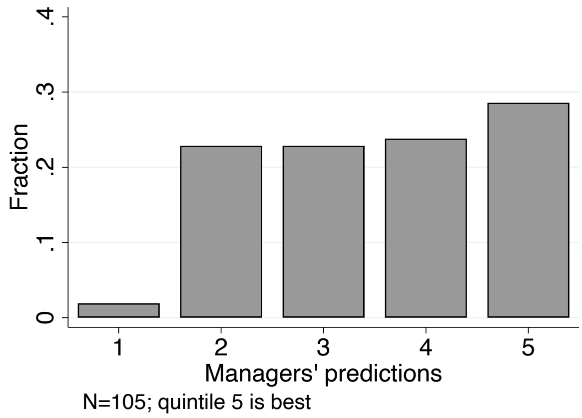
Table H1: Summary of robustness checks on manager predictions vs. rule of thumb predictors

	Manager predictions vs. rule of thumb predictors			
	Fraction of managers:			
	overconfident	accurate	underconfident	N
Overconfident priors:				
Historical mode	0.44	0.31	0.25	156
Historical mode, experienced only	0.42	0.31	0.26	106
Manager non-stationarity:				
Historical mode, experienced, drop early	0.42	0.35	0.23	85
Historical mode, current store only	0.46	0.30	0.24	137
Environment non-stationarity:				
Historical mode, recent quarters only	0.41	0.33	026	126
Imperfect knowledge:				
Historical mode, excluding Q3	0.43	0.30	0.27	148
Historical mode, nationwide tournaments	0.46	0.32	0.22	139
Non-unique mode:				
Historical mode, max	0.39	0.32	0.29	202
Historical mode, min	0.52	0.28	0.20	202

Notes: The historical mode is a manager's most frequent quintile outcome from quarters before Q4 of 2015 (dropping managers with non-unique modes). The mode for experienced managers includes only those managers with more than 2 years of experience. The mode for experienced managers, dropping early signals, is for managers with more than 2 years of experience and calculates the manager's mode after dropping the manager's first 8 tournament outcomes. The mode for the current store is calculated using only tournament outcomes from the store that the manager operated as of Q4 of 2015. The mode for recent quarters is calculated using only Q3, Q2, and Q1 of 2015. The historical mode excluding Q3 of 2015 uses only earlier quarters to calculate the mode. The historical mode for nationwide tournaments is calculated using only those quarters in which there was a nationwide tournament. The max and min versions of the mode give managers with non-unique modes the max or min of the set of candidate modes, respectively.

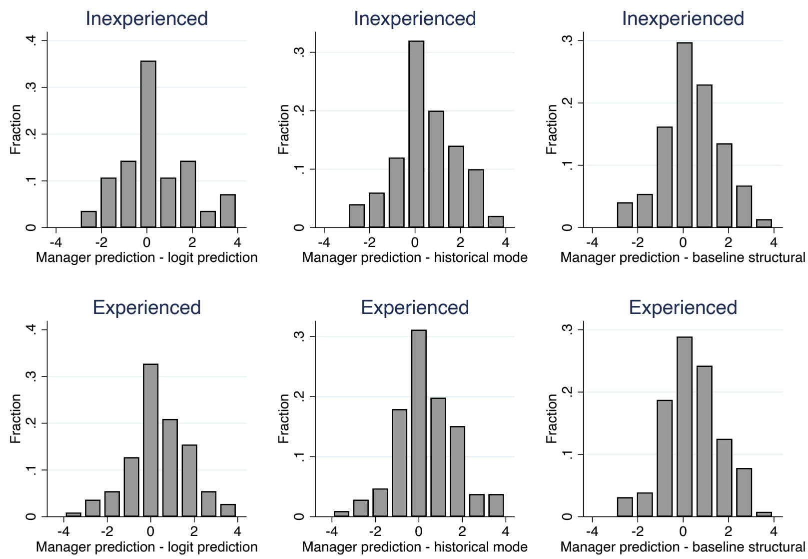
starting the job, Figure I1 shows predictions of recently hired managers. These are similar to those for the entire sample: skewed away from predicting the lowest quintile and towards predicting the best quintile. To explore whether the magnitude of ex-ante prediction errors shrinks with experience, consistent with managers having overconfident priors initially, but learning in a Bayesian way from subsequent tournament outcomes, we compare managers with substantial experience of at least two years, to managers with less experience. Figure I2 shows that magnitudes of ex-ante prediction errors are similar for inexperienced and experienced managers, regardless of which benchmark prediction model is used(see Figure I1 and Figure I2).

Figure I1: Distribution of predictions for recently hired managers, less than 1 year of experience



Notes: The figure reports the distribution of manager predictions about Q4 of 2015 for managers with less than 1 year of tenure.

Figure I2: Manager predictions vs. predictions based on tournament outcomes as a function of experience



Notes: The figure reports the differences between manager predictions about Q4 of 2015 and the respective predictors based on histories of tournament outcomes: Multinomial Logit model with 3 lags; historical mode; baseline structural model. Experienced is defined by having at least 2 years of experience at the firm, inexperienced by having less than 2 years.

We also check robustness to an explanation based on changing informativeness of tournament outcomes over the course of manager careers. If tournament outcomes for some reason become more random as managers gain experience, this could be a factor that retards learning, and conceivably leads to the preservation of rationally overconfident priors without a role for motivated beliefs. We find, however, that the average variance of performance is similar for early years in manager careers (less than or equal to 2 years) compared to later years (beyond first 2 years), 0.050 versus 0.055. One might be concerned about attrition, since some managers who contribute to the first calculation leave the firm before they can contribute to the second calculation (note however that we find little evidence of differential attrition based on manager overconfidence). Accordingly, we consider the subsample of managers who ultimately work at least four years, and compare variance for the first two years to the second two years. In this case variances are almost identical, 0.0.582 versus 0.0581.

To get more directly at the question whether tournament outcomes are becoming less informative with experience we also look at the predictive power of tournament outcomes as captured by transition matrixes, and find that these are similar for managers with less and more experience, casting doubt on this alternative explanation (Tables I1 and I2). This is also true if we eliminate possible confounds related to attrition by focusing on the subsample of managers who stay at least four years (Tables I3 and I4). To assess statistical significance of possible experience effects on informativeness of tournament outcomes we also estimated our baseline multinomial logit model with 8 lags of past performance, including a control for experience, and also interaction terms of experience with each of the lagged performances. The interaction terms are generally not statistically significant (only two of the 20 coefficients are significant), and those that are significant are in the direction of tournament outcomes becoming slightly more informative with experience (results available upon request). This is the opposite of what would be needed for the alternative explanation for lack of learning. We find similar results if we rule out attrition confounds by restricting to the subsample of managers who work for at least four years, and use interactions of lagged performance with cumulative experience, conditioning on cumulative experience being less than or equal to four years.

Table I1: Average quintile-to-quintile transition matrix $\hat{Z}$ for first two years of experience

	Fractions of managers				
	Quintile in t				
Quintile in $t-1$:	1	2	3	4	5
5	0.06	0.11	0.15	0.26	0.42
4	0.09	0.20	0.19	0.27	0.25
3	0.18	0.22	0.26	0.20	0.14
2	0.23	0.25	0.22	0.19	0.11
1	0.42	0.21	0.22	0.10	0.06

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 only quarters that come earlier in manager careers, specifically, in the first two years on the job.

Table I2: Average quintile-to-quintile transition matrix $\hat{Z}$ after first two years of experience

	Fractions of managers				
	Quintile in t				
Quintile in $t-1$:	1	2	3	4	5
5	0.04	0.11	0.16	0.27	0.43
4	0.13	0.15	0.23	0.27	0.21
3	0.13	0.24	0.29	0.21	0.13
2	0.25	0.27	0.24	0.15	0.09
1	0.46	0.26	0.18	0.08	0.03

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 only quarters come later in manager careers, specifically, after the first two years of experience.

Table I3: Average quintile-to-quintile transition matrix $\hat{Z}$ for years 1 and 2 of experience conditional on working at least 4 years

	Fractions of managers				
	Quintile in t				
Quintile in $t-1$:	1	2	3	4	5
5	0.05	0.10	0.11	0.26	0.47
4	0.05	0.18	0.19	0.29	0.28
3	0.14	0.19	0.28	0.24	0.14
2	0.24	0.23	0.22	0.22	0.09
1	0.45	0.24	0.15	0.11	0.05

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 only quarters that come from the first two years of manager careers, for the subsample of managers who ultimately work for more than four years.

Table I4: Average quintile-to-quintile transition matrix $\hat{Z}$ for years 3 and 4 of experience conditional on working at least 4 years

	Fractions of managers				
	Quintile in t				
Quintile in $t-1$:	1	2	3	4	5
5	0.03	0.11	0.16	0.29	0.40
4	0.11	0.17	0.21	0.30	0.21
3	0.13	0.21	0.32	0.20	0.14
2	0.21	0.28	0.21	0.15	0.10
1	0.46	0.26	0.19	0.06	0.02

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 only quarters that come from years three and four in managers careers, for the subsample of managers who ultimately work for more than four years.

J Robustness checks on determinants of biased memory

In a motivated beliefs explanation for our results, a deciding factor for whether a manager mis-remembers, and what they remember, should be the actual Q2 performance. In line with this explanation, the regression analysis discussed in the text (Table 2) shows that having Q2 performance below the very top of the ranking is associated with a significantly higher probability of mis-remembering and that recall errors are skewed towards being overly positive (motivated beliefs), but that memories are nevertheless significantly related to actual Q2 (reality constraints). In this section, we explore whether these conclusions are robust to including additional controls, or accounting for floor effects. It is also potentially of independent interest to explore what factors or traits might be related to having accurate memory.

In terms of robustness of results in Table 2 to additional controls, we explore various factors that might potentially help explain the probability that a manager misremembers Q2 performance. (1) *Deviation from the mean*: The extent to which Q2 performance differs from a manager's average performance might make the outcome memorable, or it might cause a manager to discount the outcome when forming predictions. (2) *Deviation from the median*: Deviation from the median is an alternative metric for whether Q2 was atypical. (3) *Variance:* A manager might see less of a value of remembering the outcome of a particular quarter if his or her performance has a high variance. (4) *Elapsed time*: Laboratory evidence suggests that forgetting negative feedback takes some time (Zimmerman, 2020), so we look at the elapsed days between the end of Q2 and the date of eliciting the memory of Q2. (5) *Valuing financial incentives*: Effort to recall accurately might be related to caring about the magnitude of financial incentives we offer for accuracy; we control for how many addition problems the manager solved in an incentivized task, with similar magnitude of incentives for a correct answer. (6) *Attentiveness*: If recall errors reflect a trait of inattentiveness rather than motivated beliefs, we might expect inaccurate recall to be related to inattention to details of the firm's incentive scheme; as a measure of attentiveness we use an indicator for knowledge about the max. and min. values of the scale used by the firm to score relative performance on the performance dimensions. (7) *Cognitive ability*: Lack of understanding of the incentive scheme might be an indicator for low cognitive ability, which could be related to memory accuracy; we use an indicator for whether the manager understands the implications of the multiplicative nature of the incentive scheme.[5] (8) *Tendency to exaggerate the truth*: To check whether stated overly positive

[5]Understanding means knowing that performance will be ranked higher if performance is equal across all four dimensions, compared to having unequal performances with the same mean across dimensions (mean-preserving spread).

memories might reflect a tendency for managers to exaggerate, in spite of potential embarrassment, and the financial incentives, we control for a (noisy) measure of willingness to mis-report: The private die role a subject reported, in an incentivized task where rolling higher numbers generates higher payments. (9) *Additional traits*: Memory accuracy might conceivably be related to other manager traits in our data: Willingness to take risks, self-reported risk attitudes, patience, competitiveness, and confidence. For more details on the control variables and measures of manager traits see Appendix E.

The corresponding results are shown in Columns (1) to (8) of Table J1. The additional controls are by and large not significantly related to the probability of mis-remembering, whereas actual Q2 performance continues to be statistically significant. One exception is manager experience, but the coefficients suggest a rather small improvement in the probability of being accurate, around 0.08, accuracy associated with a substantial 3.7 year (1 s.d.) increase in experience, and this is no longer significant once all controls are added. The lack of significant coefficients for other factors does not prove that these do not matter for memory, but only that we cannot detect such effects given our sample. Notably, the lack of a relationship to elapsed time is perhaps unsurprising given that the shortest time is already more than one month; Zimmerman (2020) finds in the lab that memories of negative feedback are already suppressed after the passage of one month's time.

Table J1: Inaccurate memory as a function of actual Q2 performance and additional controls

	Inaccurate memory							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Performance percentile in Q2 of 2015	-0.12***	-0.12***	-0.12***	-0.13***	-0.12***	-0.10**	-0.12***	-0.07*
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Performance percentile in Q3 of 2015	0.01	-0.01	0.02	0.02	0.02	0.01	0.03	-0.01
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Mean performance percentile pre- Q2 of 2015	-0.02	-0.00	-0.02	-0.02	-0.02	-0.02	-0.03	-0.01
	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Female	0.04	0.04	0.04	0.04	0.04	0.06	0.01	-0.04
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)	(0.07)
Age	0.02	-0.00	0.02	0.02	0.02	-0.00	0.01	-0.02
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
Experience	-0.08**	-0.07**	-0.08**	-0.08**	-0.08**	-0.06	-0.07**	-0.04
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)
Abs. dev. of Q2 from historical mean percentile	0.02							-0.17**
	(0.03)							(0.08)
Abs. dev. of Q2 from historical median percentile		0.05						0.25***
		(0.04)						(0.08)
Variance of historical performance percentiles			0.02					0.01
			(0.03)					(0.05)
Days elapsed between Q2 and memory measurement				-0.02				0.05
				(0.03)				(0.03)
Addition problems solved in incentivized task					0.01			0.00
					(0.03)			(0.03)
Knows scale values for incentive scheme						-0.03		-0.02
						(0.07)		(0.06)
Understands implications of mult. incentive scheme						-0.01		0.02
						(0.08)		(0.08)
Risk all in incentivized measure							-0.02	-0.02
							(0.02)	(0.02)
Die roll in incentivized lying task							0.03	0.01
							(0.03)	(0.03)
Self-assessed willingess to take risks							0.07**	0.07**
							(0.03)	(0.03)
Self-assessed competitiveness							-0.03	-0.04
							(0.03)	(0.03)
Self-assessed relative confidence							0.00	-0.00
							(0.03)	(0.03)
Self-assessed patience							0.03	0.01
							(0.03)	(0.03)
Observations	149	138	149	149	149	120	147	112
Pseudo R^2	0.126	0.141	0.126	0.125	0.123	0.087	0.171	0.226

Notes: All columns report marginal effects from probit regressions. The dependent variable 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). Independent variables are standardized, so coefficients give the change in the probability of mis-remembering associated with a 1 standard deviation 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. Knowledge of scale values is an indicator for whether the manager knows the min. and max. possible values for the scale to evaluate performance on individual dimensions. Understanding of the multiplicative nature of the scheme is an indicator for whether the manager understands that performance will be ranked higher if performance is equal across all dimensions, compared to having unequal performances with the same mean across dimensions (mean-preserving spread). Risk taking in the incentivized task is how much money the manager invested in a risky rather than a safe asset. Reporting a higher die roll is a (noisy) indicator of willingness to exaggerate. Self-assessments are on an 11-point scale, with higher values indicating greater willingness to take risks, etc.. Robust standard errors are in parentheses.

In terms of potential determinants of the specific rank that a manager remembers for Q2, we explored whether managers might construct memories based on various moments of the distribution of past performance besides the mean – mode, median, variance, maximum, and minimum – and whether memories might be related to other manager traits. As shown in Table J2, these are largely unrelated to the remembered performance, or the probability of having an overly positive memory. One exception is manager experience, where greater experience is associated with recalling lower performances. This translates into a modest reduction in the probability of overly positive memories, and an increase in the probability of overly negative memories.[6] In all regressions, actual Q2 performance continues to be a statistically significant explanatory factor for what a manager remembers.

A concern could be that a combination of two factors might explain the finding of asymmetric recall errors for managers with worse Q2 performance, rather than motivated beliefs: (1) Errors more likely for managers with worse Q2 performance because worse performance indicates low cognitive ability; (2) floor effects for managers at the bottom of the performance distribution mean that errors must be in the better-than-actual direction. In combination, this could generate a pattern of errors for managers with worse performances, and asymmetric recall errors.

We find several pieces of evidence against (1). One is the fact that Q2 of 2015 performance matters in a similar way for explaining recall errors about Q2 of 2015 even if we control for manager ability using summary statistics like average pre-Q2 performance, seemingly a better proxy for cognitive ability than just one quarter (see Table 2). This is reinforced by the robustness checks discussed above, in which results are similar adding a range of other types of controls for manager past performance and traits, including proxies for attentiveness and cognitive ability. These findings suggest that Q2 of 2015 performance has an impact on recalled performance for Q2 of 2015, which is not explainable by manager cognitive ability.

[6] We speculate about one possible explanation for this time trend in memory, which is that managers might have a greater need to need to constantly maintain overly positive memories when they are newly on the job, to implement a strong posterior of being a good type. Once they are "secure" in their beliefs, however, they might not need to work as hard to maintain overly-positive memories, at least for a while. If non-distorted memory is still subject to some noise in recollection, these relatively experienced managers may still have noisy recall, but with more symmetric recall errors.

Table J2: Recalled Q2 performance as a function of actual Q2 performance and additional controls

	Recalled Q2 per. percentile		Flattering mem.		Unflattering mem.	
	(1)	(2)	(3)	(4)	(5)	(6)
Performance percentile in Q2 of 2015	0.42***	0.41***	-0.17**	-0.17**	0.08	0.11**
	(0.12)	(0.13)	(0.07)	(0.07)	(0.06)	(0.06)
Performance percentile in Q3 of 2015	0.12	0.14	0.04	0.08	-0.01	0.01
	(0.09)	(0.10)	(0.06)	(0.06)	(0.05)	(0.05)
Mean performance percentile pre- Q2 of 2015	-0.18	-0.29	-0.17	-0.21	0.18*	0.34***
	(0.18)	(0.19)	(0.12)	(0.13)	(0.10)	(0.11)
Female	0.17	0.18	0.12	0.11	-0.06	-0.05
	(0.14)	(0.18)	(0.09)	(0.10)	(0.08)	(0.08)
Age	0.07	0.02	0.03	-0.01	0.01	0.02
	(0.10)	(0.12)	(0.05)	(0.05)	(0.05)	(0.05)
Experience	-0.28**	-0.33**	-0.13*	-0.13*	0.07	0.12*
	(0.12)	(0.13)	(0.07)	(0.08)	(0.06)	(0.07)
Maximum historical performance percentile	0.23	0.22	0.09	0.09	-0.16	-0.22**
	(0.20)	(0.21)	(0.11)	(0.12)	(0.10)	(0.10)
Minimum historical performance percentile	0.01	-0.03	0.05	0.05	-0.07	-0.11
	(0.11)	(0.14)	(0.09)	(0.10)	(0.08)	(0.09)
Modal historical performance quintile	0.06	0.03	0.06	0.03	-0.09	-0.07
	(0.17)	(0.20)	(0.10)	(0.12)	(0.09)	(0.09)
Median historical performance percentile	0.10	0.23	0.12	0.13	-0.10	-0.21
	(0.26)	(0.28)	(0.16)	(0.16)	(0.14)	(0.14)
Variance of historical performance percentiles	-0.06	-0.09	0.01	0.01	0.06	0.06
	(0.15)	(0.17)	(0.09)	(0.10)	(0.08)	(0.08)
Days elapsed between Q2 and memory measurement		-0.08		0.00		0.09**
		(0.08)		(0.05)		(0.04)
Addition problems solved in incentivized task		-0.01		-0.01		0.02
		(0.08)		(0.05)		(0.05)
Knows scale values for incentive scheme		0.10		-0.04		-0.06
		(0.15)		(0.10)		(0.09)
Understands implications of mult. incentive scheme		-0.01		0.10		-0.10
		(0.19)		(0.12)		(0.10)
Risk all in incentivized measure		-0.03		-0.05		0.05*
		(0.06)		(0.03)		(0.03)
Die roll in incentivized lying task		-0.02		0.02		0.05
		(0.09)		(0.05)		(0.05)
Self-assessed willingess to take risks		-0.12		-0.02		0.09**
		(0.10)		(0.06)		(0.04)
Self-assessed competitiveness		0.05		0.05		-0.07
		(0.09)		(0.06)		(0.04)
Self-assessed relative confidence		-0.01		-0.02		0.04
		(0.09)		(0.05)		(0.04)
Self-assessed patience		0.10		0.05		-0.00
		(0.06)		(0.05)		(0.05)
Constant	-0.13	-0.12				
	(0.11)	(0.14)				
Observations	121	97	121	97	121	97
R^2	0.450	0.517				
Pseudo R^2			0.077	0.152	0.098	0.215

Notes: Columns (1) and (2) report OLS estimates and the dependent variable is the standardized recalled performance percentile for Q2. Columns (3) and (4) report marginal effects from probit regressions, and the dependent variable is an indicator for having an overly positive memory of Q2 performance by more than 10 ranks (the elicitation gave incentives to be accurate within a range of +/- 10 ranks). Columns (5) and (6) report marginal effects from probit regressions, and the dependent variable is an indicator for having an overly negative memory of Q2 performance by more than 10 ranks (the elicitation gave incentives to be accurate within a range of +/- 10 ranks). Independent variables are standardized, so coefficients give the change in the dependent variable associated with a 1 standard deviation 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 estimation sample only includes managers with a unique historical mode. Knowledge of scale values is an indicator for whether the manager knows both the min. and max. possible values for the scale to evaluate performance on individual dimensions. Understanding of the multiplicative nature of the scheme is an indicator for whether the manager understands that performance will be ranked higher if performance is equal across all dimensions, compared to having unequal performances with the same mean across dimensions (mean-preserving spread). Risk taking in the incentivized task is how much money the manager invested in a risky rather than a safe asset. Reporting a higher die roll is a (noisy) indicator of willingness to exaggerate. Self-assessments are on an 11-point scale, with higher values indicating greater willingness to take risks, etc.. Robust standard errors are in parentheses.

