Can information alleviate overconfidence? A randomized experiment on financial market predictions*

Takanori Ida† and Ryo Okui‡

October 7, 2019

Abstract

In this study, we examine how information provision affects the degree of overconfidence using an online experiment. The 4,210 experimental participants engaged in stock market prediction exercises in 2014 were asked to evaluate their absolute and relative performance. We conducted a randomized controlled trial such that randomly selected participants obtained information about their own performance and/or the distribution of others' performances before evaluating their performances. We find that while participants exhibit overconfidence bias, this can be alleviated by information provision and that the effect of the elimination of overconfidence is stronger when only partial information, rather than complete information, is provided. Further, we show that the mere provision of information, even if it is consistent with prior beliefs, decreases the degree of overconfidence.

Keywords: Overconfidence, information provision, randomized controlled trial, online experiment, stock market prediction.

JEL classification codes: C91, D83, D91, G41.

1 Introduction

Overconfidence is a human characteristic.[1] Indeed, we all tend to overestimate our own abilities and achievements. There is a large body of research on the nature of the structural bias in our decision-making as citizens, consumers, investors, and entrepreneurs attributable to overconfidence as well as how the anomalies derived from overconfidence bias affect the allocation of market

*We thank Yasuhiro Oku, Jiaying Yang (Joey), and Yue Yu for their excellent research assistance. We acknowledge Ayako Tanaka and Yuki Ishida at MyVoice for helping us run the web-based experiments. Ken Yamada, Eric Set, and the seminar participants at NYU Shanghai provided valuable comments. We would like to thank Editage (www.editage.jp) for English language editing. Financial support from the New Faculty Startup Grant of Seoul National University and the Housing and Commercial Bank Economic Research Fund for the Institute of Economic Research of Seoul National University are greatly acknowledged. A part of the research was conducted while Okui was at Kyoto University and NYU Shanghai.

†Graduate School of Economics, Kyoto University, Yoshida, Sakyo, Kyoto 606-8501, Japan. Email: ida@econ.kyoto-u.ac.jp

‡Corresponding author: Department of Economics and the Institute of Economic Research, Seoul National University Building 16, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, South Korea; and Department of Economics, University of Gothenburg, P.O. Box 640, SE-405 30 Gothenburg, Sweden. Email: okuiryo@snu.ac.kr

[1]Section 2 reviews extant research on overconfidence.

resources. However, it is puzzling how such overconfidence can survive in the current era of information overload, particularly as we are inundated with materials through which to assess our own abilities and achievements, as well as the abilities and achievements of others.

This study uses a randomized controlled trial (RCT) to investigate the degree to which overconfidence is affected by the provision of information on an individual's own performance and/or that of others. We first conducted an online experiment involving the prediction of share prices in October 2014, with 4,210 participants. In November 2014, we conducted a second online experiment with 3,000 of the participants from the first online experiment to examine the lasting effects of the interventions in the first experiment. The demographic characteristics of participants varied in terms of their region, gender, and age and corresponded to those of ordinary citizens. They were asked to predict the share prices of companies in various industries. Specifically, they were shown the time series of the share prices and asked to make three-month-ahead predictions. There were 20 questions based on the time series data of actual share prices. We then asked participants to evaluate their own absolute and relative performance. We intervened by randomly providing them with information on their own performance and/or the distribution of other people's performances.

In particular, we asked participants about both their absolute performance (i.e., their proportion of correct predictions) and relative performance compared with others. These questions reflect two concepts within overconfidence. The first is "overestimation", in which humans estimate their own scores too highly in absolute terms. When asking about the participant's own test results in absolute terms, we intervened by either providing or withholding the distribution of other people's performances. The second concept is "overplacement", in which humans estimate the relative position of their own score too highly. This element can be investigated by asking participants to estimate the probability of their own performance being higher than that of a randomly selected opponent. As interventions, we provided information on the participant's own performance and/or the distribution of other people's performances and examined how this affected overplacement to investigate the effect of information provision on these two evaluations.

Our aim in conducting this RCT was to verify the role of cognitive bias in overconfidence, which may be caused by various forms of such bias. If cognitive bias giving rise to overconfidence simply derives from a lack of information, then correcting that cognitive bias by providing information on the test performances of participants and others could mitigate overconfidence. However, if cognitive bias originates not only from a lack of information, but also from more complex psychological processes, then simply providing information would not necessarily correct overconfidence.

In particular, we consider the effects of "limited attention" and "confirmation bias." Limited attention, also called "information friction" and "cognitive cost" has recently garnered attention for explaining the complex relationship between information and cognitive bias (Handel and Kolstad 2015; Chetty et al. 2009; Shah et al. 2012). It refers to the cognitive cost of focusing on the information provided. When limited attention exists, cognitive bias can be corrected by providing information on an option or comparing one option with another; however, the correction remains incomplete (Bertrand, Karlan, et al. 2010; Bertrand and Morse 2011; Stango and Zinman 2015).[2] On the contrary, confirmation bias induces a more complex form of information processing. This refers to the tendency to interpret subsequent evidence to maintain that initial belief and can both

[2] Chetty et al. (2009) use the concept of limited attention to explain that consumer demand changes depending on whether sales tax is included in the price shown or shown separately.

confirm evidence as relevant and reliable as well as disconfirm evidence as irrelevant and unreliable (Lord et al. 1979). Individuals with confirmation bias may strengthen their overconfidence (or underconfidence) after receiving information depending on the nature of the information, as indicated by the formal model developed by Rabin and Schrag (1999).