Regarding (2), the floor effect part of the concern, Figure 3 in the text casts doubt on this, showing that recall errors are strongly asymmetric in the middle of the performance distribution, where there is equal room to have errors in either direction. Thus, the asymmetry is not driven by managers at the bottom of the performance distribution, as it would have to be if due only to floor effects. We check this more rigorously in Table J3 below, where we estimate the same regression specifications as in Columns (7) and (8) of Table 2 in the text, but excluding managers at the extremes. Specifically, Columns (1) and (2) of Table J3 show that recall errors are strongly asymmetric if we only estimate the regression using managers in the middle three quintiles of the Q2 performance distribution, and Columns (3) and (4) show that results are very similar if we just exclude managers in the worst quintile, i.e., those closest to the floor. Floor effects thus do not appear to be crucial for the asymmetry in recall errors.

Table J3: Inaccurate memory and recall errors as a function of actual Q2 performance, excluding boundary effects

	Recalled - actual performance			
	(1)	(2)	(3)	(4)
Inaccurate memory	43.73***	39.07***	33.36***	35.02***
	(7.15)	(10.34)	(6.53)	(7.83)
Performance percentile in Q3 of 2015		15.30**		7.75
		(6.04)		(5.16)
Mean performance percentile pre- Q2 of 2015		1.24		-0.40
		(6.17)		(4.95)
Female		3.95		3.90
		(14.11)		(11.33)
Age		-1.94		-1.04
		(8.04)		(7.02)
Experience		-6.32		-7.12
		(10.30)		(8.89)
Constant	-1.33	-3.19	0.39	-5.84
	(1.15)	(8.53)	(0.86)	(7.23)
Observations	114	100	149	126
Adjusted R^2	0.052	0.064	0.042	0.042

Notes: Columns (1) to (4) report OLS estimates. The dependent variable 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. The estimation sample for Columns (1) and (2) excludes managers who were in the best and worst quintiles of performance in Q2 of 2015, to check robustness to eliminating managers near to the boundaries. The estimation sample for Columns (3) and (4) excludes managers who were in the worst quintile of performance in Q2 of 2015. 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.

K Analysis of measures of manager memories about sub-metrics determining rank

In this section we analyze additional memory measures included in the lab in the field study, which asked managers to remember different sub-metrics from Q2 of 2015 that determined overall rank (for details on wording and format see Part 10 of the instructions provided in Appendix U). Although less directly related to the question of how managers sustain overconfidence about future rank, analyzing these can provide another type of robustness check, allowing us to check whether there is evidence of asymmetric recall errors for other performance metrics besides just rank. They can also shed some light on an additional comparative static of motivated beliefs models, that biased memory should be less pronounced for metrics that are less diagnostic of own performance. At the end of the section we also briefly discuss some other questions included in the study, which asked managers try recall other types of outcomes from Q2 of 2015.

An interesting comparative static of some models of motivated beliefs is that individuals might have more biased memories about performance metrics that are more tightly linked to personal rather than group success. This can arise if individuals are mainly motivated to have positive beliefs about the self, as opposed to positive beliefs outcomes in general (general optimism). Our setting offers some potential ways to shed light on this hypothesis, because performance is measured with different metrics that vary in terms of much they depend on the performance of the manager's own individual store versus an average across a group of stores.

To investigate this hypothesis we consider three performance metrics. The Base Bonus (BB) is the core metric of the incentive scheme, which involves ranking the manager's own store relative to all other stores on each of the four performance dimensions (customer service, regional manager evaluation, profit, and sales growth), assigning a numerical score based on where it falls in the distribution, and then multiplying these four scores together.[7] A second metric, the Area Bonus (AB), is group-based; it is constructed in the same way as the BB, except that it averages the performance of the manager's store with other stores in the near geographic area for each of the four dimensions, assigns scores based on how the area ranks relative to other areas, and then multiplies the four scores (in Q2 of 2015 the average area had about 8 stores). A third metric, the Final Bonus (FB), has both individual and group-based components; it is the product of the BB and AB, except for roughly the top ten percent of managers who receive additional Top Performer multipliers that increase their scores (the FB score is the basis of assigning a manager's ultimate rank in the tournament, with differences

[7]The numerical scores range from 0.65 to 1.2 for each dimension.

arising due to tie-breaking).[8] A manager's BB, FB, and rank are thus metrics that are less group-based than the AB.[9]

The study asked managers to remember their FB, and their AB, and we can infer a recalled BB because the study asked managers to recall their scores on each of the four dimensions; multiplying these gives the recalled BB. A caveat is that a relatively high fraction of managers, about 14 percent on average, give infeasible values for some of these latter questions about the individual dimensions, as well as the AB, e.g., recalling values that are outside the support of possible values; this almost never happens for FB or rank.[10] We interpret this as some managers either misunderstanding the question, or not having not paid attention to the exact numerical values that the firm uses for sub-components of the incentive system, or both. For the cleanest comparison of the nature of recall errors across different performance metrics we exclude managers who give such responses for any of the measures, which ensures that we compare the same sample of people across all of the measures.[11]

Table K1 provides results on the frequencies of different types of recall errors about BB, AB, and FB; for comparison the table also presents results on recalled rank. A first set of observations is that, for all of the performance metrics, we see a similar pattern: (1) recall errors in the flattering direction are substantially more frequent than recall errors in the other direction; (2) errors tend to be larger in the flattering than the unflattering direction; (3) the average error is significantly different from zero in the flattering direction (the same patterns are also seen for each of the individual dimensions underlying the Base Bonus Score).[12] As shown in the bottom panel of the table, results are similar

[8]In Q2 of 2015 there were 17 managers who received an extra bonus for being at the very top of the performance distribution.

[9]All of the performance metrics depend to some extent on the performance of others, e.g., the workers in a manager's store contribute to the BB, FB, and rank, but the AB is different because it also depends on performances of other managers and stores.

[10]The rates of recalling such values is about 1 percent for FB and rank. One reason for such responses on the individual dimensions appears to be managers giving their memory about their raw performance result instead of their band, e.g., reporting a number -0.01 that could make sense as a sales growth rate, rather than reporting their band value of 0.65 assigned based on relative sales growth rate comparison. This could indicate that the manager misunderstood the question, or more plausibly in our view, could indicate that they felt very uncertain about their band, and more uncertain about the raw performance, and reported the latter to show that they knew something.

[11]We think it makes sense to exclude these responses because they potentially reflect managers answering a different question than what was asked. We do include managers giving values that are within the support of possible values but slightly different from any of the values assigned to bands, e.g., the manager recalls 0.67 but the nearest feasible values are 0.65 or 0.75. We treat this as a source of noise coming from managers trying to answer the question but not being fully informed about the details of the bands, with the noise working against finding statistically significant results about, e.g., the asymmetry of recall errors. We are also conservative in classifying recall errors, counting manager recall as being correct if it is within a window of +/- .1 around the true value.

[12]The average error is also statistically significantly different from zero in the flattering direction for each of the dimensions, except for profit and sales.

if we do not restrict the sample to be constant across measures. These findings indicate that managers tend to have overly-positive memories about performance metrics in general, not just rank (if anything the asymmetry is most pronounced for recalling FB). This is consistent with a pervasive influence of motivated beliefs on memories of various metrics of past performance.

Table K1: Summary of manager recollections of Q2 of 2015 performance metrics

	Manager recollection vs. actual score						
	Fraction of error type:			Average size of error		T-Test for	
	flattering	accurate	unflattering	flatt. error (in s.d.)	unflatt. error (in s.d.)	ave. error = 0 (p-value)	N
Sample held constant:							
Recalled Base Bonus Score	0.43	0.43	0.14	0.93	-0.52	$p < 0.01$	73
Recalled Area Bonus Score	0.30	0.59	0.11	0.97	-0.70	$p < 0.01$	73
Recalled Final Bonus Score	0.62	0.34	0.04	1.03	-0.18	$p < 0.01$	73
Recalled Rank	0.59	0.22	0.19	1.10	-0.44	$p < 0.01$	73
Sample not held constant:							
Recalled Base Bonus Score	0.40	0.43	0.17	0.91	-0.55	$p < 0.02$	91
Recalled Area Bonus Score	0.31	0.55	0.14	0.86	-0.67	$p < 0.01$	119
Recalled Final Bonus Score	0.58	0.36	0.06	0.97	-0.31	$p < 0.01$	149
Recalled Rank	0.59	0.16	0.25	1.10	-0.48	$p < 0.01$	154

Notes: The Base Bonus depends on the performance of the manager's store on four dimensions – customer service, manager review, profit, and sales growth – relative to other stores. The Area Bonus is constructed in the same way but using the average of the manager's store performance with the performances of other stores in the near geographic area, and ranking area performance relative to other areas. Final Bonus is the product of the Base Bonus and the Area Bonus, excluding top-performing managers who received extra bonuses. Managers were asked to recall their score for each of the four individual dimensions, with the product yielding an implied recalled Base Bonus. Managers were also asked to recall Area Bonus, Final Bonus, and rank. Recall is treated as accurate if it is within +/- 0.1 of the true performance. Average recall errors, conditional on being flattering or unflattering, are given in terms of standard deviations to allow comparing magnitudes across measures. T-tests are two-sided tests against the null of zero average recall error. In the top panel the sample is comprised of managers who gave responses within the range of feasible values for all of the measures, so recall errors are being compared across different aspects of performance for the same group of managers. The bottom panel does not hold constant the sample across measures.

A second observation about Table K1 is that managers are particularly likely to be accurate in remembering the AB, more so than about the other performance metrics, and this increase in accuracy comes mainly from a reduction in frequency of flattering errors. This pattern is statistically significant, as shown in Table K2. The table reports multinomial logit regressions, where the dependent variable takes on values 1, 2, or 3 corresponding to flattering, accurate, or inaccurate memory. The key independent variables are dummy variables for recollection about the BB and the FB, respectively, with AB as the omitted category. The results show a significantly higher probability of having flattering recall errors, and lower probabilities of accurate memories, when recalling the BB or FB, compared to recalling the AB. A possible interpretation of these findings is that fewer managers are motivated to distort memories of the AB, because it is less diagnostic of own performance. One caveat could be that there could be something that makes it easier for managers to recall the AB than other metrics, e.g., the fact that it is less variable over time might make it more likely that a manager who does not recall, and just guesses the mean, will be correct (this could explain greater accuracy, but not why recall errors are more symmetric). A robustness check, however, casts doubt on this particular alternative explanation: We estimate a probit regression for each performance metric, where the dependent variable is an indicator for having accurate recall for that metric, and the key independent variable is the variance a manager has experienced in that metric over their career; the coefficients on the corresponding variances are far from significant in each of the regressions. [13]

[13]Another approach is to add controls for the variances managers have experienced, for the different performance metrics, to the regressions shown in Table K2. These controls are not significantly related to the probability of being accurate in memory, again suggesting that lower variance does not drive accuracy of recall through making it easier to guess. The coefficients of interest, on the dummies for recalling BB and FB also remain similar with the addition of these controls (although the coefficient for BB for flattering errors is not quite significant, $p < 0.103$). Another caveat is that we incentivized managers to within 0.1 of the true score for the FB, but required being within 0.02 of the true score for AB, since the latter is relatively less variable. This could potentially explain why manager recollections are closer to the truth for AB than FB, but would not explain why errors are more symmetric than for the FB. Furthermore, for the individual dimensions, we did not give managers a margin for error, but rather asked them to report their scores exactly, because of the greater discreteness of the measures, but we see more asymmetry in recall errors for the implied BB than for the AB. Thus, asymmetric errors do not appear to be a function of allowing a greater margin for error for the incentives. Note that in our analysis accuracy is always assessed using +/- 0.1 band so that we are using the same yardstick to evaluate all measures.

Table K2: Comparing probabilities of different types of recall errors about different types of performance metrics

	Flattering (1)	Accurate (2)	Unflattering (3)	Flattering (4)	Accurate (5)	Unflattering (6)
Recalling BB	0.13*	-0.16**	0.03	0.16**	-0.19**	0.02
	(0.08)	(0.08)	(0.04)	(0.08)	(0.09)	(0.04)
Recalling FB	0.31***	-0.22***	-0.09	0.35***	-0.16*	-0.18*
	(0.07)	(0.08)	(0.06)	(0.07)	(0.10)	(0.10)
Performance percentile in Q3 of 2015				-0.02	0.03	-0.01
				(0.05)	(0.05)	(0.03)
Performance percentile in Q2 of 2015				-0.14***	0.13***	0.01
				(0.05)	(0.04)	(0.03)
Mean performance percentile pre- Q2 of 2015				0.09**	-0.08***	-0.01
				(0.04)	(0.03)	(0.02)
Female				0.04	-0.09	0.05
				(0.08)	(0.07)	(0.05)
Age				0.01	-0.03	0.02
				(0.05)	(0.04)	(0.03)
Experience				0.03	0.02	-0.05*
				(0.05)	(0.05)	(0.03)
Observations		219			289	
Pseudo R^2		0.045			0.10	

Notes: The sample is comprised of the managers participating in our lab in the field study, with three observations per manager. The three observations record each manager's recall errors about the Base Bonus, Final Bonus, and Area Bonus, respectively. The table reports marginal effects from multinomial logit regressions. The key independent variables are dummy variables for the recall error being about the Base Bonus or the Final Bonus, with error about the Area Bonus being the omitted category. Columns (4) to (6) include additional controls for manager past performance and traits. Robust standard errors are in parentheses, clustering on manager.

The lab in the field study also included a few additional questions that asked managers to recall other types of outcomes in Q2 of 2015. For completeness, we briefly sketch here the main findings from these measures. One set of measures asked managers to recall the highest and lowest Final Bonus achieved in Q2 of 2015. A substantial number of managers are approximately correct, about half, but the remainder are incorrect to varying degrees. Thus, managers do pay attention to outcomes besides their own performance, but memory is far from perfect. We also asked managers to recall the score above and below the one that they achieved themselves, on each of the four dimensions. This was intended as a check on internal consistency of memories. Manager answers to these questions are strongly positively correlated with memories of own performance, typically close to 0.80, showing substantial although imperfect internal consistency in responses.

L Robustness checks on the link between overconfidence and biased memories

This section explores robustness of the reduced form result that positive memories of Q2 of 2015 are associated with making overconfident predictions about Q4 of 2015 performance.

One set of concerns has to do with the definition of the dependent variable. We explore whether the result holds using a range of alternative benchmarks for defining the indicator variable for overconfidence that is the dependent variable. Another question is whether overly-negative memories are associated with a binary indicator of underconfidence, which would be another indication that manager predictions about the future are linked to memories of past signals. A different potential issue is whether the results are robust to non-binary dependent variables that measure the difference between manager predictions and reduced form predictors. Tables L1, L2, and L3 show the results measuring overconfidence and underconfidence, binary and non-binary, relative to different rule of thumb predictors, and Tables L4, L5, and L6 show analogous results for different multinomial logit models. Focusing on the 42 regression specifications that include the full set of controls, 41 have a coefficient for the measure of manager memory that is of the expected sign, and 30 are statistically significant. See also Section 3.4 in the text, and table notes for more details on the estimations.

A different type of concern is whether the relationship of overconfidence to memories, shown in Table 3 in the text, might reflect omitted variable bias. We explore a range of different possibilities. (1) *Predictions and memories formed from the same summary statistic:* Predictions and memories might be correlated if they are both based on some summary statistic of past performance besides the mean (the main analysis already controls for the mean). We therefore explore adding controls for various moments of the distribution of past performance: median, variance, maximum, minimum, and mode. (2) *Elapsed time:* Another concern could be that the coefficient on recalled Q2 performance is picking up a time effect, if memory is correlated with the elapsed time between the arrival of Q2 information and the memory elicitation. Thus, we add a control for this elapsed time. (3) *Valuing financial incentives for accuracy:* In case heterogeneity in valuing the magnitude of incentives we offer is relevant for precision of predictions, we control for performance on an incentivized addition task. Each unit of the task entails adding a set of 5 two-digit numbers, with similar magnitude of incentives for correct answers as our recall measure, \$2 per correct answer. To the extent that managerial ability at addition is relatively homogeneous, the measure can be a proxy for being motivated by incentives. (4) *Inattentiveness and low cognitive ability:*

In case inaccurate memory and inaccurate predictions might be correlated to due an omitted variable of inattentiveness or low cognitive ability we control for proxies based on a measure of attentiveness to details of the firm's incentive scheme, and understanding of the multiplicative nature of the scheme.[14] (5) *Tendency to exaggerate the truth:* To check whether stating overconfident predictions might be related to overly positive recall due to a willingness to exaggerate the truth, we control for a measure of this tendency: The number a manager reported rolling on an incentivized, private die roll. (6) *Additional traits:* Another possibility is that some additional manager traits are relevant for both memories and predictions, so we include controls for manager traits: Willingness to take risks in an incentivized task, and self-assessments of risk attitudes, patience, competitiveness, and relative confidence. Appendix E provides more details on these measures.

Table L7 at the end of this section shows that results are robust to adding these additional controls; there remains a statistically significant relationship between overconfidence about the future, and overly positive memories of the past. Most of the additional controls are not consistently statistically significant across specifications. One exception is self-assessed relative confidence, which is significantly related to making more confident predictions; this is consistent with overconfident managers noticing that they are confident about relative performance (but not necessarily realizing that they are overconfident).

[14]Understanding means knowing that performance will be ranked higher if performance is equal across all four dimensions, compared to having unequal performances with the same mean across dimensions (mean-preserving spread).