The results of our study reveal that while the provision of information on the distribution score does correct overestimation, it does so to an insufficient degree. Even after controlling for individual attributes, the correction effect is only in the region of 10–20%, with 80–90% of the overestimation remaining after the intervention. An examination of the quantile treatment effects (QTEs) indicates that the effect is homogeneous across people such that information provision decreases the degree of confidence regardless of whether one is initially overconfident or underconfident. Interestingly, participants tend to declare the information on other people's performances to be "as expected" (e.g., not higher or lower than their prior prediction).

With regard to overplacement, when information on the test performances of both themselves and others is provided, we do not find a strong effect on the level of overplacement. In this situation, it should be possible for participants to accurately calculate the probability of winning. However, participants seem unable to fully correct their overplacement. The QTE estimation results suggest that overconfident people correct their confidence downward, while underconfident people correct their confidence upward, whereas the degree of correction is insufficient. Moreover, our results suggest that the threshold for the change in correction direction lies in underplacement and that people whose prior beliefs have no underlying bias tend toward underplacement. In short, we find that while the direction of correction is appropriate, the degree of correction is insufficient.

It is surprising that if only one type of test performance information—either the participant's own performance or that of others—is provided, the correction of overplacement is greater than when full information is provided. When only information on the test performance of others is provided, overplacement is corrected by around 85%, with only 15% of the initial overplacement remaining. When only information on their own test performance is provided, participants' overplacement is corrected almost entirely. The QTE analysis indicates that everyone corrects their confidence downward when only other people's test performances are provided, with those who already display underplacement exaggerating their underplacement bias. These findings indicate that the mere provision of information decreases confidence in addition to raising the accuracy of beliefs.

Several other interesting patterns of overconfidence are also observed. As suggested by extant research (e.g., Mobius et al. 2014), when there is real competition between a participant's score and that of his or her opponent, we find that information on winning or defeat results in the probability of winning being updated in the correct direction; however, the scale of the update is only 30–50%, a more conservative figure than that implied by the Bayesian rule. The results of the second session conducted one month after the first session do not provide evidence that the treatments given in the first session have lasting effects.

Our findings may be interpreted in the framework of limited attention theory as follows. First, we find evidence of the existence of limited attention. Even if the provided information is complete, it is not possible for participants to adequately process it. Second, providing an opportunity to enhance attention may alleviate overconfidence. We find that the mere provision of information decreases the degree of confidence and that this effect is homogeneous. This may indicate that

providing information, even if it is as expected, may work as a "nudge" to enhance attention and correct overconfidence. Partial information may thus be a more effective nudge than complete information. On the contrary, it also makes underconfident people more underconfident. While this finding is consistent with confirmation bias theory, we do not find that overconfident people become more overconfident after receiving information. Hence, confirmation bias theory fails to explain our results fully. Lastly, we find little evidence that the correction of overconfidence by information provision has a lasting effect.

Another contribution of this study is solving the problems of internal and external validity. First, we are among the few studies to use an RCT in the context of overconfidence research, which solves the problem of internal validity. Given the growing recognition of the importance of RCTs, we argue that this study's illustration of the impact of information on overconfidence is an important contribution to the literature. Second, using an online test, we were able to recruit members of the public with a variety of characteristics as participants. We accordingly addressed the problem of external validity.

The remainder of the paper is as follows. Section 2 provides a literature review and discusses the relation of this study to that research. The experimental designs are provided in Section 3. The experimental results are explained in Section 4. Section 5 discusses robustness checks with detailed discussions included in the supplemental materials. We discuss the theoretical implications of the results from our experiments in Section 6. Section 7 concludes. The mathematical arguments related to our experimental design and additional empirical results are included in the appendices. The supplemental materials of this paper contain the instructions used in the experiment and detailed discussions of the robustness checks.

2 Relation to the literature

This study examines the presence and mechanisms of overconfidence in a relatively difficult task of investment decision-making and the means of correcting it. Among the abundance of research on overconfidence, relatively few researchers employ difficult tasks. Moreover, few studies consider how to correct overconfidence, and this study constitutes one of the first to use an RCT.

Several types of overconfidence exist and each type appears in a different context. Moore and Healy (2008) and Moore and Schatz (2017) summarize the different types of overconfidence, particularly the difference between overestimation (absolute overconfidence) and overplacement (relative overconfidence). Moore and Schatz (2017) find that few studies consider both absolute and relative overconfidence, as we do in this study.

Whether overconfidence or underconfidence is observed depends on the difficulty of the task as well as whether the underlying judgment is about absolute or relative performance. In overestimation, as Moore and Healy (2008) illustrate, people tend to underestimate their ability with easier tasks and overestimate their ability with harder tasks. With regard to overplacement, however, Moore and Healy (2008) and Ertac (2011) suggest that people tend to overplace their ability with easier tasks and underplace their ability with harder tasks. In other words, the directions of overestimation bias are the opposite of those of overplacement. This pattern has been observed in other studies, too. For example, analyzing the four combinations of paid/unpaid easy/hard tasks, Hoelzl and Rustichini (2005) find that overconfidence is weak for paid hard tasks. Our results also

Table 1: Previous Studies of Over-Confidence

		Overestimation (Absolute Confidence)	Overplacement (Relative Confidence)
Straightforward Tasks	Easy Level	Moore and Healy (2008) Grossman and Owens (2012) Coffman (2014) etc.	Hoelzl and Rustichini (2005) Moore and Healy (2008) Eil and Rao (2011) Ertac (2011) Mobius et al. (2014) Brandts et al. (2015) Charness et al. (2018) etc.
	Hard Level	Moore and Healy (2008) etc.	Hoelzl and Rustichini (2005) Moore and Healy (2008) Ertac (2011) etc.
Complex Tasks		Malmendier and Tate (2005) Landier and Thesmar (2009) Malmendier and Tate (2008) Deshmukh et al. (2013) Korn et al. (2014) etc.	Camerer and Lovallo (1999) Barber and Odean (2001) etc.

indicate the presence of overplacement but the evidence is not robust (see Section 5).

Table 1 divides previous studies of overconfidence according to the type of overconfidence and difficulty of the task. It shows that overplacement has received more attention than overestimation and that these studies tend to involve straightforward easy tasks. A high number of overconfidence studies use straightforward tasks including quizzes and calculations that are not directly related to economic activity (or even realistic). These studies have included tasks such as genre-based quizzes (Coffman 2014), Mensa quizzes (Grossman and Owens 2012), double-digit calculations (Brandts et al. 2015), and Raven's Advanced Progressive Matrices tests (Charness et al. 2018). In addition, some studies have classified tasks according to their degree of difficulty. For example, Moore and Healy (2008) use genre-based quizzes of differing degrees of difficulty and Hoelzl and Rustichini (2005) use vocabulary problems with differing degrees of difficulty, while Ertac (2011) uses (easy) calculations and (hard) vocabulary problems.

Our share price prediction task is economically relevant. In particular, it examines overconfidence during a relatively difficult task. For example, stock investment by general investors, such as defined contribution pension plans, is becoming important in many countries. Accordingly, one of the most important goals of monetary policy is to control people's expectations, particularly overconfidence bias, under such circumstances. There are few examples of overconfidence studies that use complex tasks, with most researchers concentrating on CEOs' overconfidence during decision-making.[3] Malmendier and Tate (2015), for instance, define "overconfidence" as the tendency of CEOs to overestimate the value that they can create. Other studies of overconfidence in CEOs include those by Malmendier and Tate (2005) and Malmendier and Tate (2008), Malmendier, Tate, and Yan (2011), and Deshmukh et al. (2013). Relatedly, Landier and Thesmar (2009) examine entrepreneurs' optimism. CEO research is not exactly comparable to our research because CEOs are expert professionals and the basis of their judgment and quality of their decision-making are likely to be superior to those of the general public. In addition, those studies are observational,

[3]Previous studies that use complex tasks include those by Camerer and Lovallo (1999) and Barber and Odean (2001) that deal with overplacement.

while ours is a randomized experiment.

Existing research documents that belief updating is conservative and depends on the nature of information. According to several studies (Ertac 2011; Eil and Rao 2011; Grossman and Owens 2012; Mobius et al. 2014; Charness et al. 2018), participants are highly conservative when updating their beliefs when a task is self-relevant, but less so when the task concerns others. Moreover, according to Grossman and Owens (2012), when a session is repeated, the learning effect is low when the task is self-relevant, but high if it concerns others. Research also shows the asymmetry between good news and bad news. According to Eil and Rao (2011) and Mobius et al. (2014), beliefs are often updated in response to favorable information and seldom in response to unfavorable information, while Ertac (2011) makes the opposite argument.[4] However, these results were not obtained using an RCT. Mobius et al. (2014) show that overplacement can be corrected through repeated comparison with others. Our contribution is to use an unconditional RCT as well as examine both overestimation and overplacement.

Some studies examine the possible causes of overconfidence. In general, the literature on information friction (Handel and Kolstad 2015; Chetty et al. 2009; Shah et al. 2012) does not typically consider overconfidence to be its consequence. Caplin (2016) reviews the literature on limited attention, while Rabin and Schrag (1999) find that confirmation bias results in overconfidence. Strictly speaking, however, the type of overconfidence considered by Rabin and Schrag (1999) should be regarded as "overprecision", which means that participants incorrectly believe their information to be highly accurate.[5] Nonetheless, confirmation bias theory provides a useful framework within which to investigate the role of information and how it affects overestimation and overplacement. Studies also argue that overconfidence can be consistent with rationality. Benoît and Dubra (2011) point out that data that appear to indicate overconfidence can be compatible with Bayesian probability updating. H. Zhang (2013) also demonstrates that if an agent is risk-averse, even overconfident beliefs can be explained by rational utility maximization.

3 Experimental design

In this section, we first provide general information about the experiment, particularly the online survey platform on which our experiment was conducted. We also outline our experimental procedures. We then explain the information provision interventions used to examine the mechanisms of overconfidence.

3.1 General information about the experiment

Our experiment relied on the Internet research and survey company, MyVoice.[6] One million Japanese adults are registered as "monitors" with MyVoice, enabling them to carry out online surveys that reflect the composition of the Japanese population in terms of region, gender, and age

[4] Korn et al. (2014) observe that healthy people are optimistic in the face of bad luck and such probability updating is significant, while depressed people show a different trend.

[5] Several studies investigate the mechanisms that result in overprecision, including that by Ortoleva and Snowberg (2015), who consider "correlation neglect" to be one such overconfidence mechanism. Enke and Zimmermann (2019) discuss the nature of the correction of "correlation neglect." The present study does not consider the role of correlation.

[6] See https://www.myvoice.co.jp/.

among other attributes. The experiment was conducted in Japanese.[7] We rewarded participants with "points" that can be used on a shopping website associated with MyVoice.[8] The exchange rate was 1 point = JPY 1. The experiment involved a number of tasks and participants received points in each case.[9]

The experiment consisted of two sessions. The first session was conducted in October 2–14, 2014. Invitations were sent to MyVoice monitors with the outline of the experiment. After screening the time taken to respond to the invitation (eliminating those either too short or too long), 4,210 valid participants were finally obtained. The second session was conducted one month after the first, in November 7–14, 2014, with a sample of 3,000 respondents recruited from the 4,210 who participated in the first session.[10] In both sessions, participation earned 500 points. Participants carried out the experiment on a computer, presumably at their home. The average time taken to complete the session was around 40 minutes.