Table L1: Alternative rule of thumb indicators for overconfidence as a function of flattering memories

	Manager prediction overconfident relative to rule of thumb prediction													
	Historical mode		Experienced		Drop early		Current store		Recent		No Q3		National	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Flattering memory about Q2 of 2015	0.18**	0.15*	0.17*	0.15*	0.23**	0.20**	0.30***	0.35***	0.22***	0.25***	0.18**	0.19***	0.15*	0.18*
	(0.08)	(0.08)	(0.10)	(0.08)	(0.10)	(0.10)	(0.08)	(0.08)	(0.08)	(0.09)	(0.08)	(0.07)	(0.08)	(0.09)
Performance percentile in Q2 of 2015	-0.07	0.01	-0.01	0.06	-0.02	0.03	-0.09**	0.01	-0.14***	-0.05	-0.08*	0.04	-0.14***	-0.10*
	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.07)	(0.05)	(0.07)	(0.04)	(0.07)	(0.04)	(0.05)	(0.04)	(0.05)
Performance percentile in Q3 of 2015		0.06		0.03		-0.03		-0.02		-0.07		0.09**		0.12**
		(0.04)		(0.05)		(0.07)		(0.05)		(0.07)		(0.04)		(0.05)
Mean performance percentile pre- Q2 of 2015		-0.19***		-0.25***		-0.13*		-0.09**		-0.05		-0.22***		-0.08*
		(0.03)		(0.04)		(0.07)		(0.04)		(0.05)		(0.03)		(0.04)
Female		-0.04		-0.07		0.04		-0.03		-0.02		0.02		-0.13
		(0.08)		(0.09)		(0.11)		(0.09)		(0.10)		(0.08)		(0.09)
Age		-0.07		-0.05		-0.09		-0.08		-0.05		-0.04		-0.02
		(0.05)		(0.06)		(0.07)		(0.05)		(0.05)		(0.05)		(0.06)
Experience		-0.01		-0.05		-0.01		0.05		0.03		-0.01		0.01
		(0.05)		(0.06)		(0.07)		(0.06)		(0.06)		(0.05)		(0.06)
Observations	128	120	90	89	75	75	108	102	100	93	130	110	125	105
Pseudo R^2	0.044	0.187	0.025	0.211	0.045	0.112	0.120	0.160	0.117	0.125	0.052	0.301	0.085	0.116

Notes: The table reports marginal effects from Probit regressions. The dependent variables are equal to 1 if the manager's prediction is overconfident relative to a given rule of thumb predictor and zero otherwise. Independent variables are standardized so the coefficients show the change in the probability of being overconfident associated with a 1 s.d. increase in the independent variable. In columns (1) and (2) the rule of thumb predictor is the historical mode. In columns (3) and (4) the predictor is the mode but the sample is restricted to managers with more than two years of experience. In columns (5) and (6) the mode is calculated for experienced managers with at least 16 quarters of experience, dropping their first 8 tournament outcomes. In columns (7) and (8) the mode uses only outcomes from the current store as of Q4 of 2015. In columns (9) and (10) the mode is calculated using only outcomes fo Q3, Q2, and Q1 of 2015. In columns (11) and (12) the mode excludes outcomes from Q3 of 2015. In columns (13) and (14) the mode is calculated using only outcomes from quarters with national tournaments. Robust standard errors are in parentheses.

Table L2: Alternative rule of thumb indicators for underconfidence as a function of unflattering memories

	Manager prediction underconfident relative to rule of thumb prediction													
	Historical mode		Experienced		Drop early		Current store		Recent		No Q3		National	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Unflattering memory about Q2 of 2015	0.10	0.12	0.12	0.15*	0.12	0.15	0.11	0.18*	0.23***	0.26***	0.20**	0.27***	0.21***	0.17**
	(0.08)	(0.08)	(0.10)	(0.08)	(0.11)	(0.11)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.08)	(0.07)	(0.07)
Performance percentile in Q2 of 2015	0.02	-0.03	-0.01	-0.09**	0.03	0.01	0.09**	0.04	0.12***	0.10	0.04	-0.03	0.07*	0.02
	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.04)	(0.06)	(0.04)	(0.06)	(0.04)	(0.05)	(0.04)	(0.04)
Performance percentile in Q3 of 2015		-0.07*		-0.08*		-0.04		-0.03		0.05		-0.10**		-0.08**
		(0.04)		(0.04)		(0.06)		(0.05)		(0.06)		(0.04)		(0.04)
Mean performance percentile pre- Q2 of 2015		0.18***		0.25***		0.09		0.12***		0.02		0.20***		0.15***
		(0.03)		(0.04)		(0.07)		(0.04)		(0.04)		(0.03)		(0.04)
Female		0.07		0.06		0.02		0.12		0.08		-0.02		-0.04
		(0.07)		(0.07)		(0.10)		(0.08)		(0.09)		(0.08)		(0.07)
Age		0.03		0.03		0.06		0.03		0.07		0.01		0.08**
		(0.05)		(0.05)		(0.05)		(0.05)		(0.05)		(0.04)		(0.04)
Experience		0.01		0.00		-0.09		-0.02		-0.03		-0.03		-0.01
		(0.05)		(0.06)		(0.06)		(0.06)		(0.05)		(0.05)		(0.04)
Observations	128	120	90	89	75	75	108	102	100	93	130	110	125	105
Pseudo R^2	0.010	0.166	0.014	0.262	0.019	0.073	0.041	0.126	0.090	0.112	0.032	0.238	0.065	0.251

Notes: The table reports marginal effects from Probit regressions. The dependent variables are equal to 1 if the manager's prediction is underconfident relative to a given rule of thumb predictor and zero otherwise. Independent variables are standardized so the coefficients show the change in the probability of being overconfident associated with a 1 s.d. increase in the independent variable. In columns (1) and (2) the rule of thumb predictor is the historical mode. In columns (3) and (4) the predictor is the mode but the sample is restricted to managers with more than two years of experience. In columns (5) and (6) the mode is calculated for experienced managers with at least 16 quarters of experience, dropping their first 8 tournament outcomes. In columns (7) and (8) the mode uses only outcomes from the current store as of Q4 of 2015. In columns (9) and (10) the mode is calculated using only outcomes fo Q3, Q2, and Q1 of 2015. In columns (11) and (12) the mode excludes outcomes from Q3 of 2015. In columns (13) and (14) the mode is calculated using only outcomes from quarters with national tournaments. Robust standard errors are in parentheses.

Table L3: Size of manager deviation from rule of thumb predictor as a function of memory deviation

	Manager prediction for Q4 performance quintile - rule of thumb prediction													
	Historical mode		Experienced		Drop early		Current store		Recent		No Q3		National	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Recalled minus actual Q2 performance	0.27**	0.27**	0.15	0.28*	0.29**	0.29*	0.29*	0.33*	0.38***	0.32**	0.26	0.34**	0.32*	0.27
	(0.14)	(0.12)	(0.16)	(0.14)	(0.13)	(0.17)	(0.17)	(0.17)	(0.13)	(0.15)	(0.16)	(0.15)	(0.17)	(0.18)
Performance percentile in Q2 of 2015		0.28		0.50***		-0.00		0.06		-0.08		0.45**		-0.13
		(0.18)		(0.19)		(0.32)		(0.21)		(0.17)		(0.21)		(0.22)
Performance percentile in Q3 of 2015		0.30**		0.23		0.20		0.15		-0.17		0.39**		0.50**
		(0.15)		(0.17)		(0.27)		(0.20)		(0.18)		(0.17)		(0.22)
Mean performance percentile pre- Q2 of 2015		-0.85***		-1.15***		-0.45*		-0.51***		-0.12		-1.05***		-0.55***
		(0.13)		(0.18)		(0.27)		(0.16)		(0.10)		(0.14)		(0.17)
Female		-0.21		-0.31		-0.11		-0.17		-0.04		0.02		-0.37
		(0.23)		(0.26)		(0.37)		(0.28)		(0.23)		(0.26)		(0.31)
Age		-0.21*		-0.07		-0.13		-0.19		-0.07		-0.16		-0.19
		(0.12)		(0.15)		(0.18)		(0.15)		(0.13)		(0.14)		(0.18)
Experience		0.05		-0.09		0.10		0.09		0.09		0.03		0.01
		(0.14)		(0.19)		(0.20)		(0.16)		(0.15)		(0.16)		(0.19)
Constant	-0.21	-0.00	-0.20	0.14	-0.06	-0.00	-0.15	-0.02	-0.27**	-0.19	-0.19	-0.13	-0.02	0.17
	(0.13)	(0.17)	(0.16)	(0.21)	(0.17)	(0.27)	(0.13)	(0.20)	(0.11)	(0.18)	(0.14)	(0.20)	(0.14)	(0.23)
Observations	131	120	92	89	76	75	112	102	103	93	132	110	127	105
Pseudo R^2	0.009	0.120	0.003	0.134	0.009	0.030	0.012	0.057	0.030	0.054	0.006	0.157	0.010	0.064

Notes: The table reports marginal effects from interval regressions. The dependent variables are manager prediction about the most likely quintile in Q4 of 2015 minus the prediction of the corresponding rule of thumb predictor. Independent variables are standardized so the coefficients show the change in the probability of being overconfident associated with a 1 s.d. increase in the independent variable. In columns (1) and (2) the rule of thumb predictor is the historical mode. In columns (3) and (4) the predictor is the mode but the sample is restricted to managers with more than two years of experience. In columns (5) and (6) the mode is calculated for experienced managers with at least 16 quarters of experience, dropping their first 8 tournament outcomes. In columns (7) and (8) the mode uses only outcomes from the current store as of Q4 of 2015. In columns (9) and (10) the mode is calculated using only outcomes fo Q3, Q2, and Q1 of 2015. In columns (11) and (12) the mode excludes outcomes from Q3 of 2015. In columns (13) and (14) the mode is calculated using only outcomes from quarters with national tournaments. Robust standard errors are in parentheses.

Table L4: Alternative multinomial logit indicators for overconfidence as a function of flattering memories

	Manager prediction overconfident relative to multinomial logit predictor													
	Historical mode		Experienced		Drop early		Current store		Recent		No Q3		National	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Flattering memory about Q2 of 2015	0.20**	0.20**	0.12	0.11	0.30***	0.28***	0.12*	0.11	0.18**	0.17**	0.12	0.11	0.06	0.05
	(0.10)	(0.10)	(0.07)	(0.07)	(0.09)	(0.09)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)
Performance percentile in Q2 of 2015	-0.14***	-0.14**	-0.23***	-0.25***	-0.13***	-0.14**	-0.23***	-0.27***	-0.22***	-0.22***	-0.23***	-0.25***	-0.22***	-0.23***
	(0.05)	(0.06)	(0.03)	(0.03)	(0.05)	(0.06)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Performance percentile in Q3 of 2015		0.00		0.09**		0.01		0.09**		0.06		0.09**		0.07*
		(0.06)		(0.04)		(0.06)		(0.04)		(0.04)		(0.04)		(0.04)
Mean performance percentile pre- Q2 of 2015		-0.11*		-0.04		-0.05		-0.02		-0.05		-0.04		-0.08**
		(0.06)		(0.03)		(0.06)		(0.03)		(0.04)		(0.03)		(0.04)
Female		-0.14		-0.05		-0.03		-0.02		-0.03		-0.05		-0.04
		(0.11)		(0.08)		(0.11)		(0.07)		(0.08)		(0.08)		(0.08)
Age		-0.00		-0.01		0.02		0.03		-0.02		-0.01		0.01
		(0.07)		(0.05)		(0.06)		(0.04)		(0.05)		(0.05)		(0.05)
Experience		-0.06		0.02		-0.08		-0.01		0.02		0.02		-0.01
		(0.08)		(0.05)		(0.07)		(0.05)		(0.05)		(0.05)		(0.05)
Observations	75	75	129	127	75	75	129	127	129	127	129	127	129	127
Pseudo R^2	0.115	0.152	0.226	0.269	0.162	0.178	0.249	0.284	0.234	0.247	0.226	0.269	0.195	0.236

Notes: The table reports marginal effects from Probit regressions. The dependent variables are equal to 1 if the manager's prediction is overconfident relative to a given multinomial logit predictor and zero otherwise. Independent variables are standardized so the coefficients show the change in the probability of being overconfident associated with a 1 s.d. increase in the independent variable. In columns (1) and (2) the predictor is the 8 lag model. In columns (3) and (4) the predictor is the 3 lag model. In columns (5) and (6) the predictor is the 8 lag model, estimated without the first 8 tournament outcomes of the sample of experienced managers. In columns (7) and (8) the predictor is estimated using tournament outcomes from the current store as of Q4 of 2015. In columns (9) and (10) the predictor is estimated using only outcomes fo Q3, Q2, and Q1 of 2015. In columns (11) and (12) the predictor is estimated excluding outcomes from Q3 of 2015. In columns (13) and (14) the predictor is estimated using only outcomes from quarters with national tournaments. Robust standard errors are in parentheses.

Table L5: Alternative multinomial logit indicators for underconfidence as a function of unflattering memories

	Manager prediction underconfident relative to multinomial logit predictor													
	Historical mode		Experienced		Drop early		Current store		Recent		No Q3		National	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Unflattering memory about Q2 of 2015	-0.19	-0.18*	0.16**	0.19**	-0.16	-0.17	0.20**	0.19**	0.17**	0.17**	0.16**	0.19**	0.20***	0.20***
	(0.13)	(0.11)	(0.08)	(0.07)	(0.13)	(0.11)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)
Performance percentile in Q2 of 2015	0.09**	0.13***	0.18***	0.17***	0.12***	0.16***	0.16***	0.18***	0.16***	0.17***	0.18***	0.17***	0.19***	0.19***
	(0.04)	(0.05)	(0.03)	(0.04)	(0.04)	(0.05)	(0.03)	(0.04)	(0.03)	(0.04)	(0.03)	(0.04)	(0.03)	(0.04)
Performance percentile in Q3 of 2015		-0.09*		-0.04		-0.08		-0.09**		-0.10**		-0.04		-0.05
		(0.05)		(0.04)		(0.05)		(0.04)		(0.04)		(0.04)		(0.04)
Mean performance percentile pre- Q2 of 2015		0.06		0.08**		0.00		0.05		0.08**		0.08**		0.05
		(0.05)		(0.03)		(0.05)		(0.04)		(0.04)		(0.03)		(0.03)
Female		0.07		-0.02		-0.01		-0.01		0.02		-0.02		-0.01
		(0.09)		(0.07)		(0.08)		(0.07)		(0.07)		(0.07)		(0.07)
Age		0.05		0.05		0.04		0.04		0.06		0.05		0.05
		(0.05)		(0.04)		(0.05)		(0.04)		(0.04)		(0.04)		(0.04)
Experience		-0.01		-0.05		0.01		-0.06		-0.05		-0.05		-0.04
		(0.06)		(0.04)		(0.05)		(0.05)		(0.05)		(0.04)		(0.04)
Observations	75	75	129	127	75	75	129	127	129	127	129	127	124	122
Pseudo R^2	0.097	0.180	0.184	0.242	0.141	0.200	0.140	0.200	0.130	0.212	0.184	0.242	0.220	0.255

Notes: The table reports marginal effects from Probit regressions. The dependent variables are equal to 1 if the manager's prediction is underconfident relative to a given multinomial logit predictor and zero otherwise. Independent variables are standardized so the coefficients show the change in the probability of being overconfident associated with a 1 s.d. increase in the independent variable. In columns (1) and (2) the predictor is the 8 lag model. In columns (3) and (4) the predictor is the 3 lag model. In columns (5) and (6) the predictor is the 8 lag model, estimated without the first 8 tournament outcomes of the sample of experienced managers. In columns (7) and (8) the predictor is estimated using tournament outcomes from the current store as of Q4 of 2015. In columns (9) and (10) the predictor is estimated using only outcomes fo Q3, Q2, and Q1 of 2015. In columns (11) and (12) the predictor is estimated excluding outcomes from Q3 of 2015. In columns (13) and (14) the predictor is estimated using only outcomes from quarters with national tournaments. Robust standard errors are in parentheses.

Table L6: Size of manager deviation from multinomial predictor as a function of memory deviation

	Manager prediction for Q4 performance quintile - multinomial logit prediction													
	Historical mode		Experienced		Drop early		Current store		Recent		No Q3		National	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Recalled minus actual Q2 performance	0.45**	0.12	0.62***	0.26*	0.54***	0.22	0.63***	0.29**	0.59***	0.26**	0.63***	0.28**	0.50***	0.17
	(0.20)	(0.16)	(0.14)	(0.14)	(0.18)	(0.16)	(0.16)	(0.14)	(0.14)	(0.13)	(0.14)	(0.14)	(0.15)	(0.15)
Performance percentile in Q2 of 2015		-0.85***		-1.00***		-0.85***		-0.96***		-0.86***		-0.97***		-0.89***
		(0.25)		(0.15)		(0.24)		(0.14)		(0.14)		(0.16)		(0.16)
Performance percentile in Q3 of 2015		0.47*		0.55***		0.47**		0.59***		0.55***		0.55***		0.55***
		(0.24)		(0.14)		(0.23)		(0.14)		(0.14)		(0.14)		(0.15)
Mean performance percentile pre- Q2 of 2015		-0.34*		-0.21**		-0.19		-0.17		-0.27***		-0.24**		-0.26**
		(0.18)		(0.10)		(0.17)		(0.11)		(0.10)		(0.11)		(0.11)
Female		-0.42		-0.15		-0.28		-0.02		-0.04		-0.14		-0.06
		(0.35)		(0.23)		(0.35)		(0.23)		(0.22)		(0.23)		(0.24)
Age		0.01		-0.08		0.07		-0.05		-0.13		-0.07		-0.06
		(0.18)		(0.15)		(0.16)		(0.14)		(0.13)		(0.14)		(0.14)
Experience		-0.44*		0.01		-0.40*		-0.02		-0.00		0.00		-0.04
		(0.25)		(0.16)		(0.23)		(0.16)		(0.15)		(0.16)		(0.16)
Constant	-0.05	0.32	-0.19	-0.07	-0.16	0.12	-0.34***	-0.30*	-0.24*	-0.19	-0.19	-0.06	-0.25*	-0.18
	(0.18)	(0.27)	(0.13)	(0.17)	(0.17)	(0.26)	(0.13)	(0.17)	(0.12)	(0.16)	(0.13)	(0.17)	(0.14)	(0.19)
Observations	75	75	129	127	75	75	129	127	129	127	129	127	129	127
Pseudo R^2	0.021	0.093	0.036	0.149	0.033	0.099	0.040	0.147	0.039	0.154	0.038	0.152	0.025	0.122

Notes: The table reports marginal effects from interval regressions. The dependent variables are manager prediction about the most likely quintile in Q4 of 2015 minus the prediction of the corresponding multinomial logit predictor. Independent variables are standardized so the coefficients show the change in the dependent variable associated with a 1 s.d. increase in the independent variable. In columns (1) and (2) the predictor is the 8 lag model. In columns (3) and (4) the predictor is the 3 lag model. In columns (5) and (6) the predictor is the 8 lag model, estimated without the first 8 tournament outcomes of the sample of experienced managers. In columns (7) and (8) the predictor is estimated using tournament outcomes from the current store as of Q4 of 2015. In columns (9) and (10) the predictor is estimated using only outcomes fo Q3, Q2, and Q1 of 2015. In columns (11) and (12) the predictor is estimated excluding outcomes from Q3 of 2015. In columns (13) and (14) the predictor is estimated using only outcomes from quarters with national tournaments. Independent variables are standardized so the coefficients show the impact of a 1 s.d. increase in the independent variable on the probability of being overconfident. Robust standard errors are in parentheses.