We obtained the demographic information of participants from the MyVoice database. Participants had input their demographic information upon registration with MyVoice. We did not ask them questions about their demographics during the experiment.

3.2 Design of the first session

The first session was split into three sections. The first section involved questions to elicit the economic and psychological characteristics, including financial knowledge, of participants. In the second section, participants were asked to predict share prices. The third section consisted of several belief elicitation tasks in which participants were asked to evaluate their performances in the second section in both absolute and relative terms.

There were seven sets of economic psychology questions and 10 questions about financial investment in the first session. The economic psychology questions were related to competitiveness with others (set 1),[11] risk preference (sets 2–4),[12] and time preference (sets 5–7).[13] The 10 questions on financial investment required participants to distinguish the correct answer from a choice of four options.[14] The maximum score for the financial knowledge test was 10.

[7] The questionnaire translated into English is provided in the online supplemental materials.

[8] According to MyVoice, using points induces more effort from participants than using cash to reward them. J. Zhang and Breugelmans (2012) find that reward points promotion is more effective for consumers than price discounts.

[9] According to Azrieli et al. (2018), it is desirable to make a payment based on performance in a randomly chosen case to avoid any hedging behavior across tasks. However, it is difficult to implement such a system because of the communication difficulty with participants. After consultation with the online survey platform, we decided not to implement that scheme and instead provided an award in each case.

[10] We examine the effect of attrition in Section 5 and the supplemental materials.

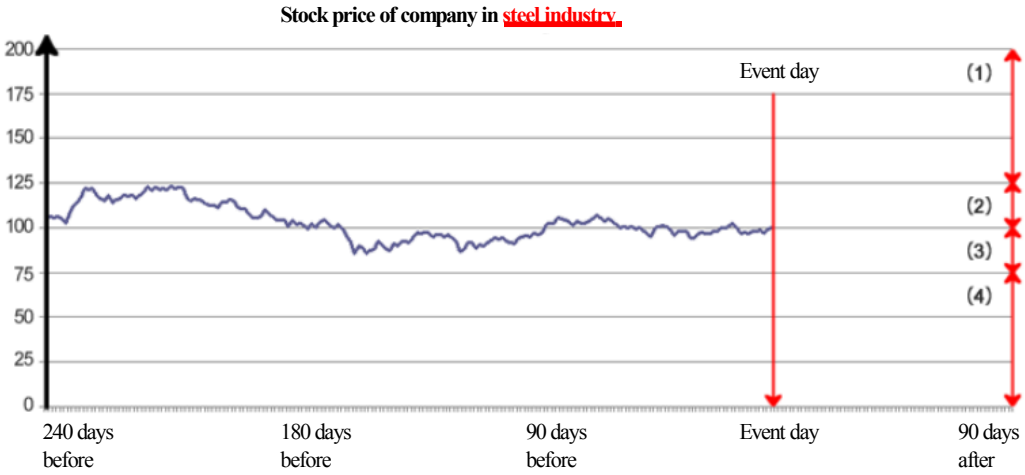
[11] Respondents were asked to say whether 10 statements such as "I am happy if I win when competing with someone else" and "I am upset if I lose to someone I am competing with" applied to themselves on a seven-point scale. These questions were based on the questionnaire used in Horino (1987).

[12] Respondents were asked to choose between Options A and B for 11 questions (e.g., Option A: "50% probability of 30,000 yen and 50% probability of no money"; Option B "100% probability of 10,000 yen").

[13] Respondents were asked to choose between Options A and B for 11 questions (e.g., Option A: "Receive 30,000 yen in one year's time"; Option B "Receive 10,000 yen immediately").

[14] We used relevant questions from the Nikkei Test (http://ntest.nikkei.jp/about/ in Japanese), which is used to examine the economic knowledge of businesspeople. An example of one questions is as follows: "Which of the following statements about bond credit ratings is incorrect?" 1) The rating shows whether the bond's principle and interest can be expected to be paid properly; 2) if a company can command a high credit rating, its cost of fund procurement will be lower; 3) if a bond's rating is below a certain level it is called a junk bond in the United States; and 4) between AAA and AA, AAA is more creditworthy.

Figure 1: Stock price prediction test



Sample question:
Based on the stock price movement shown in the chart below, predict what the stock price would be three months (90 days) after the event shown.
Select your answer from among ranges (1) to (4) shown in the chart.
The values on the vertical axis are proportional relative to an arbitrary value of 100 at the time of the event.
For example, if you think the stock price will be in range (1), select (1) as your answer.

The second section involved the share price prediction test.[15] Participants were asked to predict the share prices of 20 companies three months (90 days) ahead. We used data on actual historical share price movements. For each company, we presented the time series of the share prices for 240 days as well as the name of the industry; however, the name of the company was withheld. We therefore think that the share prediction is a difficult task for general investors. Participants were asked to select one of four alternatives: The share price 90 days after the event day was (1) 25–100% higher, (2) 0–25% higher, (3) 0–25% lower, or (4) 25–100% lower. Five points were available for each share price, with a maximum score of 100. Figure 1 provides the sample share price question used as an illustration.

In the third section, participants were asked to predict their own score in the share price prediction test and compare it with those of other participants. Beliefs were elicited in two ways: direct asking and an incentive-compatible mechanism. First, we directly asked them about their beliefs about their own score using the following format:

Bonus Challenge A: A computer will automatically select one of your 20 answers. If that answer is correct, you will win a prize.

Question: What is the probability you will win that prize?

One question was selected at random from the 20 questions and respondents received 20 bonus points if they answered it correctly. Since the question was chosen at random, the same numerical

[15] The design of this test was inspired by the task used in Offerman et al. (2009).

value as the respondent's predicted score (5 points $\times$ 20 questions = maximum score of 100) was entered. For example, if the respondent had predicted their own score as 40 points, she/he should respond that the probability of winning the prize is 40%. We then used an incentive-compatible mechanism framed in the following manner:

Bonus Challenge B: The computer will randomly select a probability and you will win a prize at that probability.