Table L7: Manager predictions and overconfidence as a function of recalled Q2 performance and additional controls

	Manager prediction		Overconfident (rel. to mult. logit)		Underconfident (rel. to historical mode)	
	(1)	(2)	(3)	(4)	(5)	(6)
Recalled performance quintile for Q2 of 2015	0.41**	0.52***				
	(0.16)	(0.16)				
Flattering memory about Q2 of 2015			0.26***	0.52***	0.16**	0.20***
			(0.10)	(0.12)	(0.07)	(0.06)
Performance percentile in Q2 of 2015	0.13	0.26	-0.06	-0.07	0.07	0.17***
	(0.21)	(0.21)	(0.08)	(0.09)	(0.06)	(0.05)
Performance percentile in Q3 of 2015	0.58***	0.51***	-0.06	-0.07	0.13**	0.23***
	(0.19)	(0.19)	(0.07)	(0.08)	(0.05)	(0.07)
Mean performance percentile pre- Q2 of 2015	-0.23	0.00	-0.41	-0.38*	0.01	0.04
	(0.33)	(0.32)	(0.29)	(0.23)	(0.11)	(0.08)
Female	-0.21	-0.37	-0.18*	-0.50***	-0.04	-0.19**
	(0.25)	(0.26)	(0.11)	(0.18)	(0.07)	(0.08)
Age	-0.07	-0.10	-0.03	-0.34***	-0.02	-0.01
	(0.13)	(0.13)	(0.07)	(0.09)	(0.04)	(0.04)
Experience	-0.19	-0.11	-0.08	0.26**	-0.10*	-0.04
	(0.24)	(0.27)	(0.08)	(0.13)	(0.05)	(0.05)
Maximum historical performance percentile	0.45	0.41	0.11	-0.23	0.17**	0.14*
	(0.34)	(0.34)	(0.18)	(0.21)	(0.09)	(0.08)
Minimum historical performance percentile	-0.22	-0.21	0.01	-0.54**	-0.12*	-0.26***
	(0.26)	(0.27)	(0.22)	(0.22)	(0.07)	(0.09)
Modal historical performance quintile	0.33	0.45	0.10	0.30**	-0.37***	-0.35***
	(0.37)	(0.41)	(0.10)	(0.12)	(0.09)	(0.08)
Median historical performance percentile	-0.04	-0.37	0.12	0.59***	-0.03	-0.17*
	(0.49)	(0.51)	(0.24)	(0.20)	(0.14)	(0.10)
Variance of historical performance percentiles	-0.41	-0.29	-0.06	-0.02	-0.19***	-0.31***
	(0.26)	(0.28)	(0.13)	(0.11)	(0.06)	(0.08)
Days elapsed between Q2 and memory measurement		0.21		0.16***		0.08**
		(0.13)		(0.05)		(0.04)
Addition problems solved in incentivized task		0.05		0.44***		0.08**
		(0.16)		(0.13)		(0.04)
Knows scale values for incentive scheme		-0.04		-0.29**		-0.00
		(0.27)		(0.14)		(0.05)
Understands implications of mult. incentive scheme		-0.17		-0.34**		-0.30***
		(0.34)		(0.16)		(0.10)
Risk all in incentivized measure		-0.13		-0.04		-0.07
		(0.12)		(0.04)		(0.04)
Die roll in incentivized lying task		0.14		0.02		0.20***
		(0.14)		(0.04)		(0.06)
Self-assessed willingness to take risks		0.25		0.19*		0.15***
		(0.16)		(0.10)		(0.05)
Self-assessed competitiveness		-0.19		-0.33***		-0.11***
		(0.18)		(0.10)		(0.04)
Self-assessed relative confidence		0.27		0.20***		0.11***
		(0.18)		(0.06)		(0.04)
Self-assessed patience		-0.01		-0.13**		-0.08**
		(0.13)		(0.06)		(0.03)
Constant	3.03***	3.18***				
	(0.20)	(0.23)				
Observations	120	97	62	55	120	97
Pseudo R^2	0.162	0.235	0.194	0.482	0.369	0.635

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 of probit regressions. The dependent variable for Columns (3) and (4) is an indicator for whether a manager predicted a higher quintile than the quintile predicted by the baseline (8 lag) multinomial logit model. The dependent variable for Columns (5) and (6) is an indicator for whether a manager predicted a higher quintile than their historical modal quintile. Independent variables are standardized, so coefficients give the change in the dependent variable 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 estimation sample only includes managers with a unique historical mode. Knowledge of scale values is an indicator for whether the manager knows both the min. and max. possible values for the scale to evaluate performance on individual dimensions. Understanding of the multiplicative nature of the scheme is an indicator for whether the manager understands that performance will be ranked higher if performance is equal across all dimensions, compared to having unequal performances with the same mean across dimensions (mean-preserving spread). Risk taking in the incentivized task is how much money the manager invested in a risky rather than a safe asset. Reporting a higher die roll is a (noisy) indicator of lying. Self-assessments are on an 11-point scale, with higher values indicating greater willingness to take risks, etc.. Robust standard errors are in parentheses.

Table L8: Manager predictions and overconfidence as a function of recalled Q2 performance with experience interaction terms

	Manager prediction		Overconfident (rel. to mult. logit)		Underconfident (rel. to historical mode)	
	(1)	(2)	(3)	(4)	(5)	(6)
Recalled performance quintile for Q2 of 2015	0.55***	0.50***				
	(0.17)	(0.16)				
Performance percentile in Q2 of 2015	0.41**	0.22	-0.14***	-0.13**	-0.07	0.02
	(0.18)	(0.17)	(0.05)	(0.06)	(0.04)	(0.05)
Performance percentile in Q3 of 2015		0.62***		0.00		0.07
		(0.16)		(0.06)		(0.04)
Mean performance percentile pre- Q2 of 2015		0.06		-0.11*		-0.19***
		(0.10)		(0.06)		(0.03)
Female		-0.10		-0.15		-0.04
		(0.23)		(0.11)		(0.08)
Age		-0.08		-0.02		-0.07
		(0.12)		(0.07)		(0.05)
Experience		0.13		-0.09		-0.03
		(0.14)		(0.08)		(0.06)
Recalled Q2 performance*Experience		0.31**				
		(0.12)				
Flattering memory about Q2 of 2015			0.20**	0.18*	0.18**	0.17**
			(0.10)	(0.10)	(0.08)	(0.08)
Flattering memory about Q2*Experience				0.14		0.06
				(0.14)		(0.08)
Constant	3.08***	3.23***				
	(0.11)	(0.19)				
Observations	170	148	75	75	128	120
Pseudo R^2	0.101	0.170	0.115	0.163	0.044	0.190

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 of probit regressions. The dependent variable for Columns (3) and (4) is an indicator for whether a manager predicted a higher quintile than the quintile predicted by the baseline (8 lag) multinomial logit model. The dependent variable for Columns (5) and (6) is an indicator for whether a manager predicted a higher quintile than their historical modal quintile. Independent variables are standardized, so coefficients give the change in the dependent variable 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 estimation sample only includes managers with a unique historical mode. Risk taking in the incentivized task is how much money the manager invested in a risky rather than a safe asset. Reporting a higher die roll is a (noisy) indicator of lying. Self-assessments are on an 11-point scale, with higher values indicating greater willingness to take risks, etc.. Robust standard errors are in parentheses.

M Details on the structural analysis

M.1 Details on estimation of the baseline Bayesian model

As discussed in the body of the paper the baseline structural model explores whether manager predictions might be rationalizable by a mechanism in which fully Bayesian managers learn about an underlying "type" based on individual histories of tournament outcomes. We first recapitulate our summary in the body of the paper. We suppose that there are a finite number of periods $t = 1, 2, ...T$ corresponding to quarters. We suppose that the manage has a type a_k that takes on a value between 1-5 and is time invariant. The assumption of five types is arbitrary, but has a natural interpretation in terms of reflecting individuals' quintiles in terms of ability.[15] One can imagine that 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 N). In extensions of the baseline model, discussed in robustness checks, type is partly endogenous because it also depends on the manager's effort, $e_{k,t}$, with $a_{k,t} = \Gamma(\theta_k, e_{k,t})$. For now we assume type does not depend on effort.

Every period a public signal $s_{k,t}$, which the rank quintile, is generated for each manager, taking on an integer value between 1 and 5.[16] We assume that $s_{k,t}$ 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.[17]

f is the belief distribution of the manager over their own 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

[15]Adding more types makes it less likely that the data could be rationalized in a Bayesian way. AsBenoît 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.

[16]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 inBenoît and Dubra (2011).

[17]We suppose that the distribution over signals may depend on time, but not the individual. In other words the functional form of $s_{k,t}$ may depend on t but not on k.

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. Betting behavior is denoted $b_{k,\tau}(j)$.

Estimation, described below, requires some identifying assumptions, which ensure that variation in signals over time is just due to (mean zero) noise, and it is possible to back out from a set of noisy signals the exogenous part of a manager's quality, θ_i.

Our interpretation of the baseline model is that manager type is time invariant, , i.e. $a_{k,t} = a_k$, 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. We assume that P is also time invariant. 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.[18] Our assumptions regarding time invariance ensure that variation in signals over time is just due to (mean zero) noise, and it is possible to back out from a set of noisy signals the exogenous part of a manager's quality, θ_i.

The estimation of the baseline Bayesian model proceeds in three steps.

STEP 1: The first step is to estimate the unobserved matrix P, based on observable data about a transition matrix, Z. The transition matrix gives the probabilities of observing each of the possible signals (quintiles) in quarter $t+1$, for each possible quintile outcome in quarter t. Formally $Z_{i,j,}$, the i,j^{th} entry of Z, is the probability that that signal j occurs in period $t+1$, conditional on signal i having occurred in period t. In our model, there is a mapping from P to the elements of Z. To see this, note that since signals and types are quintiles, we know that the belief about an individual k being type i, after observing a single signal, j, must be $\frac{\frac{1}{5}P_{i,j}}{\frac{1}{5}}$. The probability of observing signal $\hat{j}$ in the next period is then $Z_{j,\hat{j},} = \sum_i P_{i,j}P_{i,\hat{j}}$. Thus, in our model, Z is polynomial function (of degree 2) of the entries in P; moreover, it is symmetric. In our data, for each period t

[18]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.

we have an empirically observable transition matrix Z_t. In total our data yield 33 Z_ts, which are noisy observations about Z.

We can estimate the P that best fits these data using a minimum distance estimator which minimizes $\sum_t \sum_j \sum_{\hat{j}} [Z_{j,\hat{j},t} - \sum_i P_{i,j} P_{i,\hat{j}}]^2$, subject to several constraints.[19] Formally, the i,j^{th} entry of P, denoted $P_{i,j}$, is $p(j|i)$. Two constraints in the estimation reflect the fact that rows and columns of P must sum to 1, respectively. First, $\sum_j P_{i,j} = 1$ since the rows of P give conditional probabilities, conditional on the same event. Second, because the signals are quintiles $\sum_i P_{i,j} = 1$.[20] We denote the resulting estimate by $\hat{P}$.

Unfortunately, proving clean identification is difficult. This is for two reasons. The first is that even if we have a single set of Z_t's that we were trying to match, proving the uniqueness of a solution is difficult. Although there are techniques developed for analytically solving systems of polynomial equations, and showing uniqueness, such approaches are not computationally feasible given the number of variables we have (i.e., the 25 entries in the P matrix). A second issue is that, if we consider our estimation procedure, the objective function $\sum_t \sum_j \sum_{\hat{j}} [Z_{j,\hat{j},t} - \sum_i P_{i,j} P_{i,\hat{j}}]^2$ is not globally concave.

To address these issues and verify our solution we generate 1,000 initial P matrices with random entries (satisfying our constraints). For each of these sets of starting values, we then numerically solve the constrained minimization problem for a (potentially local) minimum. One could also imagine doing a grid search over all possible values to find the minimum, but given the number of parameters we need to estimate, even with a coarse grid such an approach is not computationally feasible.[21] We find, however, that the estimated P does not depend on the initial values.[22]

Table M1 provides a first result from the baseline structural model, which is the estimate of $\hat{P}$. The matrix is well-behaved in that it satisfies the Monotone Ratio Likelihood Property: Better types have higher probabilities of observing better signals. It also shows that the baseline model is predictive, in the sense that knowing a manager's type delivers a relatively large mass for the modal quintile. For example, the worst and best types have probabilities .60 and .62 of ending up with the worst and best quintiles,

[19]Note that the approach here, and elsewhere, is essentially an simulated methods of moments approach. The moments are the entries of Z, and the implicit weighting matrix is the identity matrix.

[20]The third constraint is a matter of convenience. Because types are unobservable, and so have no objective meaning, the rows of P are interchangeable. Thus there are multiple equivalent matrixes that contain the same probability distributions for types but only differ in the order of rows. We focus on the matrix that involves an easy to understand ordering of rows; we require that the estimation make type 1 (row 1) the type with the highest probability of signal 1, type 2 (row 2) the type with highest probability of signal 2, and so on. In the case that two rows generate the same signal with the highest chance, we assign the row that assigns that signal with a higher chance (this does not occur).

[21]A simulated annealing method would be an alternative approach to finding the global optimum.

[22]We only do this procedure once, for our baseline estimates, and do not repeat the random choice of initial starting values when we bootstrap, or for the robustness checks.

respectively.

Table M1: Estimated matrix $\hat{P}$ for the baseline model

	Signal					
	1	2	3	4	5	Total
Type 5	0.017	0.042	0.095	0.225	0.621	1
Type 4	0.030	0.146	0.166	0.401	0.257	1
Type 3	0.086	0.221	0.408	0.212	0.073	1
Type 2	0.261	0.394	0.192	0.123	0.031	1
Type 1	0.606	0.197	0.140	0.039	0.018	1
Total	1	1	1	1	1	

Notes: Estimated probability distributions across signals, by type. Signals correspond to quintiles in the performance distribution, with 5 being the best. Types are ordered from lower to higher ability. Rows sum to 1 because these are probability distributions. Columns sum to 1 because types are uniformly distributed.

STEP 2: The second step is to use $\hat{P}$, and each manager's history of tournament outcomes, to derive a posterior belief about a manager's most likely outcome for Q4 of 2015. Starting from a uniform prior about managers' types, we update these priors using P, a manager's history of tournament outcomes and Bayes' rule. Formally, suppose we are at the beginning of period τ and the individual has a history of signals $s_{k,\tau'}, s_{k,\tau'+1}, ..., s_{k,\tau-1}$ where $\tau^{'} < \tau$. By Bayes' rule the posterior belief in period τ that k is type $a_k = i$ is

$$f_{k,\tau}(i) = \frac{f_{k,0}(i)\Pi_{t=\tau'}^{\tau-1} P_{i,s_{k,t}}}{\sum_{\hat{i}} f_{k,0}(\hat{i})\Pi_{t=\tau'}^{\tau-1} P_{\hat{i},s_{k,t}}} = \frac{\Pi_{t=\tau'}^{\tau-1} P_{i,s_{k,t}}}{\sum_{\hat{i}} \Pi_{t=\tau'}^{\tau-1} P_{\hat{i},s_{k,t}}}$$

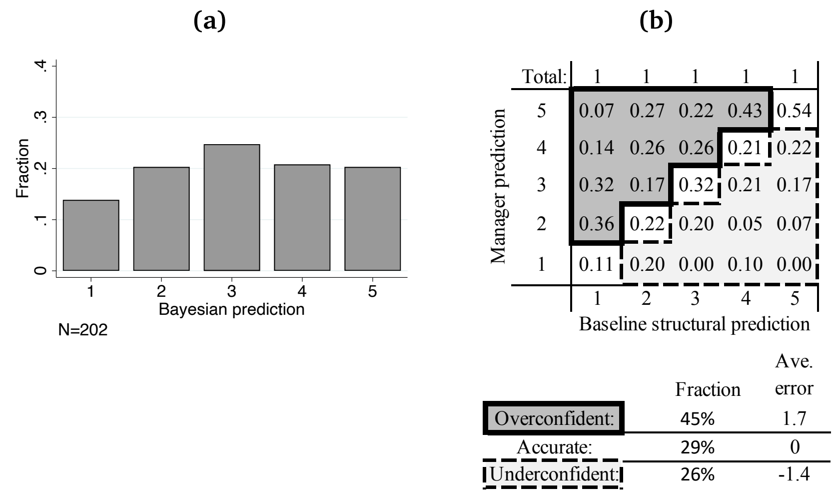
.

STEP 3: In the third step, we use the posterior distributions to identify each manager's modal quintile signal for Q4 of 2015. In particular, we know that manager beliefs about the likelihood of a given signal, j, is given by $g_\tau(j) = \sum_i f_{k,\tau}(i) P_{i,j}$. We denote this as the "Bayesian prediction" for a manager. Betting behavior is denoted $b_{k,\tau}(j)$. When there is a unique maximum in the $g_{k,\tau}$ vector, i.e., a unique modal quintile, then the vector describing betting behavior has $b_{k,\tau}(j) = 1$ if $g_\tau(j) = \max_{\hat{j}} g_\tau(\hat{j})$ and 0 otherwise. In the case when there isn't a unique optimum, $b_{k,\tau}(j) = 0$ if j is not a maximizer of $g_{k,\tau}$ and the $\sum_J b_{k,\tau}(j) = 1$, where J is the set of signals that are the maximizers of $g_{k,\tau}$ (in practice we never need to use this tie-breaking rule). We suppose that an individual, when asked to bet on what signal will occur in period τ, predicts the signal that is most likely.

M.2 Results on overconfidence relative to the baseline Bayesian model

Panel (a) of Figure M1 shows the distribution of Bayesian predictions. The distribution of predictions is hump-shaped, which 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.[23] Turning to a comparison with manager predictions, Panel (a) shows that predictions of the baseline structural model are less skewed towards predicting high quintiles than manager predictions (recall Figure 1). Furthermore, Panel (b) of Figure M1 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.

Figure M1: 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.

[23]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. Rule of thumb predictors, and multinomial logit predictors, did not have this extra step of inference about a manager's (assumed) underlying type.

One notable insight from the estimated $\hat{P}$ concerns the speed of learning: Tournament outcomes are quite informative about a manager's type, and thus learning should be fast. One implication is that even relatively extreme overconfident priors should be corrected within just two or three quarters.[24] This finding complements the evidence from the reduced form analysis, underlining that it is hard for overconfident priors alone to explain the persistence of overconfidence after 8 quarters of experience. In robustness checks on the structural model, discussed in the appendix, we also see a similar invariance of overconfidence with respect to the baseline structural model, as experience increases.

M.3 Bootstrapping the Bayesian structural model

In order to assess the statistical significance of the difference between manager and model predictions re-estimate P 100 times using a moving block bootstrapping design. Importantly, we want to take into account the noise in the signals used to estimate $\hat{P}$, posteriors about manager types, and the associated bets $b_{k,\tau}(\hat{P})$, in order to have a confidence interval around the Bayesian predictions.[25]

We sample the noise in our data using bootstrapping. We implement a moving block-bootstrap estimation using blocks (sequences) with lengths of 3 periods. We use the moving block approach as our observations are time series data which appear (as discussed previously) to be stationary. We conduct 100 bootstraps, each time generating a sample of 33 Z_t's (11 blocks) and estimating a $\tilde{P}$. We denote the n^{th} estimated $\tilde{P}$ from the bootstrap as $\widetilde{P}_n$. Given an estimated $\widetilde{P}$, we denote the betting vector induced by $\tilde{P}$ as $b_{k,\tau}(\widetilde{P})$. We then calculate the distance between each bootstrapped distribution of bets, $b_{k,\tau}(\widetilde{P}_n)$, and the central tendency of the model, i.e., the distribution of bets obtained using the original sample, $b_{k,\tau}(\hat{P})$. Given any two betting vectors b and $b^{'}$ we denote the Euclidean distance between them as $D(b, b^{'})$. We calculate the distances between bootstrapped bets and the central tendency as $\sum_k D(b_{k,\tau}(\hat{P}), b_{k,\tau}(\widetilde{P}_n))$. This yields a distribution of distances, denoted $\tilde{d}$, which provides a measure of the size of the errors in assigning managers particular bets.

The bootstrapping also allows a statistical test of the baseline Bayesian model. We can calculate one more distance, namely the distance of observed manager bets, denoted $b_{k,O}$, from the distribution of bets derived from $\hat{P}$. We can see where this dis-

[24]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.

[25]Given our interpretation of the model, this noise is also partly present in managers' Bayesian beliefs about their types, as they are learning over time.

tance, $\sum_k D(b_{k,\tau}(\hat{P}), b_{k,O})$, lies in the distribution of bootstrapped distances $\tilde{d}$. If it lies far in the tail of $\tilde{d}$ we will reject that the manager's predictions are consistent with the baseline model, even allowing for noise in the estimation process.

As shown in the next subsection, we can reject the hypothesis that the observed data is consistent with the baseline Bayesian model. The Euclidean distance of manager predictions from the predictions based on $\hat{P}$ lies far in the tail of the bootstrapped distances.[26] 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.[27]

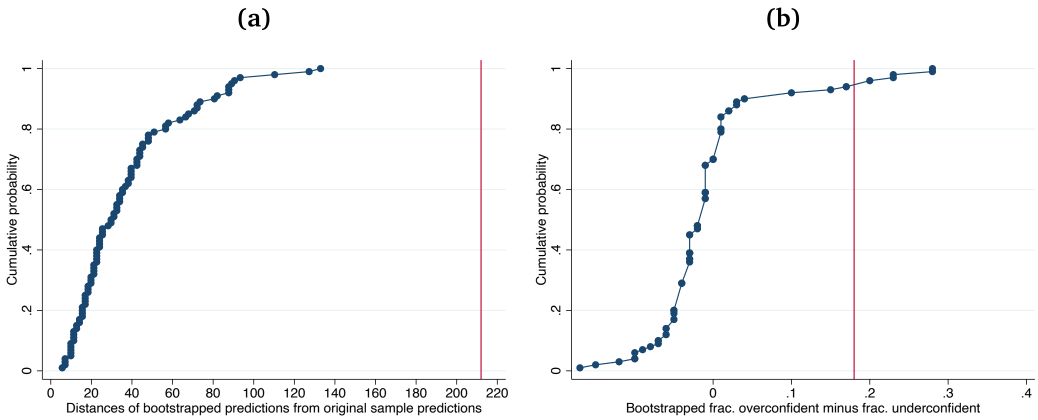
[26]If we run a χ-squared test of the difference between model and manager predictions we obtain $p < 0.05$.

[27]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.

M.4 Statistical test of the baseline structural model

This section provides the cumulative distribution functions from the bootstrapping of the baseline structural model. The results show that we can statistically reject at conventional levels that the baseline structural model matches individual manager predictions, or the skew of manager predictions towards overconfidence, as measured by the fraction of managers overconfident minus the fraction underconfident.

Figure M2: Statistical test of manager predictions vs. baseline structural model predictions



Notes: The connected (blue) dots in Panel (a) show the cumulative distribution of Euclidean distances between the bootstrapped structural predictions and predictions based on the original sample and $\hat{P}$. See Section 4.1 in the text for discussion of the bootstrapping. The vertical (red) line in Panel (a) shows the Euclidean distance of manager predictions from the predictions of the structural model using the original sample. The connected (blue) dots in Panel (b) show the cumulative distribution of the differences, for all of the bootstrapped predictions, of the fraction overconfident relative to the predictions based on the original sample and $\hat{P}$ minus the fraction underconfident. The vertical (red) line in Panel (b) shows the fraction of managers overconfident relative to the predictions of the structural model using the original sample minus the fraction of managers underconfident.

N Robustness checks for structural analysis

The structural model makes a number of identifying assumptions. This section considers whether these assumptions hold, and whether results are robust to relaxing these assumptions. Many of these assumptions, such as time stationarity and uniform priors 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. There are also some robustness checks on the structural model that do not have a close analogue in the reduced form analysis. For example, we evaluate whether allowing for choice errors in manager betting behavior (e.g. McFadden, 1974) can rationalize the data. The results of the robustness checks are summarized in Table N1.

Table N1: Summary of robustness checks on manager predictions vs. structural model predictors

	Manager vs. Bayesian structural predictors					
	Fraction of managers:			P-values		
	overconfident	accurate	underconfident	different	frac. overconf. - frac. underconf.	N
Overconfident priors:						
Baseline	0.44	0.31	0.25	p<0.01	p<0.05	202
Experienced only	0.47	0.29	0.24	p<0.01	p<0.07	109
Manager non-stationarity:						
Experienced, drop early	0.49	0.35	0.17	p<0.01	p<0.01	109
Current store only	0.48	0.30	0.22	p<0.01	p<0.01	202
Environment non-stationarity:						
Recent tournaments	0.44	0.32	0.24	p<0.01	p<0.01	202
Recent tournaments, recent P	0.42	0.31	0.27	N.A.	N.A.	202
Imperfect knowledge:						
Excluding Q3 tournament	0.44	0.30	0.26	p<0.01	p<0.06	201
Nationwide tournaments	0.47	0.31	0.22	p<0.01	p<0.01	194

Notes: The structural model uses all data back to Q1 of 2008 to estimate P, unless otherwise specified. Different robustness checks vary which tournament outcomes are combined with P to form predictions. The baseline model uses all manager signals prior to Q4 of 2015 to form predictions. P-values test whether manager predictions are different from the model predictions, and whether they are more skewed towards overconfidence. See text for details on bootstrapping. Predictions for experienced managers focus on the subset of managers with at least 8 tournament outcomes, using all of their outcomes to form predictions. Dropping early tournaments means predictions are formed without using an experienced manager's first 8 tournaments. The prediction based on the current store is based only on tournament outcomes from the store that the manager operated as of Q4 of 2015. Predictions based on recent tournaments use the outcomes from Q3, Q2, and Q1 of 2015 to form predictions, but P is estimated using all of the historical data. Recent P refers to estimating P using only the signal-to-signal matrixes (Z_t's) for Q3, Q2, and Q1 of 2015. In this case there are too few quarters to do meaningful bootstrapping of P. Predictions dropping Q3 use all tournament outcomes except Q3 of 2015. Predictions based on nationwide tournaments base predictions only on outcomes from quarters with nationwide tournaments.

One assumption of the structural model is uniform priors, but managers might enter the job with (rationally) overconfident priors. Even after incorporating a few signals from tournament outcomes, the posteriors could still be skewed towards predicting high performance quintiles. As managers gain experience, however, the impact of overconfident priors should wane if managers are Bayesian.[28] Focusing on predictions for the sub-sample of managers with more than two years of experience, the prevalence and extent of overconfidence in predictions is similar to the sample of managers as a whole, suggesting that the results are not driven by relatively inexperienced managers with overconfident priors (see Table N1).

Another identifying assumption of the baseline model is that the manager type, a_k, is time invariant. This might not hold if type is partly endogenous, e.g. affected by manager effort, and managers are not fully informed about θ_k. Denoting effort by $e_{k,t}$ we have with $a_{k,t} = \Gamma(\theta_k, e_{k,t})$. In this case, as they learn over time about θ_k, managers would adjust effort, leading to time varying $a_{k,t}$. Unobserved, changing effort levels would confound our efforts to infer a manager's fixed quality θ_k from tournament outcomes. Over time, however, with repeated feedback, managers would learn θ_k, and eventually settle on a steady state effort level appropriate to their quality. Thus, as managers become more experienced, the predictions would converge to those of the baseline model with time invariant a_k.

We can capture this case with an alternative version of our model, in which type depends partly on endogenous manager effort $e_{k,t}$. Time stationarity is ensured by assuming that managers have already learned the immutable component of type, θ_k, with certainty, and by assuming a stationary environment in terms of the distribution of other managers' θ's, so there is no learning.[29] In this version, managers can choose an effort level $e_{k,t}$ in each period. This, combined with their (known) underlying θ_k generates $a_{k,t}$: $a_{k,t} = g(e_{k,t}, \theta_k)$.[30] Individuals are fully informed about their characteristics and fully informed about the characteristics of all other managers (the distribution of which is time invariant). Given a distribution of other managers' θs and effort levels in a given time period (which we denote θ_{-k} and $e_{-k,t}$) any given individual has a best response function that provides an optimal level of effort $e^*_{k,t}(\theta_k, \theta_{-k}, e_{-k,t})$. If there exists a pure strategy Nash Equilibrium which the managers play every period then $e^*_{k,t}(\theta_k, \theta_{-k}, e_{-k,t})$ is time invariant.[31] Thus, $a_{k,t}$ is also time invariant. Similarly, if the managers play the same mixed strategy Nash Equilibrium, then every period the predicted distribution of effort levels is stationary.[32] Although managers know their types and there is no learn-

[28]If managers are Bayesian this is true even if some of the managers who learn that they are low types leave the company and are missing from the sample of experienced managers; managers who remain should still be learning and have relatively precise predictions.

[29]Without learning, overconfidence must be driven by priors.

[30]Because we have only 5 types, this implies a coarseness of the mapping of θ_k and $e_{k,t}$ to type.

[31]This will happen, for example, whenever the one-shot version of the game has a unique pure-strategy equilibrium.

[32]In the case of mixed strategy equilibria it is important for identification that the realization of manager randomization for period t only occurs at the very end of period $t-1$. Then managers' predictions

ing process, managers can still make prediction errors, as tournament outcomes are a stochastic function of ability. This version of the model is probably not applicable to inexperienced managers, who may still be learning about their types and adjusting effort over time, but is a more plausible description for experienced managers who have observed a substantial number of signals. For this reason, we evaluate this version of the model by looking at the sub-sample of experienced managers, and dropping signals from early in these managers' tenures when we form predictions.

Specifically, we take the estimated $\hat{P}$ and the bootstrapped $\widetilde{P}_n$'s from the baseline analysis. We then drop all managers who have fewer than 8 periods of signals (the median manager has 10 periods of signals). For all other individuals, those who have strictly more than 8 signals, we estimate behavior dropping their first 8 signals. As shown in Table N1, manager predictions were overconfident relative to the model predictions in this case: 49% predicted a higher quintile than the model, compared to 17% predicting a lower quintile. Furthermore, the distance of manager predictions from the model predictions is far in the tail of the bootstrapped failure rates (see text for more details on bootstrapping) and it is possible to reject the model at the 1-percent level. Results are similar dropping the first 4 signals for all managers (and dropping all managers with fewer than 4 signals). It is also possible to reject at the 1-percent level that the model can explain the larger fraction of overconfident versus underconfident predictions for managers.

Another source of within-manager non-stationarity could be the switching of managers from one store to another over time, if store characteristics matter for performance. A corresponding robustness check uses only the signals from a manager's store as of Q4 of 2015 to form predictions; signals from previous stores are not used. Relative to this benchmark, 48% of managers were overconfident, compared to 22% being underconfident, and the model can be rejected at the 1-percent level.

Another identifying assumption is that there are no shocks to the informativeness of the type-to-signal matrix P over time; that is $P_t = P$ for all t. In contrast to the previous concerns, which were about individual non-stationarity, this concern is about environmental non-stationarity. P is not observable directly, but if the observable Z_t's indicate that the signal-to-signal matrix Z is not time-invariant, this would imply that P is not time-invariant. There is little evidence for time variation looking at the Z_t's (Appendix H).[33] As an additional robustness check, the model can be re-calculated

about effort in t (and so ability and thus signals) will occur before the realization of the strategy for period t, and so should agree with the time-averaged predictions generated by the model. If the strategy is realized before the managers make their predictions, then stationarity will be violated. Similarly, if managers switch between one shot Nash equilibrium across periods then our stationarity assumptions would be violated.

[33]Recall a time invariant P implies that Z is symmetric. However, if P is time varying, then Z may not

using all quarters to estimate P, but only signals from the last three quarters to estimate manager types. Compared to this benchmark based on recent signals, 44% of managers predicted a higher quintile, compared to 24% predicting a lower quintile, and the model is different from the data at the 1-percent level. An additional robustness check is re-estimating P but using only the three most recent transitions matrices to estimate P, from Q1, Q2, and Q3 of 2015. Moreover, we only use signals from Q1 to Q3 of 2015 to update managers' beliefs. Table N2 shows the estimated $\hat{P}$. The estimated $\hat{P}$ is very similar to the estimate we obtain when using the full sample. In this case there is little value in bootstrapping the model because the sample is too small to make this viable, but there are similar results in terms of manager predictions being overconfident relative to the model predictions. Results are also similar if we construct predictions that omit signals from Q3 of 2015, or from quarters with regional tournaments.

Table N2: Estimated matrix $\hat{P}$ using only the most recent 3 periods

	Signal					
	1	2	3	4	5	Total
Type 1	0.657	0.248	0.044	0.044	0.008	1
Type 2	0.260	0.304	0.297	0.118	0.020	1
Type 3	0.055	0.293	0.380	0.180	0.093	1
Type 4	0.021	0.105	0.160	0.479	0.236	1
Type 5	0.008	0.050	0.120	0.180	0.642	1
Total	1	1	1	1	1	

Notes: Estimated probability distributions across signals, by type. Signals correspond to quintiles in the performance distribution, with 5 being the best. Types are ordered from lower to higher ability. Rows sum to 1 because these are probability distributions. Columns sum to 1 because types are uniformly distributed.

O Augmenting the structural model with choice errors

The estimation technique allows for noise in the estimates of P. It assumes, however, that given a P and a sequence of signals (and so a posterior belief), an individual always bets on the signal that has the highest chance of occurring. It is possible that individuals may not always choose this, due to some form of choice errors (bounded rationality).[34]

be symmetric. Thus we looked at all entries in the Z matrix.

[34]As discussed in the text, there could be fully rational reasons why a manager would not bet on the most likely signal, namely a hedging motive. Managers receive a higher payoff when they obtain a higher signal, due to the workplace incentives. At the same time the lab in the field study offers a payment if a given signal arises. As discussed above, this implies that individuals may have insurance or hedging

In particular, it could be that managers are subject to choice errors, as in the typical discrete choice model (e.g. McFadden, 1974). Drawing on this literature, one could suppose that given a belief vector about the probability of signals $g_{k,\tau}$ individuals bet on signal j with probability $\frac{e^{\lambda g_{k,\tau}(j)}}{\sum_{\hat{j}} e^{\lambda g_{k,\tau}(\hat{j})}}$. Here $\lambda > 0$ is a parameter that captures how "random" choice is. If $\lambda = 0$ then each signal is bet on with a uniform chance. As $\lambda \to \infty$ the signal that has the highest chance of occurring is chosen with certainty. To incorporate such errors into the model, every time we estimate $\hat{P}$ and each $\widetilde{P}$, and then construct beliefs in period τ, we draw betting behavior from the distribution induced by $g_{k,\tau}$ and the probability distribution $\frac{e^{\lambda g_{k,\tau}(j)}}{\sum_{\hat{j}} e^{\lambda g_{k,\tau}(\hat{j})}}$. We then use these simulations of betting vector to construct our distances for the significance test of the model.

Observe that as $\lambda \to 0$ the data we observe must be rationalized. This is because in the limit each betting vector will simply be a draw from a uniform distribution over each signal. Thus, we expect, for any given individual that there is an 80% chance that betting predictions from the model, $b_{k,\tau}(\hat{P})$, and manager bets, $b_{k,O}$, disagree. Similarly, there is always a .8 chance that baseline model predictions, $b_{k,\tau}(\hat{P})$, and any simulation of betting behavior, $b_{k,\tau}(P)$, disagree. Importantly, the distance between $b_{k,\tau}(\hat{P})$ and $b_{k,O}$ changes with λ, as well as the distribution of simulated distances $\tilde{d}$. Thus, for each λ we consider we compare each simulation to the average across all simulations, and look at the upper end of the distribution of both the distance and the difference between over and underconfident behavior. We similarly compute both those statistics comparing the average simulation to observed manager behavior.[35] For $\lambda = 0, 1, 10, 100, 1000$ the maxima of the distance distributions across simulations are 258.8, 253.14, 175.36, 140.1 and 134.35 respectively, and the maximum differences between over and underconfident behavior are .11, .08, .09, .28 and .28 respectively.[36] The distances between actual behavior and the average simulation are 191.02, 189.95, 189.76, 204.97, 205.45; and the differences between over and underconfident behavior (comparing actual behavior to simulated behavior) are .20, .21, .19, .19 and .19. Thus,

motives: risk averse individuals will want to smooth earnings across different potential signal realizations. Although it is commonly observed that subjects narrowly bracket these kinds of experimental payoffs, which would imply that subjects ignore insurance motives, it is still important to discuss the implications. Importantly, insurance motives would lead subjects to bet on signals that otherwise pay less. This goes in the direction of underconfidence. Given that the data show overconfidence, it seems that managers are unlikely to be distorting bets due to hedging, and to the extent that they are, this strengthens the conclusion that managers are overconfident.

[35]In order to compute the over or underconfidence of a betting vector with only 1's and 0's as entries (e.g. observed behavior) and a vector with entries anywhere between 0 and 1 (e.g. an average simulated vector), we denote the entry in the former vector that contains a 1 as E. Then we compute compute the probability mass above, and below, E in the second vector.

[36]We do not explore $\lambda > 1000$ because numerical estimation procedures run up against the issue that larger λs imply that the objective function is not very smooth — a small change in beliefs can induce a large change in betting behavior.

for λ small enough (we find approximately the cutoff is approximately 5 when looking at a grid of λs) the observed distance is within the observed distribution of distances.[37] However, for small λs the model fails to account for the amount of overconfidence in the data. Similarly, although larger λs can potentially generate skewness in terms of over versus underconfidence (although not on average), they fail to account for the size of the overconfidence.

In conclusion, 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. [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 Q.

P Augmenting the structural model with private signals

In this section we give details on how we incorporate the possibility of private signals into the structural model, and the extent to which they can rationalize behavior in our data.

Given a posterior belief vector $f_{k,\tau}$, derived using the public signals, we suppose that each individual also observes one private signal in the final period before making predictions. In fact, even if managers receive a sequence of conditionally (on type) i.i.d. private signals (and they all receive the same number of signals), then without loss of generality we can simply reclassify each sequence of signals as a single private "signal."[39] Thus, our approach is quite general. Moreover, recall that we suppose that

[37]The degree of randomness in choice for $\lambda < 5$ is, however, rather extreme. For example, suppose $\lambda = 5$ and a manager knows he or she is the best type. In this case the probability of the best signal is 60%, wheres the next most likely signal occurs with only 20% chance. Choice error induced by $\lambda = 5$ causes the individual to only choose the best signal with probability 70-75%, despite it being at least three times as likely as other signals.

[38]An alternative hypothesis is that individuals bet incorrectly due to a distortion of probabilities unrelated to motivated beliefs. For example, distorting 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]Such a trick would not be possible if we had asked individuals to bet multiple times across different quarters, since then we would need to take a stand on how much private information they had gained

there are 5 potential signals (as we mentioned in the body of the text, supposing 5 signals is sufficient to test whether private information can generate our result).

The private signal structure is summarized by Q, a 5 by 5 type-to-signal matrix where $Q_{i,j}$ gives the probability that type i observes signal j. The new posterior about a manager's type, after receiving the private signal σ_k, is denoted $\dot{f}$ and given by:

$$\dot{f}_{k,\tau}(i|\sigma_k) = \frac{f_{k,\tau}(i)Q_{i,\sigma_k}}{\sum_{\hat{i}} f_{k,\tau}(\hat{i})Q_{\hat{i},\sigma_k}}$$

The belief about next period's performance quintile, conditional on the given private signal, is then

$$\dot{g}_{k,\tau}(j|\sigma_k) = \sum_i \dot{f}_{k,\tau}(i|\sigma_k)P_{i,j}$$

For technical reasons, we add a version of the discrete choice rule discussed above, where individuals make errors in which choice they make, conditional on beliefs. This ensures that there is a smooth mapping between Q and the bets. The distribution of choice probabilities conditional on a given private signal is

$$\gamma_{k,\tau}(s|\sigma_k) = \frac{e^{\lambda \dot{g}_{k,\tau}(s|\sigma_k)}}{\sum_{\hat{s}_{k,\tau}} e^{\lambda \dot{g}_{k,\tau}(\hat{s}|\sigma_k)}}$$

The goal is to estimate Q that brings choice behavior predicted by the model as close to the data as possible. We do not observe the realizations of private signals, however, to feed into the model and generate choice predictions, so instead we average across signals to generate the expected values of manager choices. The expected choice behavior is derived from averaging across different possible private signals and their associated choice probabilities, and averaging across the possible types. This expectation is given by

$$\dot{\gamma}_{k,\tau}(s) = \sum_i f_{k,\tau}(i) \sum_{\sigma_k} \gamma_{k,\tau}(s|\sigma_k)Q_{i,\sigma_k}$$

We then turn to estimating Q by minimum distance estimator. This means we estimate the Q that minimizes the distance between what the model predicts the managers do on average and actual betting behavior. Specifically, the Euclidean distance for a given individual is: $\sum_{\tilde{s}}(\dot{\gamma}_{k,\tau}(\tilde{s}) - b_{k,O}(\tilde{s}))^2$. We can then sum over all individuals to obtain $\sum_k \sum_{\tilde{s}}(\dot{\gamma}_{k,\tau}(\tilde{s}) - b_{k,O}(\tilde{s}))^2$. We estimate Q to minimize this.

We have fewer restrictions on Q than on P: We simply need the rows to sum to 1, i.e., $\sum_j Q_{i,j} = 1$. The columns do not need to sum to 1. As happened when estimating the P matrix, it is difficult to analytically prove identification because our objective

between the two elicitations.

function is not well behaved (i.e., not globally concave). In order to verify our solution we randomly generate 1,000 initial matrices on which to begin our estimation procedure of Q.[40] For each, we then numerically solve the constrained (potentially local) minimization problem and find the associated Q. We then consider the 100 initial matrices whose solution generates the smallest distances between observed behavior and the model-predicted behavior, and their associated Qs (observe that these Qs may not be unique). We focus on the solution that generates the smallest distance, but the statements regarding the lack of fit between observed data and the model predictions are true for all 100 of the Qs that have the smallest distances. Below we provide the best fitting Q when $\lambda = 1000$ to run our estimation (as discussed previously, larger λs can generate computational problems as the objective function becomes much less smooth).

Table P1: Estimated matrix Q for the baseline model

	Signal					
	1	2	3	4	5	Total
Type 1	0.014	0.014	0.269	0.508	0.194	1
Type 2	0.163	0.001	0.394	0.351	0.091	1
Type 3	0.411	0.009	0.331	0.072	0.177	1
Type 4	0.158	0.113	0.375	0.093	0.261	1
Type 5	0.113	0.434	0.136	0.097	0.219	1

Notes: Estimated probability distributions across private signals, by type, assuming relatively small choice errors ($\lambda = 1000$). Types are ordered from lower to higher ability. Rows sum to 1 because these are probability distributions. Columns need not sum to 1.

We next turn to understanding whether the private information structure we estimate can help rationalize the observed behavior. The estimated Q does not allow the model to match the data exactly. But since private signals are generated probabilistically, it could be that manager predictions are different from the expected value due to a particular realization of private signals that is different from the average due to chance. To assess whether the difference between the model and manager predictions falls within the bounds of this randomness, we simulate the model.