Question: What probability in Bonus Challenge B would cause you to choose it over Bonus Challenge A? Please indicate the probability at which Bonus Challenge B is more attractive.

The response was used as the elicited belief about one's own score. The prize was given according to Bonus Challenge B if the probability generated by the computer from the uniform distribution was higher than the number provided by the participant and according to Bonus Challenge A otherwise. In Bonus Challenge B, 20 points were awarded to prizewinners with some probability regardless of the correctness of the respondent's answers. As the probability of winning was chosen at random by the computer, if the probability of winning with Bonus Challenge B was greater than that of winning with Bonus Challenge A, then Bonus Challenge B was preferable. As such, the likelihood of winning that should be entered here was the rate of the correct scores that the respondent predicted for him- or herself. This mechanism was originally proposed by Grether (1981), Holt (2007), and Karni (2007), which has also been used in Mobius et al. (2014) under the name of the crossover mechanism. Appendix A.1 provides a theoretical justification of this mechanism.

Next, we elicited participants' beliefs about to which quantile of the score distribution their own score belonged (i.e., their relative performance). Each participant's own score in the share price prediction test was compared with the score of a randomly selected opponents from 100 other participants. We first directly asked them about their probability of winning:

Question: What is your predicted probability of your score being higher (a tie is not included) than the score of a randomly chosen opponent?

We then used the incentive-compatible mechanism. Here, we presented the question used in the first round.

Bonus Challenge A: You will win a prize if your score is higher (a tie is not included) than the score of opponent 1.

Bonus Challenge B: The computer will randomly select a probability and you will win a prize at that probability.

Question: What probability in Bonus Challenge B would cause you to choose it over Bonus Challenge A? Please indicate the probability at which Bonus Challenge B is more attractive.

The response was used as the elicited belief about relative performance. The prize was given according to Bonus Challenge B if the probability generated by the computer from the uniform distribution was higher than the number provided by the participant and according to Bonus Challenge A otherwise. In Bonus Challenge A, participants would not receive an award in the case of a tie. In other words, our focus is on the probability of strict winning.

Table 2: Treatment status and information structure for each group

Group	I	II	III	IV	V
Before evaluating absolute performance	O	N	N	N	N
Before evaluating relative performance	P	S	O	N	N
At the end of the session	P	S	O	P	N
# of participants in the first session	1400	550	550	500	1000

Note: P: informed of both one's own score and the distribution of others' scores (perfect information). S: informed of one's own score. O: informed of the distribution of others' scores. N: no information.

This belief elicitation was repeated four times, with a different opponent in each case. Four opponents were thus randomly chosen from the sample of 100. After each comparison with a randomly and independently chosen opponent, participants were informed whether their score was higher than that of their opponent. We also asked them about the probability of successive wins both in a direct manner and using the incentive-compatible mechanism. Specifically, before each of the first to third rounds, the probability of winning in both the next two rounds was asked. See the questionnaire in the supplemental materials for the details.

3.3 Interventions in the first session

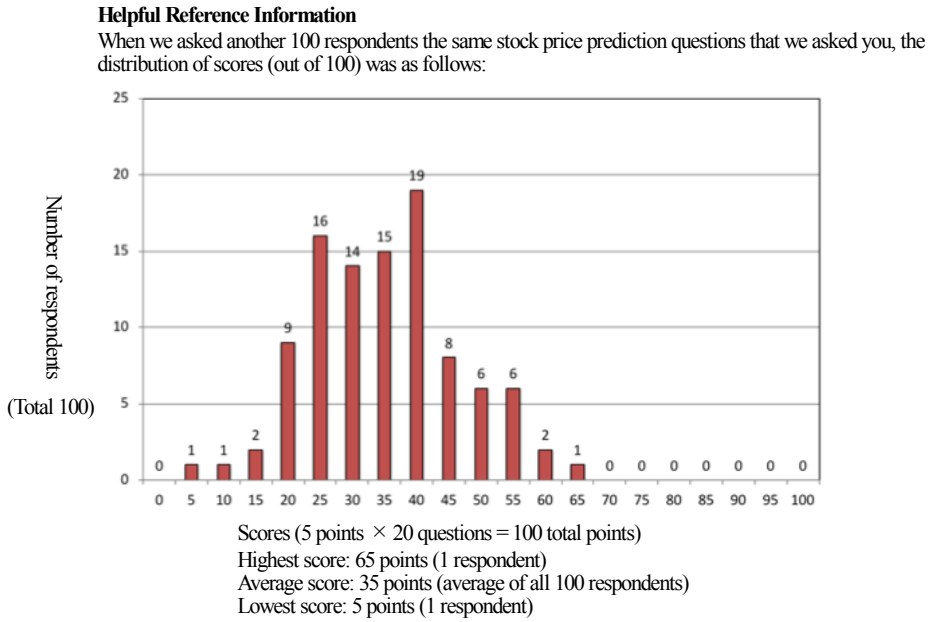
Participants were randomly assigned to one of five groups that differed in the type of information they received and the timing of receiving it. There were two types of information; their own performance and the distribution of others' performances.