To conduct simulations we first want to come up with the probability of individual k getting private signal $\varsigma_k = 1, 2, 3, 4, 5$, given a posterior belief vector about types (derived using all the public signals) of $f_{k,\tau}$, and our estimated matrix Q. The proba-

[40]One could also imagine doing a grid search over all possible values to find the minimum. However, given the number of parameters we need to estimate, even with a coarse grid, such an approach is not feasible.

bility that an individual with distribution over types $f_{k,\tau}$ observes signal $\varsigma_{k,t}$ is $r(\varsigma_k) = \sum_i f_{k,\tau}(i) Q_{i,\varsigma_{(k,t)}}$.

For each simulation, for each individual, we can conduct a draw from this distribution. Call the simulated signal $\varsigma_k^{Sim,n}$ where n denotes the simulation. We can then conduct Bayesian updating using this signal, using $\mathring{f}$ to denote the beliefs after the private signal:

$$\mathring{f}_{k,\tau}(i|\varsigma_k^{Sim,n}) = \frac{f_{k,\tau}(i) Q_{i,\varsigma_k^{Sim,n}}}{\sum_{\hat{i}} f_{k,\tau}(\hat{i}) Q_{\hat{i},\varsigma_k^{Sim,n}}}$$

The belief about next period's performance quintile is then

$$\mathring{g}_{k,\tau}(j|\varsigma_k) = \sum_i \mathring{f}_{k,\tau}(i|\varsigma_k) P_{i,j}$$

In order to ensure consistency with our estimating model, we again allow for choice errors in the betting behavior as in models of logit choice. Results are very similar if we exclude choice errors when we simulate behavior. Thus, given the betting behavior $b_{k,\tau}$ (a 1 by 5 vector) the entry j equals 1 if $\mathring{g}_{k,\tau}(j|\varsigma_k) = \max_{\hat{j}} \mathring{g}_{k,\tau}(\hat{j}|\varsigma_k)$ and 0 otherwise.

A particular realization of private signals could have generated manager bets that are different from the predictions of average betting behavior generate by the model. Thus, we want to assess whether any difference between actual and expected betting behavior lies within the bounds of the noise entailed in the private signals (and also the noise coming from public signals). We do this by simulating the model 100 times. For each, we start with one of the 100 bootstrapped $\widetilde{P}$s that we estimated for the baseline model. This incorporates noise in posterior beliefs about manager type that arises from the random component of public signals, and also generates, via Q an associated probability distribution over the possible private signals. For each of these sets of public posteriors we add noise from private signals, by drawing from the appropriate probability distribution over private signals, updating posteriors, and calculating betting behavior. With the 100 simulations in hand we find the average betting vector induced for each individual across all simulations, for a distribution of average betting behavior.[41] We then calculate: (i) the distances between these average simulated betting vectors and the observed betting vectors, and sum across all individuals; and (ii) for each simulation, the distances between the betting vectors for that simulation and the average simulated betting vectors, summed across all individuals. We then compare the distance calculated in (i) to the distribution of 100 distances derived in (ii). The distance of the observed betting behaviors from the average simulated behavior, which

[41]In the limit this average is the same as the expected choice (i.e., betting behavior), $\dot{\gamma}_{k,\tau}(s)$, derived above which was used to estimate Q.

is 200, lies beyond the 99th percentile of the cumulative distribution of distances of the simulated betting vectors, and thus we reject the model at the 1-percent level. We similarly compare the difference between the fraction of overconfident managers and the fraction of underconfident managers to the distribution of fractions derived from the simulations. The observed difference lies beyond the 99th percentile of the distribution of simulated differences.

Q Augmenting the structural model with biased memory

In this section we provide details on estimation of the structural model with biased memory. As a first step we check whether manager overconfidence (and underconfidence) relative to the baseline structural model goes hand in hand with biased memory of past performance. Table Q1 shows that this is indeed the case, which suggests incorporating heterogeneity in biased memory may help the model explain heterogeneity in overconfidence.

We incorporate a technology for memory distortion in the form of a memory matrix M. If a manager is motivated to distort memories, $M_{\kappa,j}$ gives the probability that, conditional on having actually observed signal κ in period t, the individual remembers it as signal j (here a "signal" is the quintile of performance). Thus, the rows of M must sum to 1. We use the empirical frequencies from the data on manager recall to calibrate the probabilities. M is displayed in Table Q2.

All managers are assumed to have access to the same M, but only managers who are "motivated" will use M to distort memories of past signals. Managers who are "unmotivated" do not use M and always remember signals correctly. Managers are assumed to update beliefs based on remembered signals (the remembered signal could be the same as the actual signal).

To assess how close the model comes to matching the data we start from the 100 bootstrapped $\tilde{P}$'s used for the baseline model (so that the model incorporates the noise in the actual signals), and for each bootstrap, simulate memories for each manager who is motivated to distort. We index the number of the simulation with ι. In each simulation, for each individual, for each signal, $s^{k,t}$ we look at row $s^{k,t}$ in M. We then conduct a draw among the columns of M using the distribution induced by row $s^{k,t}$ of M. This generates a remembered signal $\widehat{s}^{k,t,\iota}$. We repeat this process for each signal, for each individual, till we have generated for each k a set of remembered signals $\{\widehat{s}^{k,t}\}^{\iota}$. Managers who are unmotivated to distort remember all signals correctly.

Table Q1: Overconfidence and underconfidence relative to the structural model as a function of biased memory

	Overconfident (rel. to structural)		Underconfident (rel. to structural)		Manager prediction - structural model prediction	
	(1)	(2)	(3)	(4)	(5)	(6)
Flattering memory about Q2 of 2015	0.12*	0.17**				
	(0.07)	(0.07)				
Unflattering memory about Q2 of 2015			0.13*	0.17**		
			(0.07)	(0.07)		
Recalled minus actual performance					0.26**	0.32***
					(0.13)	(0.12)
Performance percentile in Q2 of 2015		0.08*		-0.02		0.28**
		(0.04)		(0.04)		(0.13)
Performance percentile in Q3 of 2015		0.02		-0.08**		0.23*
		(0.04)		(0.04)		(0.13)
Mean performance percentile pre- Q2 of 2015		-0.22***		0.15***		-0.77***
		(0.03)		(0.04)		(0.11)
Female		-0.15**		0.04		-0.27
		(0.07)		(0.07)		(0.19)
Age		-0.06		0.04		-0.14
		(0.05)		(0.04)		(0.11)
Experience		0.02		0.04		0.04
		(0.05)		(0.04)		(0.12)
Constant					-0.19*	0.02
					(0.11)	(0.16)
Observations	174	148	174	148	174	148
Pseudo R^2	0.011	0.204	0.015	0.173	0.008	0.119

Notes: Columns (1) to (4) present marginal effects of Probit regressions. Columns (5) and (6) are marginal effects from interval regressions. The dependent variable in columns (1) and (2) equals 1 if a manager's prediction was overconfident relative to the baseline structural model prediction and zero otherwise. The dependent variable in columns (3) and (4) is the corresponding indicator for underconfidence. The dependent variable in columns (5) and (6) is the manager prediction about the most likely quintile in Q4 of 2015 minus the prediction of the model. Independent variables are standardized so the coefficients show the impact of a 1 s.d. increase in the independent variable. Robust standard errors are in parentheses.

Table Q2: Memory matrix M

	Empirical frequencies					Total
Actual Q2 signal						
1	0.46	0.23	0.03	0.29	0.00	1
2	0.15	0.25	0.10	0.33	0.18	1
3	0.13	0.16	0.24	0.22	0.24	1
4	0.02	0.10	0.02	0.36	0.50	1
5	0.08	0.00	0.00	0.08	0.84	1
	1	2	3	4	5	
	Recalled Q2 signal					

Notes: The entries are based on empirical frequencies of managers remembering different performance quintiles for Q2 of 2015, conditional on being in a given actual quintile.

To specify how managers who distort memory update beliefs, we need to make a distinction between sophisticated or naïve managers. If individuals are naïve then we conduct Bayesian updating using a uniform prior about manager types, $f^{k,0}$ and the bootstrapped $\tilde{P}$, just as we did for the baseline model, but using remembered signals $\{\widetilde{s}^{k,t}\}^{\iota}$ rather than actual signals. This generates, ultimately, a vector $\hat{b}^{k,\iota}(\hat{P})$ for each individual k, for particular simulation ι.

Sophistication adds a twist: now individuals know they distort their memories. Thus, they observe their recalled signal, but know that it isn't what actually happened. They then try to backwards induct what actually happened, and update according to that. Suppose an individual remembers signal j. The probability that they actually observed signal κ is $\omega_{\kappa,j} = \frac{M_{\kappa,j}}{\sum_{\hat{\kappa}} M_{\hat{\kappa},j}}$. Given a prior $f^{\iota}_{k,t}(i)$ the posterior belief about being type i if they remember signal j is the average posterior over all the signals they could have observed:

$$f^{\iota}_{k,t+1}(i) = \sum_{\hat{\kappa}} \omega_{\hat{\kappa},j} \frac{f^{\iota}_{k,t}(i) P_{i,\hat{\kappa}}}{\sum_{\hat{i}} f^{\iota}_{k,t}(\hat{i}) P_{\hat{i},\hat{\kappa}}}$$

We start out at a uniform prior $f^{\iota}_{k,0} = .2$ and then simply iterate forward to period τ.[42] We then obtain our posterior beliefs in period τ, $f_{k,\tau^{\iota}}$, as well as $g^{\iota}_{k,\tau}$ and betting behavior $b^{\iota}_{k,\tau}$.

The prevalence of memory distorters, naïve or sophisticated, and potentially also individuals who are unmotivated to distort, is an empirical question. Rather than just impose an assumption about these, we use the data to infer which assumption best describes each manager. For each individual we check to see goodness of fit of each of the three different assumptions. We start with the 100 bootstrapped $\tilde{P}_n$'s from the

[42]Unfortunately, there is no closed form way to write this out.

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. Specifically, we conduct 100 simulations for an individual under the assumption that the individual is naïve. For each simulation, we suppose that an individual randomly replaces their true signal with a remembered signal (and uses one of the bootstrapped $\tilde{P}$ matrices derived when testing the baseline model). Second, we conduct the same exercise, but under the assumption of sophistication. Last, we conduct the same exercise, but supposing individuals remember their signals perfectly. For each individual we then pick out the assumption that best matches behavior, i.e., generates the smallest Euclidean distance between the average predicted bet across all 100 simulations and the observed behavior. We find that there are 85 naïve, 67 sophisticated and 61 "unmotivated" managers.[43]

Having assigned categories, we then re-run the 100 simulations for all managers, having managers distort memories, and update beliefs, according to their category. 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. We find the average betting vector induced for each individual across all simulations, i.e., the expected choice behavior. We then find (i) the sum of distances between managers' average simulated betting vectors and managers' actual bets, and (ii) for each simulation, the sum of distances between managers' average simulated betting vectors and managers' bootstrapped betting vectors. The distance calculated in (i) turns out to be 135. Comparing this to the distribution of distances generated in (ii), we cannot reject the model at conventional significance levels. Similarly, we cannot reject at conventional levels that the model could generate the observed fraction of managers who are overconfident minus the fraction underconfident. Results are discussed in the text.

One concern is that the model comes closer to the data because the various sources of heterogeneity give extra degrees of freedom. We therefore also checked robustness to

[43]Our approach supposes that there are individuals who are unmotivated to distort their memory in a biased way, according to the data we observe. However, the memory matrix we use includes the memories of all individuals, including those we classify as unmotivated. We could, alternatively, try to only use, in our memory matrix, those individuals who we do not classify as unmotivated. However, even individuals who we classify as ummotivated may still have imperfect memories, so long as they are roughly "symmetric" around the true memories. Thus, it isn't necessarily clear whether to drop all memories of individuals who we classify as unmotivated or only those who also remember correctly. In order to ensure that our assumptions "distort" the inputs into our model as little as possible, we simply keep all managers in the memory matrix.

restricting heterogeneity in various ways: (1) optimally assigning managers to be either sophisticated or naïve, using the procedure described above, but without the possibility of unmotivated managers; (2) assuming all managers are motivated and naïve; (3) assuming all managers are motivated and sophisticated; (4) randomly assigning the three categories of naïve, sophisticated, and unmotivated. The resulting distances are 161, 188, and 190 for (1) to (3), respectively. Across a range of different proportions for (4) the distances lie between 185 and 190. The assumption of 100 percent sophisticates (a case of zero heterogeneity in manager types, with the worst fit) is still better than the baseline structural model or structural model with private information, where the distances are at least 200.

Related perceptual biases

We can also potentially detect the effects of memory distortions, or more generally, perceptual distortions of the environment by attempting to estimate what kind of P individuals use. To make this connection concrete, suppose that there are three signals, L, M and H. Suppose that individuals distort memories of L up to memories of M, and always remember the other two signals correctly. Thus, if we supposed individuals correctly remembered signals, they would act as if they are using the incorrect P. The P matrix that best fits their behavior would involve mixing the columns of L and M in the true P matrix. Thus, individuals using a different P than estimated for the baseline model can be indicative of distorting memories, when direct data on memories is not available. Of course, individuals could also directly distort how they think remembered signals transform into predictions, exhibiting a form of motivated beliefs that does not work through memory. In other words, they understand the signals they observe, update according to Bayes' rule, but have a motivation to conceive the signal as conveying information different from what our objective estimation procedure says it does.

In order to understand whether this could explain behavior, we take an conservative approach, assuming managers mis-perceive P in a way that is as favorable as possible to the model. Specifically, we estimate a new matrix P that comes closest to being able to explain observed manager bets, denoted $\ddot{P}$. We then ask how close the model comes to explaining behavior, when we assume managers use $\ddot{P}$, in conjunction with Bayesian updating.

Recall that the posterior belief for individual k attached to a given signal $s_{k,\tau}$ in period τ is

$$\sum_i \frac{\Pi_{t=\tau'}^{\tau-1} P_{i,s_{k,t}}}{\sum_{\hat{i}} \Pi_{t=\tau'}^{\tau-1} P_{\hat{i},s_{k,t}}} P_{i,s_{k,\tau}}$$

If we suppose that an individual always bets on the signal that has the highest probability, we have problems estimating this equation, as the chance of choosing a particular signal jumps from 1 to 0 (or vice versa). Thus, in order to suppose we have a smooth function to estimate, we revisit incorporating choice errors. We suppose that the chance that an individual chooses a signal $s_{k,\tau}$ with probability

$$\gamma_{k,\tau}(s) = \frac{e^{\lambda \sum_i \frac{\Pi_{t=\tau'}^{\tau-1} P_{i,s_{k,t}}}{\sum_{\hat{i}} \Pi_{t=\tau'}^{\tau-1} P_{\hat{i},s_{k,t}}} P_{i,s_{k,\tau}}}}{\sum_{\hat{s}_{k,\tau}} e^{\lambda \sum_i \frac{\Pi_{t=\tau'}^{\tau-1} P_{i,s_{k,t}}}{\sum_{\hat{i}} \Pi_{t=\tau'}^{\tau-1} P_{\hat{i},s_{k,t}}} P_{i,\hat{s}_{k,\tau}}}}$$

For any individual k and matrix P, we will compute the Euclidean distance between the $\gamma_{k,\tau}$ and the actual behavior: $\sum_{\widetilde{s}}(\gamma_{k,\tau}(\widetilde{s}) - b_{k,O}(\widetilde{s}))^2$. We can then sum over all individuals to obtain $\sum_k \sum_{\widetilde{s}}(\gamma_{k,\tau}(\widetilde{s}) - b_{k,O}(\widetilde{s}))^2$. We then find the P that minimizes this distance subject to the same constraints as in the baseline model. Again, as $\lambda \to \infty$ we recover the fully rational choice model. For $\ddot{P}_\lambda$ we then calculate $\sum_k D(b_{k,\tau}(\ddot{P}_\lambda), b_{k,O})$, We then conduct a bootstrapping procedure to generate confidence intervals for our estimates. Because our data are now individual manager's predictions, our bootstrapping procedure involves randomly sampling (with replacement) from the set of managers. Once we have our bootstrapped sample, we re-estimate the information matrix, and derive behavior. The bootstrapped behavior is then compared to the behavior derived using the full set of managers in the same way as previously (i.e. the distance is the sum over all managers of the Euclidean distance between the induced betting vectors (behaviors). We use $\lambda = 1000$. As discussed previously, larger λs can generate computational problems as the objective function becomes much less smooth.[44] We find that the distance of observed behavior to the predicted behavior is 181.02; the fraction overconfident in the data is 0.33, and the fraction underconfident is 0.27, implying the gap is 0.06. This falls at the 99th percentile of the distribution of distances when comparing the bootstraps to the baseline predictions, and the difference between over and underconfidence is at the 87th percentile. Notably, however, the bootstrapped gaps between over and underconfident range from -0.37 to 0.31, implying a lot of noise in the estimates. Overall this model seems to match the data somewhat less well than the memory model, as well as generating a larger "noise" in the simulation procedure.

[44]When we derive behavior, we do so supposing that individuals are fully rational, in other words, we only use λ as a nuisance parameter to assist with the estimation.

R Overconfidence and future performance

This section explores whether manager overconfidence in the present is related to better or worse performance in future quarters. On the one hand, overconfidence in the present might be associated with worse performance in subsequent quarters, because making decisions based on biased beliefs leads to mistakes (Brunnermeier and Parker, 2005). On the other hand, some models predict that overconfidence could have offsetting benefits for some aspects of performance, for example if it counteracts self-control problems, or otherwise improves the production function for performance (Benabou and Tirole, 2002; Compte and Postlewaite, 2004). Notably, negative and positive effects need not be be mutually exclusive, in that overconfidence might be associated with better outcomes for some aspects of performance and worse outcomes for others. It is important to keep in mind, however, the caveat about endogeneity of overconfidence in the case of motivated beliefs, discussed at the end of the main text.

The empirical strategy is to regress a manager's standardized performance percentiles in future quarters (Q1 and Q2 of 2016) on manager predictions about Q4. With controls for past tournament outcomes in the regressions, the coefficient on manager predictions captures overconfidence (or underconfidence) in manager beliefs relative to what one would have predicted for Q4 using past public signals. We also try specifications in which we collapse manager predictions into various binary indicators for overconfidence relative to different predictors for Q4: Historical mode, multinomial logit, and baseline structural model prediction. The analysis does not include Q4 performance in the dependent variable, because predictions were formed partway into Q4; some of the variation in Q4 captures information that informed manager predictions, so using Q4 as a dependent variable would raise reverse causality concerns. To rule out that overconfident managers have different future performances because they switch to different types of stores over time, the analysis is restricted to managers who have the same store from Q4 of 2015 to Q2 of 2016. To account for the possibility that overconfident managers are assigned to stores with systematically different characteristics during that time period, the regressions control for additional store characteristics. We have also explored whether manager characteristics, including overconfidence, are correlated with store characteristics, and find little evidence for this. Although Q4 performance may be partly an outcome of confidence about Q4, we include Q4 performance as a control variable. This is intended to be conservative, and ensure that a relationship between manager predictions for Q4, and performance in 2016, is not just picking up Q4 performance, which predicts 2016 performance due to serial correlation. Results are similar, however, if the regressions exclude the control for Q4 performance.

Table R1 shows regressions explaining a manager's overall aggregate performances in Q1 and Q2 of 2016. Column (1) uses manager predictions as the key independent variable, controlling for performance in recent quarters and the mean of pre-Q2 performance. Variation is manager beliefs thus captures deviations relative to what one would predict based on recent and historical mean performance. Columns (2) to (4) use various binary indicators for overconfidence, relative to different benchmark predictions: historical mode, multinomial logit model, and baseline structural model predictions, respectively. These specifications also control for the levels of the corresponding predictors. The table shows that greater manager overconfidence about Q4 of 2015 did not have a statistically significant relationship to performance in early 2016, and point estimates are generally close to zero.

Table R2 through Table R5 present similar analyses for the four individual dimensions of performance that make up the overall performance measure. The results show that managers who made more confident predictions about Q4 of 2015 do have significantly different outcomes than other managers on these individual dimensions. Specifically, overconfidence is associated with significantly higher profits, but also lower customer service scores, with these two differences working in opposite directions and contributing to the weak relationship with aggregate performance. Statistical significance is weaker for some of the specifications using binary indicators, which may partly reflect the reduced variation in the explanatory variable that arises from binarizing manager predictions. It is possible that the relationship of manager predictions to future performance reflects some private information about early 2016, but this seems unlikely. The analysis has shown that it is difficult to explain manager predictions for Q4 of 2015 with private information. For sales growth and regional manager review scores, there is no statistically significant relationship to manager confidence. Although the results are correlational, the findings are consistent with overconfidence being a two-edged sword for manager performance, leading to better performance on some dimensions but worse performance on others.