Table 2 summarizes the five interventions.[16] Participants assigned to Group I obtained information about the distribution of the test performance of others (Treatment O) before they predicted their own score. Moreover, information about their own test performance was provided before they competed with a randomly chosen opponent. Group I thus obtained perfect information (Treatment P) when they were asked to evaluate their relative performances in the sense that they should be able to provide the correct probability of winning. Group II participants were informed of their own score (Treatment S) before they competed with a randomly chosen opponent. The distribution of the test performance of others was not provided. For Group III, the distribution of the test performance of others (Treatment O) was provided before they competed with a randomly chosen opponent. Information on their own test performances was not provided. For Group IV, both their own test scores and the distribution of the test performance of others were provided (Treatment P) at the end of the first session. They did not obtain any information during the session. Group V was the control group in this experiment and did not obtain any information—even at the end of the first session.

Figure 2 is the English version of the figure presented as the score distribution which is shown in Treatments O and P. It is the distribution of the scores of a sample of the 100 people who answered the same share price prediction problems in a pretest. The histogram in this figure ranges from 0–100 points in five-point increments, with the added information that the average score was 35, the highest score was 65, and the lowest score was 5. When Figure 2 was presented,

[16] In deciding the sample size for each group, seven pretests and a power calculation were conducted ahead of the real test.

Figure 2: Score distribution for the stock price prediction questions



the respondent was asked to indicate whether he or she felt it was high or low on a five-point scale.

3.4 The second session

The second session followed essentially the same design as that for the first session. The only difference was that the first section, in which we collected information on participants' economic and psychological characteristics, was not included. The other two sections were the same. In particular, participants carried out the share price prediction test again. Their scores could generally be different from those in the first section, because participants could utilize information on the scores obtained in the first section.

In the second session, participants were split into two groups: Group I and Group V. First-session Group I and Group V participants were randomly allocated to either Group I or Group V in the second session. First-session Group II participants were reallocated to Group I in the second session, while first-session Group III and Group IV participants were reallocated to Group V. Table 3 summarizes these combinations of first- and second-session interventions.

3.5 Relationship between the treatments and theoretical predictions

We next discuss how our experimental settings enable us to investigate the effect of information provision on overconfidence. First, we explain our measure of overconfidence. We then discuss how the observed effect of each intervention provides us with insights into the mechanisms of

Table 3: Number of participants in each intervention

Second session \First session	Group I	Group II	Group III	Group IV	Group V	Total
Group I	525	413			375	1313
Group V	525		412	375	375	1687
Total	1050	413	412	375	750	3000

Note: See Table 2 for the interventions each group received.

overconfidence.

We measure overconfidence by the difference between the elicited belief and actual performance. For example, the difference between the elicited belief about one's own score and actual score is our measure of overestimation. A positive (negative) difference indicates overestimation (underestimation). Likewise, our measure of overplacement is the difference between the elicited belief about the probability of beating a randomly chosen opponent and true probability. We investigate how each intervention affects these differences.

The behavior under Treatment P is useful to investigate the existence of limited attention. Under Treatment P, one can compute the true probability of winning. Thus, any difference between the reported probability and true probability indicates that a participant did not try to compute the true probability.

Treatments S and O reveal how individuals react to different types of information. Because we measure the effect of Treatment O on both overestimation and overplacement, we can disentangle its effect on the belief about one's own score and on the belief about the distribution of the scores. Treatment S resolves uncertainty about one's own score. If Treatment S has an effect different from the difference between Treatments P and O, we would identify the effect of having the information about own score on the belief about the distribution of others.

Confirmation bias tends to occur when the evidence is ambiguous and aggregated from various sources (i.e., Rabin and Schrag 1999). Thus, whether Treatment S has a larger effect than Treatment O depends on the subjective ambiguity of the information each treatment provides. More importantly, this theory also indicates that Treatment P, which combines Treatments S and O, may not have a larger effect than an individual treatment. This observation is useful for distinguishing limited attention and confirmation bias. Furthermore, confirmation bias makes individuals more responsive to information that favors them and it may exaggerate over/underconfidence.

4 Results

In this section, we present the results of the experiment. We start with the summary statistics and results of the balance tests. We then investigate the effect of the interventions. We find that providing information mitigates overestimation. Similar results are observed for overplacement. However, providing complete information has a smaller effect on overplacement than providing partial information does.

Table 4: Variable Definition

Name	Definition
Treatment P	binary variable. 1 if Treatment P is given.
Treatment S	binary variable. 1 if Treatment S is given.
Treatment O	binary variable. 1 if Treatment O is given.
Group I	binary variable. 1 if she belongs to Group I. Group II to Group V are defined analogously.
Group I-V	binary variable. 1 if she belongs to Group I in the first session and Group V in the second session. Group I-I to Group V-V are defined analogously.
Female	binary variable. 1 if female
Age	age
Competitiveness	a measure of competitiveness. a high value means that that subject is more competitive.
Finance Score	score from the quiz on financial knowledge
Schooling	categorial variable. graduate school = 1. college = 2. two years college or college of technology = 3. specialized training college, post-secondary course = 4. high school = 5. junior high school = 6.
Score	score from the share price prediction test
update12	the theoretical value of the updated belief after the first round and before the second round. update23 and update 34 are defined analogously. See Appendix A.2.
win1	binary variable. 1 if she won the first round. win2 and win3 are defined analogously.
*_2	variables with "_2" are from the second session.

4.1 Summary statistics and balance tests

Table 4 defines the variables used as the regressors and Tables 5 and 6 provide the descriptive statistics of participants' characteristics in the first and second sessions, respectively. Tables 5 and 6 also report the results of the balance tests conducted to examine if the treatments were indeed assigned independently of the characteristics of the participants.

Our sample is diverse in terms of demographic characteristics. The gender ratio is 50% and the mean of age is 47.8 years old; the youngest participant is 20 years old and the oldest is 69 years old.[17] Their educational attainment ranges from graduate school to junior high school (note that the completion of junior high school is compulsory in Japan). As schooling is a categorical variable, its mean and standard deviation are hard to interpret. Further, some participants did not report their educational attainment and others had non-standard educational attainment. We drop those participants when we include schooling in the statistical analysis, and the number of observations is smaller.