Table R1: Overconfidence and future overall performance

	Performance percentile in 2016			
	(1)	(2)	(3)	(4)
Manager prediction about Q4 of 2015	0.029			
	(0.095)			
Overconf. rel. to mode		-0.248		
		(0.178)		
Overconf. rel. to mult. Logit			0.023	
			(0.292)	
Overconf. rel. to structural				0.088
				(0.156)
Performance percentile in Q2 of 2015	-0.039	-0.107	-0.270	-0.068
	(0.083)	(0.093)	(0.345)	(0.086)
Performance percentile in Q3 of 2015	-0.148	-0.123	-0.073	-0.156
	(0.102)	(0.102)	(0.207)	(0.097)
Performance percentile in Q4 of 2015	0.419***	0.486***	0.367**	0.410***
	(0.086)	(0.086)	(0.156)	(0.080)
Female	0.053	0.164	0.008	0.116
	(0.150)	(0.153)	(0.255)	(0.138)
Age	0.005	-0.060	0.146	0.019
	(0.084)	(0.096)	(0.149)	(0.077)
Experience	-0.122	-0.089	-0.284	-0.129
	(0.093)	(0.098)	(0.248)	(0.090)
Quarter shop opened	-0.010	-0.005	-0.007	-0.011*
	(0.007)	(0.007)	(0.014)	(0.006)
Mean performance percentile pre-Q2 of 2015	0.092			
	(0.071)			
Historical modal quintile		0.124		
		(0.117)		
Mult. logit predicted quintile			0.271	
			(0.282)	
Structural predicted quintile				0.208**
				(0.099)
Constant	2.492*	1.159	1.943	2.506**
	(1.304)	(1.363)	(2.591)	(1.187)
Additional store controls	Yes	Yes	Yes	Yes
Observations	227	191	113	249
Adjusted R^2	0.279	0.344	0.136	0.307

Notes: The table reports coefficients from OLS regressions. The dependent variable is the standardized value of performance percentile in Q1 or Q2 of 2016, so there are two observations per manager. Independent variables are standardized so coefficients show the impact of a 1 s.d. change in the independent variable in terms of s.d. of the dependent variable. The sample is restricted to managers who worked in all three quarters, Q4 of 2015 through Q2 of 2016, and excludes managers who switched stores, so that store characteristics are being held constant within manager over time. Column (1) uses the manager's prediction for Q4 of 2015 quintile as the indicator of manager overconfidence (underconfidence), controlling for past manager performance in Q3 and Q2, and the mean of pre-Q2 performance. Columns (2) to (4) use binary indicators for manager overconfidence about Q4 of 2015, relative to different benchmark predictors: Historical mode, multinomial logit, and baseline structural model. These models also control for the respective predictor. Additional store controls include dummy variables for the location of the store in terms of one of 38 different geographic areas, as well as age of the store. Robust standard errors are in parentheses, clustered on manager.

Table R2: Overconfidence and future profit

	Profit in 2016			
	(1)	(2)	(3)	(4)
Manager prediction about Q4 of 2015	0.142**			
	(0.068)			
Overconf. rel. to mode		0.079		
		(0.137)		
Overconf. rel. to mult. Logit			0.327**	
			(0.160)	
Overconf. rel. to structural				0.273**
				(0.117)
Performance percentile in Q2 of 2015	-0.106*	-0.094	-0.105	-0.098
	(0.064)	(0.074)	(0.234)	(0.072)
Performance percentile in Q3 of 2015	-0.070	-0.061	-0.245**	-0.079
	(0.068)	(0.076)	(0.104)	(0.070)
Performance percentile in Q4 of 2015	0.179**	0.252***	0.265**	0.183***
	(0.070)	(0.087)	(0.113)	(0.070)
Female	0.110	0.112	0.190	0.148
	(0.113)	(0.122)	(0.182)	(0.121)
Age	-0.012	-0.126	0.013	-0.036
	(0.073)	(0.080)	(0.111)	(0.068)
Experience	-0.013	0.041	-0.045	0.019
	(0.081)	(0.092)	(0.133)	(0.080)
Mean performance percentile pre-Q2 of 2015	0.152***			
	(0.045)			
Historical modal quintile		0.222***		
		(0.080)		
Mult. logit predicted quintile			0.328*	
			(0.195)	
Structural predicted quintile				0.313***
				(0.079)
Constant	3.491***	3.007**	1.629	3.524***
	(0.869)	(1.167)	(1.412)	(0.925)
Additional store controls	Yes	Yes	Yes	Yes
Observations	227	191	113	249
Adjusted R^2	0.328	0.343	0.284	0.324

Notes: The table reports coefficients from OLS regressions. The dependent variable is the standardized value of store profits in Q1 or Q2 of 2016, so there are two observations per manager. Independent variables are standardized so coefficients show the impact of a 1 s.d. change in the independent variable in terms of s.d. of the dependent variable. The sample is restricted to managers who worked in all three quarters, Q4 of 2015 through Q2 of 2016, and excludes managers who switched stores, so that store characteristics are being held constant within manager over time. Column (1) uses the manager's prediction for Q4 of 2015 quintile as the indicator of manager overconfidence (underconfidence), controlling for past manager performance in Q3 and Q2, and the mean of pre-Q2 performance. Columns (2) to (4) use binary indicators for manager overconfidence about Q4 of 2015, relative to different benchmark predictors: Historical mode, multinomial logit, and baseline structural model. These models also control for the respective predictor. Additional store controls include dummy variables for the location of the store in terms of one of 38 different geographic areas, as well as age of the store. Robust standard errors are in parentheses, clustered on manager.

Table R3: Overconfidence and future customer service score

	Customer service score in 2016			
	(1)	(2)	(3)	(4)
Manager prediction about Q4 of 2015	-0.174**			
	(0.088)			
Overconf. rel. to mode		-0.355**		
		(0.178)		
Overconf. rel. to mult. Logit			-0.323	
			(0.276)	
Overconf. rel. to structural				-0.177
				(0.134)
Performance percentile in Q2 of 2015	0.152	0.054	-0.138	0.053
	(0.095)	(0.128)	(0.351)	(0.093)
Performance percentile in Q3 of 2015	-0.105	-0.047	-0.167	-0.157
	(0.125)	(0.128)	(0.202)	(0.117)
Performance percentile in Q4 of 2015	0.222**	0.186**	0.233	0.239***
	(0.089)	(0.087)	(0.156)	(0.083)
Female	-0.114	-0.044	-0.184	-0.143
	(0.151)	(0.160)	(0.235)	(0.138)
Age	0.065	0.032	0.340**	0.078
	(0.104)	(0.118)	(0.131)	(0.095)
Experience	-0.020	-0.095	-0.330*	-0.065
	(0.096)	(0.116)	(0.197)	(0.093)
Mean performance percentile pre-Q2 of 2015	0.074			
	(0.061)			
Historical modal quintile		-0.007		
		(0.115)		
Mult. logit predicted quintile			0.283	
			(0.229)	
Structural predicted quintile				0.082
				(0.084)
Constant	2.754**	1.964	3.144	2.957**
	(1.315)	(1.302)	(2.507)	(1.263)
Additional store controls	Yes	Yes	Yes	Yes
Observations	227	191	113	249
Adjusted R^2	0.125	0.123	0.032	0.127

Notes: The table reports coefficients from OLS regressions. The dependent variable is the standardized value of customer service score in Q1 or Q2 of 2016, so there are two observations per manager. Independent variables are standardized so coefficients show the impact of a 1 s.d. change in the independent variable in terms of s.d. of the dependent variable. The sample is restricted to managers who worked in all three quarters, Q4 of 2015 through Q2 of 2016, and excludes managers who switched stores, so that store characteristics are being held constant within manager over time. Column (1) uses the manager's prediction for Q4 of 2015 quintile as the indicator of manager overconfidence (underconfidence), controlling for past manager performance in Q3 and Q2, and the mean of pre-Q2 performance. Columns (2) to (4) use binary indicators for manager overconfidence about Q4 of 2015, relative to different benchmark predictors: Historical mode, multinomial logit, and baseline structural model. These models also control for the respective predictor. Additional store controls include dummy variables for the location of the store in terms of one of 38 different geographic areas, as well as age of the store. Robust standard errors are in parentheses, clustered on manager.

Table R4: Overconfidence and future sales growth

	Sales growth in 2016			
	(1)	(2)	(3)	(4)
Manager prediction about Q4 of 2015	-0.006			
	(0.133)			
Overconf. rel. to mode		-0.226		
		(0.315)		
Overconf. rel. to mult. Logit			-0.290	
			(0.463)	
Overconf. rel. to structural				0.045
				(0.223)
Performance percentile in Q2 of 2015	-0.261**	-0.193	-0.141	-0.238*
	(0.117)	(0.126)	(0.479)	(0.122)
Performance percentile in Q3 of 2015	-0.034	-0.153	0.189	0.009
	(0.128)	(0.160)	(0.297)	(0.135)
Performance percentile in Q4 of 2015	0.473**	0.552***	0.519	0.424**
	(0.185)	(0.203)	(0.370)	(0.172)
Female	0.166	0.168	0.252	0.190
	(0.160)	(0.166)	(0.367)	(0.147)
Age	0.060	-0.048	-0.062	0.097
	(0.118)	(0.133)	(0.240)	(0.113)
Experience	-0.012	0.136	0.113	-0.034
	(0.130)	(0.173)	(0.350)	(0.121)
Mean performance percentile pre-Q2 of 2015	-0.143**			
	(0.072)			
Historical modal quintile		-0.248		
		(0.229)		
Mult. logit predicted quintile			-0.721	
			(0.717)	
Structural predicted quintile				-0.149
				(0.133)
Constant	-2.863	-4.586**	-9.917*	-1.792
	(1.884)	(1.983)	(5.730)	(1.638)
Additional store controls	Yes	Yes	Yes	Yes
Observations	227	191	113	249
Adjusted R^2	0.264	0.303	0.253	0.250

Notes: The table reports coefficients from OLS regressions. The dependent variable is the standardized value of sales growth in Q1 or Q2 of 2016, multiplied by 100, so there are two observations per manager. Independent variables are standardized so coefficients show the impact of a 1 s.d. change in the independent variable in terms of s.d. of the dependent variable. The sample is restricted to managers who worked in all three quarters, Q4 of 2015 through Q2 of 2016, and excludes managers who switched stores, so that store characteristics are being held constant within manager over time. Column (1) uses the manager's prediction for Q4 of 2015 quintile as the indicator of manager overconfidence (underconfidence), controlling for past manager performance in Q3 and Q2, and the mean of pre-Q2 performance. Columns (2) to (4) use binary indicators for manager overconfidence about Q4 of 2015, relative to different benchmark predictors: Historical mode, multinomial logit, and baseline structural model. These models also control for the respective predictor. Additional store controls include dummy variables for the location of the store in terms of one of 38 different geographic areas, as well as age of the store. Robust standard errors are in parentheses, clustered on manager.

Table R5: Overconfidence and future regional manager evaluation score

	Regional manager review score in 2016			
	(1)	(2)	(3)	(4)
Manager prediction about Q4 of 2015	-0.154			
	(0.152)			
Overconf. rel. to mode		-0.128		
		(0.275)		
Overconf. rel. to mult. Logit			-0.004	
			(0.360)	
Overconf. rel. to structural				0.055
				(0.217)
Performance percentile in Q2 of 2015	0.031	-0.022	0.378	-0.028
	(0.096)	(0.144)	(0.385)	(0.129)
Performance percentile in Q3 of 2015	-0.133	-0.153	-0.143	-0.199
	(0.117)	(0.128)	(0.204)	(0.164)
Performance percentile in Q4 of 2015	0.239*	0.237	0.051	0.156
	(0.128)	(0.144)	(0.151)	(0.147)
Female	0.164	0.236	0.221	0.173
	(0.224)	(0.247)	(0.353)	(0.232)
Age	-0.140	-0.104	-0.262	-0.067
	(0.129)	(0.154)	(0.183)	(0.132)
Experience	-0.128	-0.130	0.094	-0.101
	(0.132)	(0.158)	(0.220)	(0.136)
Mean performance percentile pre-Q2 of 2015	0.068			
	(0.087)			
Historical modal quintile		0.103		
		(0.133)		
Mult. logit predicted quintile			-0.153	
			(0.420)	
Structural predicted quintile				0.284*
				(0.154)
Constant	3.741	2.609	3.647	2.139
	(2.297)	(2.713)	(2.527)	(2.608)
Additional store controls	Yes	Yes	Yes	Yes
Observations	102	85	52	112
Adjusted R^2	0.161	-0.041	0.191	0.048

Notes: The table reports coefficients from OLS regressions. The dependent variable is the standardized value of the manager evaluation score in Q2 of 2016, multiplied by 100; the evaluation was not conducted in Q1 of 2016 so there is only one observation per manager. Independent variables are standardized so coefficients show the impact of a 1 s.d. change in the independent variable in terms of s.d. of the dependent variable. The sample is restricted to managers who worked in all three quarters, Q4 of 2015 through Q2 of 2016, and excludes managers who switched stores, so that store characteristics are being held constant within manager over time. Column (1) uses the manager's prediction for Q4 of 2015 quintile as the indicator of manager overconfidence (underconfidence), controlling for past manager performance in Q3 and Q2, and the mean of pre-Q2 performance. Columns (2) to (4) use binary indicators for manager overconfidence about Q4 of 2015, relative to different benchmark predictors: Historical mode, multinomial logit, and baseline structural model. These models also control for the respective predictor. Additional store controls include dummy variables for the location of the store in terms of one of 38 different geographic areas, as well as age of the store. Robust standard errors are in parentheses.

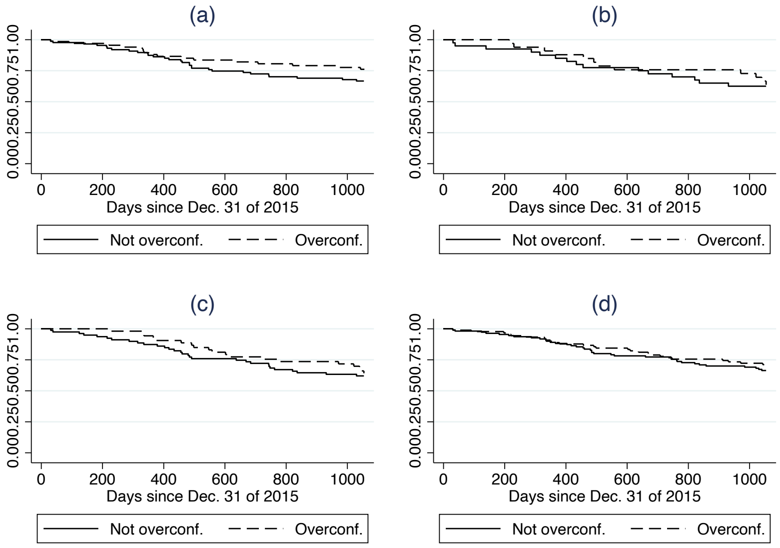
S Overconfidence and retention

This section explores whether various indicators for manager overconfidence are related to the tendency for managers to stay in or leave their jobs at the firm. If manager overconfidence causes managers to think that their earnings in the job will be greater than their outside option, this could lead to higher survival rates for overconfident managers. If manager overconfidence also affects beliefs about the outside option, however, then there might not be a strong relationship of overconfidence to survival rates.

The indicators for overconfidence are whether a manager's prediction for Q4 of 2015 was optimistic relative to a respective rule of thumb, reduced form predictor, or structural model predictor for Q4 of 2015. The analysis uses data on whether managers were still working in their same jobs at the firm in quarters subsequent to Q4 of 2015. We estimate Cox proportional hazard models, where the independent variable of interest is a binary indicator for being overconfident. The regressions also control for other factors that might affect survival rates: Past manager performance, captured by percentile of performance in Q4, Q3, and Q2 of 2015, as well as mean of percentile pre-Q2 of 2015; gender; age; store characteristics in terms of store age and store location.

The resulting estimated survival functions do show a tendency for overconfident managers to remain longer at the firm, for each of the different measures of overconfidence. These differences are relatively modest in size, however, and are not statistically significant at conventional levels.

Figure S1: Kaplan-Meier survival estimates as a function of overconfidence



Notes: The figure reports the estimated survival rates for managers since Dec. 31 of 2015, comparing managers who were overconfident about Q4 of 2015 relative to managers who were not overconfident. Panel (a) measures overconfidence relative to the historical mode predictor. Panels (b) and (c) measure overconfidence relative to the 8-lag and 3-lag multinomial logit predictors, respectively. Panel (d) measures overconfidence relative to the baseline structural model predictions. The estimates are from Cox proportional hazards models that control for other factors that might affect survival rates, such as past manager performance, gender, age, and store characteristics.

T Overconfidence and managerial style

This section tests additional hypotheses about how manager overconfidence might be related to managerial style, subject to caveats about sample size, and limited outcomes, mentioned in the discussion at the end of the main text.

First, the analysis looks at manager decisions about hiring assistant managers (AM's). For each store, the firm recommends hiring a particular number of assistant managers, typically 1 or 2, but the manager has some discretion over whether they comply with this recommendation. The hypothesis was that overconfident managers would be more likely to hire fewer than the recommended number assistant managers, because they are more confident in their abilities to manage the store without help.

Table T1 presents probit regressions in which the dependent variable is equal to 1 if a manager hired at least the recommended number of AM's, and 0 otherwise. The key independent variable is Column (1) is manager predictions about Q4 of 2015. With tournament outcomes in the regressions, variation in manager predictions captures overconfidence or underconfidence relative to what one would predict based on past signals. Columns (2) to (4) use different binary indicators for manager overconfidence relative to various benchmark predictors: Historical mode, mutlinomial logit, and baseline structural model prediction. These regressions control for the respective predictor. The regressions also control for store characteristics, most notably the number of AM's recommended for that store, but also geographic region and store age.[45] Column (1) shows that more confident managers were in fact significantly less likely to hire the number of AM's recommended by the company, and instead tended to rely on fewer AM's. In Column (2) and (4) the point estimates for the binary indicators for overconfidence are also negative, and in the latter case, statistically significant. The coefficient on the multinomial logit indicator for overconfidence in Column (3) is essentially zero and imprecisely estimated (this specification involves the fewest observations as the multinomial logit prediction is for the sub-sample of experienced managers with 8 or more lags of past performance).

Second, the analysis explores how overconfident managers approached the decision of whether or not to delegate decisions to workers, in a lab experiment conducted as part of the lab in the field study. The experiment was about task choice. The manager was randomly and anonymously matched with a real worker from the firm. Both the manager and the worker had one minute to look at two brain teaser questions. The

[45]In this case the regressions include controls for four larger regions, rather than 38 smaller geographic areas. This reflects lack of variation of the dependent variable within a substantial number of the smaller areas, which requires dropping the respective areas and leads to insufficient numbers of observations to estimate some regressions. Results are similar for those regressions that do have sufficient observations for estimation.

questions were equally difficult empirically, although participants were not informed about this. Then, the manager could decide whether to let the worker pick which problem to solve, or decide for the worker which problem to solve. To break indifference, the manager had to pay a small cost, roughly 7 cents, if they wanted to choose the problem to be solved by the worker. The payoff of the manager and the worker depended on whether the worker got the chosen problem right; a correct answer would give both the worker and the manager roughly $12 (subtracting 7 cents from the manager in case they chose the problem for the worker). The analysis tests the hypothesis that more overconfident managers were more likely to be confident in their own ability to select the best problem, as opposed to the worker's ability to select the best problem.

Table T2 shows results of probit regressions where the dependent variable is equal to 1 if the manager chose the task for the worker in the lab experiment, and 0 if the manager let the worker choose. Since the dependent variable is choice in a laboratory experiment, it is less clear that the regressions need to include controls for store characteristics, but these are included for consistency with the analysis on AM's and future performance; results are similar without store controls.[46] Column (1) shows that more confident managers were significantly more likely to control the worker task choice. The point estimates for the binary indicators of overconfidence in Column (2) through (4) are generally less precisely estimated, but they are consistently positive and substantial and size, and statistically significant in the case of the historical mode specification.

[46]The regressions control for the four larger geographic regions, but results are similar controlling for the 38 smaller areas.