Our sample is also diverse in terms of financial knowledge and psychological characteristics. The financial knowledge score ranges from 0 (the lowest possible score) to 10 (the highest possible score).[18] We also find that the sample covers a range of degrees of competitiveness.

[17] Those younger than 19 or older than 70 were not invited to participate in this experiment.

[18] For the Nikkei Test, from which the questions about financial knowledge are extracted, the historical rate of correct answers is about 52.1%, and 90% of test takers obtain the correct rate between 25% to 80%. The distribution of scores among our participants is lower than the historical distribution, perhaps because those who take the Nikkei Test voluntarily are more motivated to improve their financial knowledge or work in the financial

Table 5: Summary statistics for the first session

Group	All	I	II	III	IV	V	Balance test
	Mean	Mean	Mean	Mean	Mean	Mean	F-stat
	(std)	(std)	(std)	(std)	(std)	(std)	(p-value)
Female	0.5	0.5	0.5	0.5	0.5	0.5	0
	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(1.000)
Age	44.55	44.676	44.555	44.34	44.525	44.501	0.068
	(13.764)	(13.801)	(13.819)	(13.696)	(13.634)	(13.807)	(0.991)
Competitiveness	35.201	35.007	35.105	35.007	35.308	35.58	0.672
	(9.371)	(9.237)	(9.467)	(8.997)	(9.655)	(9.563)	(0.611)
Score	34.094	33.759	33.698	34.284	34.406	34.519	1.037
	(11.196)	(11.272)	(11.213)	(11.021)	(11.502)	(11.015)	(0.386)
Finance Score	4.154	4.122	4.174	3.986	4.315	4.201	1.994
	(2.061)	(2.063)	(2.052)	(2.064)	(2.084)	(2.044)	(0.093)
Schooling	3.103	3.1	3.089	3.144	3.115	3.086	0.192
	(1.37)	(1.36)	(1.378)	(1.379)	(1.381)	(1.37)	(0.943)
# of obs	4210	1470	580	580	530	1050	
# of obs for schooling	4154	1452	571	570	523	1038	

It must be emphasized that participants predict future stock prices better than random guesses, even though their predictions are far from perfect. The mean score is about 34 (the mean under random guesses would be 25). The distribution of scores is similar to that obtained from the pretest whose score distribution was used in Treatments P and O. Nonetheless, some participants obtained scores above the maximum score in the pretest. Fourteen participants obtained 70 points and one participant obtained 75 points, whereas the maximum score in the pretest was 65. This result is unavoidable given that there were 4210 participants in the main experiment and only 100 in the pretest. Further, the scores in the second session are not better than those in the first session, although identical questions were used in both sessions.

Tables 5–6 also include the results of the balance tests that examine whether treatments were indeed assigned independently of participants' characteristics. Specifically, we test for the difference in the means of the variables across treatments. The results show that we seldom find statistically significant differences. We find only one significant difference at the 10% level in the finance score in the first session. In particular, although participation in the second session was voluntary, the balance tests do not show any evidence that the treatments in the second session are confounded by the attrition problem.[19] While not reported here, we examine whether the scores of opponents are independently distributed from the scores of participants in a pre-session and find no evidence against it.

4.2 Absolute performance

We first investigate how participants evaluate their own absolute performance, finding that they exhibit overestimation and that the information about the distribution of performances reduces overestimation.

Table 7 summarizes the estimation results on participants' beliefs about their own absolute

sector. See http://ntest.nikkei.jp/study/about_score.html (in Japanese).

[19]Section 5 and the supplemental materials provide a more detailed discussion about the attrition problem.

Table 6: Summary statistics for the second session

Group	All	I	V	Balance test
	Mean	Mean	Mean	F-stat
	(std)	(std)	(std)	(p-value)
Female	0.5	0.499	0.501	0.012
	(0.5)	(0.5)	(0.5)	(0.912)
Age	44.446	44.48	44.42	0.014
	(13.781)	(13.839)	(13.741)	(0.907)
Competitiveness	35.223	35.343	35.129	0.392
	(9.293)	(9.329)	(9.267)	(0.531)
Score_2	34.443	34.261	34.585	0.654
	(10.87)	(10.919)	(10.833)	(0.419)
Finance Score	4.201	4.196	4.206	0.017
	(2.052)	(2.068)	(2.039)	(0.895)
Schooling	3.101	3.076	3.121	0.769
	(1.368)	(1.384)	(1.355)	(0.381)
# of obs	3000	1313	1687	
# of obs for schooling	2960	1297	1663	

performance in the first session. The dependent variable is the difference between a subject's prediction about his or her own score and the true score. Only Group I received a treatment (Treatment O) and the other groups are included in the control group.

Column (1) shows that participants predict their own score 10.821 points (see the intercept) higher than the true score on average when no information is given.[20] It also shows that once information about the distribution of score is given, the degree of overconfidence is reduced by 1.363 (see the coefficient of Treatment O). However, a sizable degree of overconfidence remains even after the treatment, and participants who received treatment A predict 9.458 ($= 10.821 - 1.363$) points higher than the true score on average.