Table T1: Overconfidence and hiring the recommended number of assistant managers

	Hires recommended number of AM's			
	(1)	(2)	(3)	(4)
Manager prediction about Q4 of 2015	-0.07**			
	(0.03)			
Overconf. rel. to mode		-0.06		
		(0.07)		
Overconf. rel. to mult. Logit			0.01	
			(0.07)	
Overconf. rel. to structural				-0.14**
				(0.06)
Performance percentile in Q2 of 2015	0.04	0.03	-0.10	0.01
	(0.03)	(0.03)	(0.08)	(0.03)
Performance percentile in Q3 of 2015	0.00	0.00	-0.03	-0.02
	(0.03)	(0.04)	(0.04)	(0.03)
Performance percentile in Q4 of 2015	0.04	-0.00	0.05	0.03
	(0.03)	(0.04)	(0.05)	(0.03)
Female	0.02	-0.00	0.01	-0.01
	(0.05)	(0.05)	(0.07)	(0.05)
Age	0.06	0.04	0.09	0.04
	(0.04)	(0.04)	(0.07)	(0.03)
Experience	-0.03	-0.01	0.03	-0.01
	(0.03)	(0.03)	(0.05)	(0.03)
Mean performance percentile pre-Q2 of 2015	0.01			
	(0.02)			
Historical modal quintile		-0.01		
		(0.02)		
Mult. Logit predicted quintile			0.10	
			(0.07)	
Structural predicted quintile				-0.01
				(0.03)
Additional store controls	Yes	Yes	Yes	Yes
Observations	148	127	74	164
Pseudo R^2	0.179	0.133	0.308	0.157

Notes: The table reports marginal effects from Probit regressions. The dependent variable is a dummy variable equal to 1 if the manager hired at least as many assistant managers as the firm recommended for the manager's particular store in Q4 of 2015, and 0 if the manager hired fewer than the recommended number of managers. Independent variables are standardized so coefficients show the change in the probability of hiring the recommended number of AM's associated with a 1 s.d. change in the independent variable. The sample is restricted to managers who worked in Q4 of 2015, so there is one observation per manager. Column (1) includes the manager's prediction for Q4 of 2015 quintile, and controls for past manager performance in Q3 and Q2, and the mean of pre-Q2 performance. Columns (2) to (4) use binary indicators for manager overconfidence about Q4 of 2015, relative to different benchmark predictors: Historical mode, multinomial logit, and baseline structural model. These models also control for the respective predictor. Additional store controls include dummy variables for one of four geographic regions, as well as age of the store, and also the number of assistant managers recommended for the manager's store by the firm. Robust standard errors are in parentheses.

Table T2: Overconfidence and controlling worker task choices

	Chooses task for the worker			
	(1)	(2)	(3)	(4)
Manager prediction about Q4 of 2015	0.10**			
	(0.05)			
Overconf. rel. to mode		0.25***		
		(0.09)		
Overconf. rel. to mult. Logit			0.07	
			(0.12)	
Overconf. rel. to structural				0.10
				(0.09)
Performance percentile in Q2 of 2015	-0.02	-0.00	-0.25***	-0.02
	(0.05)	(0.05)	(0.09)	(0.05)
Performance percentile in Q3 of 2015	-0.01	-0.05	0.15**	-0.01
	(0.05)	(0.05)	(0.06)	(0.05)
Performance percentile in Q4 of 2015	-0.00	0.04	-0.06	0.04
	(0.05)	(0.05)	(0.07)	(0.05)
Female	-0.06	-0.13	-0.05	-0.07
	(0.08)	(0.08)	(0.10)	(0.08)
Age	0.08	0.04	0.13**	0.07
	(0.05)	(0.05)	(0.05)	(0.05)
Experience	-0.04	-0.04	-0.07	-0.04
	(0.05)	(0.05)	(0.06)	(0.05)
Mean performance percentile pre-Q2 of 2015	0.05			
	(0.04)			
Historical modal quintile		0.08*		
		(0.04)		
Mult. Logit predicted quintile			0.19***	
			(0.06)	
Structural predicted quintile				0.04
				(0.04)
Additional store controls	Yes	Yes	Yes	Yes
Observations	148	127	74	164
Pseudo R^2	0.088	0.085	0.170	0.051

Notes: The table reports marginal effects from probit regressions. The dependent variable is a dummy variable equal to 1 if the manager chose to decide for the worker, which problem to solve in the lab experiment, and 0 otherwise. Independent variables are standardized so coefficients show the change in the probability of controlling worker task choice associated with a 1 s.d. change in the independent variable. The sample is restricted to managers who worked in Q4 of 2015, so there is one observation per manager. Column (1) includes the manager's prediction for Q4 of 2015 quintile, and controls for past manager performance in Q3 and Q2, and the mean of pre-Q2 performance. Columns (2) to (4) use binary indicators for manager overconfidence about Q4 of 2015, relative to different benchmark predictors: Historical mode, multinomial logit, and baseline structural model. These models also control for the respective predictor. Additional store controls include dummy variables for one of four geographic regions, as well as age of the store. Robust standard errors are in parentheses.

U Instructions for the lab-in-the-field study

In this section we provide more details on the other full set of measures used in the study. The instruction wording is paraphrased or adjusted in some places, to avoid firm-specific terminology, but without changing the sense of the instructions.

The key measures for our analysis are the prediction measure and the measure of memory about past rank. The measure of manager predictions about tournament rank in Q4 of 2015 can be found in **Part 9** of the instructions. The measure of manager memories about rank in Q2 of 2015 can be found in **Part 10**, the first part of the first question.

Academic Study on Managers

Introduction

This is academic research, by economists at Oxford and Cambridge Universities who are collaborating with company X <actual name here>.

We will be sharing the general results with company X, but will keep your individual decisions ***completely confidential.***

The study involves questions, tasks and games and to make it more fun, **you can earn money!**

It will take about 1 hour and 15 minutes.

The money

- You will automatically get **$23** as a "thank you" for participating, plus extra money you make based on your choices in the study.
- The money does not come from company X, but from an academic grant.
- The study has 11 parts. We will randomly select 1 of the first 10 parts to actually be paid. Since any part (except the last one) could be selected and determine how much you earn, you should choose carefully in every part. At the end of the session, we will publicly roll a 10-sided die to randomly select one part, so you will learn today which part will be paid.
- We will send you a check in Q1 2016, after all managers have completed the study. It will include the thank you payment and any extra earnings you have. The check will go to the address you indicate on the next page.
- Along with the check, we will tell you the outcomes of the games, how well you did in each (in money terms), and how your payments were calculated.
- Please fill in your name and store number:

Name ________________________________

Store Number ________________________________

- In a minute we will ask you to write your name, address and signature in the University of Cambridge form on the next page. Leave the rest of the form blank. This form will enable us to pay you from our grant.

Rules

- Wait to turn the page until we say it's ok—for every part of the study.
- No talking during the study.
- No mobile phones or calculators.
- After you are done, don't talk about the study to managers who haven't taken part yet.

Please wait to turn the page until we say it's OK

Part 1

How does it work?

1. We give you \$15.
2. You must decide how much of this amount you wish to bet in a lottery.
3. You can bet anything from zero to \$15.

How do you earn money?

Please remember that only one part, chosen at random, will contribute to your earnings.

If you bet more than zero, we will flip a coin:
If it lands:

heads we give you **two and a half times** whatever you bet.

tails you **lose half** of whatever you bet.

If you bet nothing, you get to keep the \$15.

Example

Suppose that out of the \$15, you decide to bet \$6, keeping \$9.
We flip the coin.

- If it lands heads then we give you two and a half times \$6 which is \$15.
  You would now have \$9 + \$15 = \$24.
  This is obviously better than \$15, so you won.
- If it lands tails then you lose half of the \$6, and get back \$3.
  This would leave you with \$9 + \$3 = \$12.
  This is obviously worse than \$15, so you lost.

Make your decision

How much of the \$15 will you bet ?

I will bet \$ ____________________

Please wait to turn the page until we say it's OK

Part 2

How does it work?

1. The envelope with a ***green*** page inside has math problems —**we will tell you when to open it.**
2. Solve as many math problems correctly as you can **in 3 minutes.**
3. For the problems you just add up 5 two-digit numbers.
4. You figure out the sum and write it in the blank box.
5. You can use scrap paper during the three minutes, but **no calculators.**

! When we tell you that the **3 minutes** are over, you must put down your pen and stop working.

Example

$$21 + 35 + 48 + 29 + 83 = \underline{\hspace{3cm}}$$

How do you earn money?

Please remember that only one part, chosen at random, will contribute to your earnings.

You will get a "piece rate": **$2 per correct answer.**

Please wait to open the envelope with the *green* sheet inside until we say it's OK

Part 3

How does it work?

1. The envelope with a ***lilac*** page inside has more math problems—**we will tell you when to open it.**
2. You will solve maths problems again for **3 minutes**, like you did for **Part 2.**

How do you earn money?

Please remember that only one part, chosen at random, will contribute to your earnings.

After all managers have done the study, we will randomly match you with *another* manager. It will be a "tournament" with the other manager.
We will check:

1. How many answers **you** got right.
2. How many answer **your matched manager** got right.

- If you get more answers right, you win and get $5 per correct answer.
- If you get less answers right, you lose and get $0.
- A coin flip determines the winner if there is a tie.

Example

- You do 5 correct answers.
- Your matched manager does 4 correct answers.
- You win! You get $5 for each answer
- You earn 5 x $5 = $25; your matched manager gets zero.

! *The randomly chosen manager can be from any store in the country, so it is almost certainly **not someone in the room today**. You and the other manager will never find out each other's identity.*

Please wait to open the envelope with the *lilac* sheet inside until we say it's OK

Part 4

How does it work ?

1. The envelope with a ***yellow*** page inside has one more set of math problems—**we will tell you when to open it.**
2. You will solve maths problems one more time for **3 minutes**, like you did for **Part 2 and Part 3.**
3. This time **you get to choose** how you will earn money.

How do you earn money?

Please remember that only one part, chosen at random, will contribute to your earnings.

Option 1:
Get **$2** per correct answer, like in Part 2.

Option 2:
Get matched with another randomly chosen manager:

- We compare whatever **new** score you get in this part, Part 4, to the score this matched manager had back in Part 3, in the tournament.
- If your **new** score is higher than the other manager's score in Part 3, **you get $5 per correct answer**.
- If you have fewer right answers, **you get $0.**
- A coin flip determines your payment if there is a tie.

Make your decision

☐ **Option 1:** **$2** per correct answer.

☐ **Option 2:** **$5** per correct answer if you win, **$0** otherwise.

Please wait to open the envelope with the *yellow* sheet inside until we say it's OK

Part 5

How does it work ?

Think back to Part 3, where you were **required** to be in a tournament with a manager.

Give your best guess, about how your score in **Part 3** ranked compared to all managers who did Part 3.

We expect roughly 300 managers to participate in this study.

! You do not have to guess the exact rank, but just which range you think is most likely to contain your rank.

How do you earn money?

Please remember that only one part, chosen at random, will contribute to your earnings.

If your actual rank falls in the range that you guessed, you will get $23.

Make your decision

Consider where you may rank when put in a league with all managers, in terms of your Part 3 score, and tick one of the options below to show your most likely position in the league.

Top 20%	Roughly, ranks 1 to 60	☐
Top Middle 20%	Roughly, ranks 61 to 120	☐
Middle 20%	Roughly, ranks 121 to 180	☐
Bottom Middle 20%	Roughly, ranks 181 to 240	☐
Bottom 20%	Roughly, ranks 241 to 300	☐

Please wait to turn the page until we say it's OK

The next three parts, part 6, 7 and 8, have tasks you do with a store worker. We will randomly match you with a store worker, and it is most unlikely to be a store worker you know. You will never learn each other's identity. The store worker will complete their part of the task in a separate session for store workers. Each store worker will be performing several tasks, and will be paid for one of them we chose at random.

Part 6

How does it work ?

1. The store worker with whom you are matched will have to try to solve one of 2 problems for you.

2. Before they do this, you get to look at the 2 potential problems for **1 minute**. You will then get to choose one of the following:

You can **let the store worker decide** which problem to try to solve	**or**	You **decide for the store worker** which one they have to solve. For this you have to pay a **10¢** "fee"

The store worker

The store worker goes through the following steps:

1. Looks at the 2 problems for 1 minute, like you did.
2. Then, if you picked a problem for them, they learn this, and work on the one you picked for 1 minute.
3. Or, if you let them choose, they choose one, and work on that one for 1 minute.

How do you earn money?

Please remember that only one part, chosen at random, will contribute to your earnings.

1. If the store worker gets the ***right*** answer:

   - The store worker gets $15.

•	You also get \$15, if the store worker chose the problem	**or**	You get \$14.9, if you chose the problem because you have to pay the 10¢ fee

2. If the store worker gets it ***wrong***: you both get \$0.

Please wait to turn the page until we say it's OK

Study the problems

You have 1 minute to look at these two problems.

Problem 1

If the length of a rectangle is increased by 25% and the width is decreased by 25%, what is the percentage change in its area from its original amount ?

1. *-10%*
2. *-6.25%*
3. *0%*
4. *6.25%*
5. *10%*

Problem 2

A man keeps his boat in a lake. A ladder is attached to the boat, with three rungs showing. The rungs are 30cm apart. After a drought, the water level in the lake sinks 3m. How many rungs of the ladder are showing now?

1. *3*
2. *8*
3. *14*
4. *20*
5. *25*

Make your decision

Now tick your choice, and indicate the problem the store worker should solve if you are deciding for the store worker.

☐ **I want the store worker to solve problem number _____ (1 or 2).**

☐ **I want the store worker to decide which one to solve.**

Please wait to turn the page until we say it's OK

Part 7

How does it work ?

1. You will be matched with a chosen **store worker** at random, like last time.
2. The store worker works on one problem for you.
3. This problem will be adding 4 two-digit numbers in 20 seconds.

How do you earn money?

Please remember that only one part, chosen at random, will contribute to your earnings.

- If the store worker gets it right, **you get $12 and the store worker gets $12.**
- If the store worker gets it wrong, **you get $8 and the store worker gets $8.**
- In addition, as the manager, you also have an **"extra budget" of $15.**
- You are free to keep the extra $15 as additional earnings, or you can give some or all to the store worker. You can decide for two cases:
  1. In case the store worker gets it right.
  2. In case the store worker gets it wrong.

After the store worker has participated, we will do what you said, for the outcome that actually happens.

! *When the store worker is working on the problem the store worker knows that you will decide about the extra budget, in the cases that the store worker gets it wrong or right, but they don't find out your exact decision until afterwards.*

Make your decision

Now choose what you want to do in each case.

If the store worker gets it **right** I want to give $ __________ (0 to 15) of the extra budget to the store worker, keeping the rest of the $15.

If the store worker gets it **wrong** I want to give $ __________ (0 to 15) of the extra budget to the store worker, keeping the rest of the $15.

Please wait to turn the page until we say it's OK

Part 8

How does it work ?

1. You will be matched with a chosen **store worker** at random, like last time.
2. The store worker gets a fixed amount of money for this part: **$15.**
3. The store worker can choose how many problems to do for you, from a list of 10 problems.
4. Each problem is adding 7 two-digit numbers.
5. There is **no time limit** for the store worker to do the problems.

How do you earn money?

Please remember that only one part, chosen at random, will contribute to your earnings.

Each problem the store worker attempts and gets right, **you get $3**.

Example

The store worker does 0, you get **$0**;

The store worker does 10 right, you get **$30**.

Your decision

Before the store worker is given the list of problems, **you decide** the following:

Require store worker to do a minimum of 2 problems, correctly, for them to get their $15.	**or**	Do not require a minimum number for the store worker to do.

The store worker

1. The store worker finds out whether or not you required them to do at least two problems, correctly, to get their $15.
2. The store worker decides how many to attempt to solve; can stop at any time

Make your decision

Tick your decision below

☐ **Require the store worker to do a minimum of 2 problems, correctly.**

☐ **Do not require a minimum number.**

Please wait to turn the page until we say it's OK

The next two parts 9 and 10 are different, because we will ask you about the **company X bonus tournament.**

Part 9

How does it work ?

Think about the company bonus tournament **in this quarter, Q4 of this year**.

We will ask for your best guess about your store's overall position (rank) in the company bonus tournament for Q4.

Company X expects roughly 300 stores to take part in the Q4 bonus tournament.

! You do not have to guess your exact position, just a range!

How do you earn money?

Please remember that only one part, chosen at random, will contribute to your earnings.

We will get information from the bonus tournament, and will pay you $23 if your actual position falls within the range you guessed.

Make your decision

Now, mark the range that you think is most likely for your overall position (rank).

Top 20%	Roughly, ranks 1 to 60	☐
Top Middle 20%	Roughly, ranks 61 to 120	☐
Middle 20%	Roughly, ranks 121 to 180	☐
Bottom Middle 20%	Roughly, ranks 181 to 240	☐
Bottom 20%	Roughly, ranks 241 to 300	☐

Please wait to turn the page until we say it's OK

Part 10

How does it work ?

We ask you eight questions about the **company bonus tournament that has already happened** in Q2 of this year.

How do you earn money?

Please remember that only one part, chosen at random, will contribute to your earnings.

You get $3 if you get all the parts of the question correct. If not we pay a proportion for the parts you got right.

A hint

The company bonus tournament results table looked like this in Q2 (but longer):

<Tournament table column titles and one row as an example here>

Please wait to turn the page until we say it's OK

In the Q2 company bonus tournament:

1. What was (a) your store's **rank** (position); (b) your **Final Bonus?** (We count your answers as right if they are within plus/minus 10 of the actual).

   <Tournament table column titles here with the required titles circled>

Rank	Final Bonus
	%

2. What **Final Bonus** did the top and bottom stores in the company have? (We count your answers as right if they are within plus/minus 10 of the actual).

Top	Bottom
%	%

3. What was your area bonus? (We count your answer as right if it is within plus/minus 2 of the actual).

   <Tournament table column titles here with the required titles circled>

Area bonus
%

4. What **score** did your store achieve in each of the four measures?

   <Tournament table column titles with the required titles circled. Below RME is Regional Manager Evaluation>

Sales Growth	Profit	Service	RME
%	%	%	%

5. For your store, what was **the score in the next band up, above the one you actually got,** in each of the measures? (if you were in the top band, write N/A and we count this as correct answer)

Sales Growth	Profit	Service	RME
%	%	%	%

6. For your store, what was **he score in the next band down, below the one you actually got,** in each of the measures? (if you were in the bottom band, write N/A and we count this as correct answer)

Sales Growth	Profit	Service	RME
%	%	%	%

7. Choose any one of the measures of bonus tournament. What were the top and the bottom scores **possible** in this measure?

Measure you choose	Top	Bottom
	%	%

8. Consider two imaginary company X stores:

   Store A falls into a a band with score Z <actual number here> in each of all four measures.

   Store B falls into a **high scoring band** on some measures, and a **low scoring band** in others, but the **average** of these scores is Z <actual number here>.

   Everything else that's relevant for the overall Final Bonus in the bonus tournament is the same for these two stores.

   Tick below which store has a higher Final Bonus:

   ☐ **Store A.**

   ☐ **Store B.**

   ☐ **Both will have the same overall Final Bonus.**

9. Lucky question! You can get extra money for Part 10, depending on the die in the cup. Shake the cup, holding your hand over the top so that it doesn't fall out. Now look to see what you rolled:

   If you roll a 6, you get $6.

   If you roll a 5, you get $5.

   If you roll a 4, you get $3.

   If you roll a 3, you get $2.

   If you roll a 2, you get $0.

   If you roll a 1, you get $0.

   Record the number you rolled here so we can pay you:

Please wait to turn the page until we say it's OK

Part 11

Please answer the following questions:

1. What is the year you were born?

2. Are you male or female?

☐ **Male**

☐ **Female**

3. How many years of managerial experience do you have (including before company X)?

4. Please circle a number below to indicate:
"Are you a person who is generally fully prepared to take risks, or do you try to avoid risks?"

Completely **unwilling** to take risks — Completely **willing** to take risks

0 1 2 3 4 5 6 7 8 9 10

5. Please circle a number below to indicate:
"Are you generally a person who is fully prepared to compete, or do you prefer to avoid competition?"

Completely **unwilling** to compete — Completely **willing** to compete

0 1 2 3 4 5 6 7 8 9 10

6. Please circle a number below to indicate:
"In general, are you a person who is confident that you can do better than others, or are you not that confident?"

Not at all confident — **Very** confident

0 1 2 3 4 5 6 7 8 9 10

7. Please circle a number below to indicate:
"How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?"

Completely **unwilling** to give up										Completely **willing** to give up
0	1	2	3	4	5	6	7	8	9	10

The study is complete, thank you!

Please wait for us to give a wrap up and collect materials.