Next, we run regressions adding the characteristics of participants.[21] In Column (2), we control for gender because the literature reports a difference in overconfidence between men and women (e.g., Barber and Odean 2001; Brandts et al. 2015; Charness et al. 2018; Coffman 2014; Ertac 2011). We find that women are much less overconfident than men, even after we control for the other demographic variables (see Column (3)). This finding is in line with the existing literature. Further, older people are more overconfident. As expected, more competitive people are more overconfident, whereas those whose actual scores are high are less overconfident. Those with better financial knowledge are more overconfident; hence, better knowledge may not necessarily enable people to evaluate their own performance more objectively. Schooling does not have a statistically significant effect. In Column (4), we examine how the treatment effect differs across the different demographic variables. Although they are not statistically significant, the magnitudes

[20] A caveat is that the degree of overconfidence observed here might be biased. This is an online experiment and it is thus difficult to force participants to engage in the task seriously. Some might have reported some random numbers. For example, suppose that a sizable proportion of people simply put random numbers from a uniform distribution between 0 and 100. In that case, the intercept would be biased upward because the mean of this uniform distribution, 50, is above the mean of scores. We examine this problem in Section 5.

[21] It is not absolutely necessary to control for covariates to evaluate the effects of the treatment. This is an RCT and the treatments are randomly assigned. On the contrary, adding covariates might raise the accuracy of the estimates if they affect the dependent variable. Moreover, it would be interesting to see how participants' characteristics affect the degree of confidence.

Table 7: Estimation results. Prediction error about absolute performance in the first session

Dependent variable: Reported score minus true score	(1)	(2)	(3)	(4)
Treatment O	-1.363*	-1.363*	-1.881**	-3.674
	(0.676)	(0.676)	(0.602)	(2.283)
Female		-1.417*	-1.360*	-2.030**
		(0.648)	(0.598)	(0.739)
Age			-0.021	-0.021
			(0.020)	(0.020)
Competitiveness			0.100**	0.099**
			(0.033)	(0.033)
Score			-0.926***	-0.928***
			(0.026)	(0.026)
Finance Score			0.932***	1.043***
			(0.148)	(0.181)
Schooling			-0.410	-0.642*
			(0.225)	(0.276)
Treatment O × Female				1.964
				(1.257)
Treatment O × Finance Score				-0.301
				(0.293)
Treatment O × Schooling				0.665
				(0.471)
Constant	10.821***	11.530***	38.086***	38.751***
	(0.405)	(0.515)	(2.032)	(2.172)
F-test				
H_0: interaction terms involving				5.271
"Treatment O" = 0				(0.000)
$\bar{R}^2$	0.001	0.002	0.230	0.231
# of obs	4210	4210	4154	4154

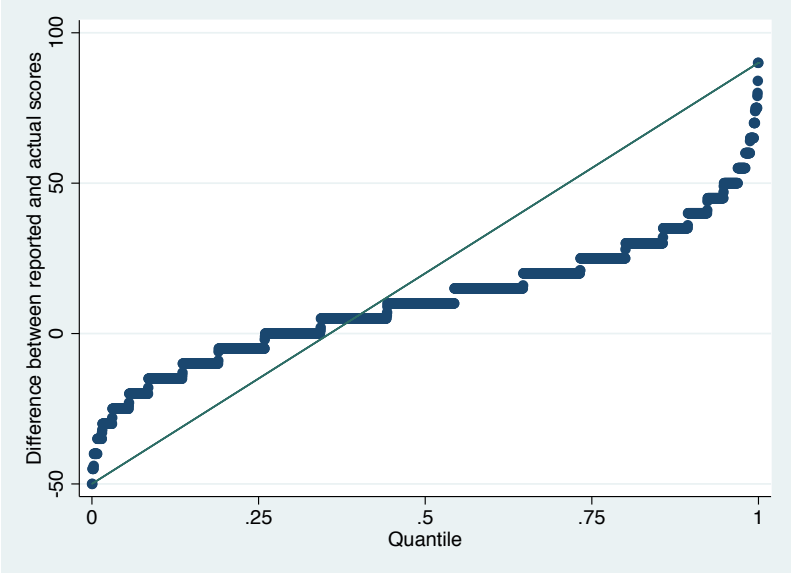
Note: Estimated by OLS. Heteroskedasticity robust standard errors are presented within parentheses below the coefficients and p-values are presented within parenthesis below the F statistic. See Table 4 for the variable definition.

of the coefficients of the interaction terms are large. For example, women are less confident when no information is provided, but this difference from men becomes small when the score distribution is presented.

We also examine the heterogeneity of the treatment effect by computing the QTE function. Before presenting the QTE, we examine the quantile curve of the difference between the reported and actual scores for the control group. The quantile function in Figure 3 shows that the distribution is diverse for both underconfident and overconfident participants. The difference ranges from -50 (the reported score is less than the actual score by 50) to 90 (the reported score is more than the actual score by 90). The difference is zero at the 35% quantile and overconfidence is present at the median. Figure 4 plots the QTE function.[22] The function is bumpy and no clear pattern is observed. Under the "rank preservation" assumption (i.e., the treatment does not change the quantile position of an individual), this result indicates that the treatment reduces the degree

[22]The QTE function is the difference between the quantile function under the treatment and that under the control. This is not the quantile of the treatment effect. See, for example, Fan and Park (2012) for more details.

Figure 3: Quantile of the difference between the reported and actual scores for the control group in the first session



Note: This figure plots the quantile curve of the difference between the reported and actual scores for the control group in the first session. The quantile curve crosses zero at the 35% quantile. The right-hand side of the 35% quantile is above zero and this corresponds to overconfidence. The left-hand side is below zero, corresponding to underconfidence.

of overconfidence of all subjects in a similar way. In other words, it makes initially underconfident subjects more underconfident, while it alleviates the bias of overconfident subjects. While the treatment alleviates overconfidence bias on average, it may not do so individually; it makes everyone including those underconfident slightly less confident.

In summary, we find evidence of overestimation, which diminishes partially once the information about the distribution of scores is given. While there appears to be substantial heterogeneity in the degree of overestimation, the treatment effect turns out to be homogeneous and does not depend on whether a subject is initially under- or overconfident. This result indicates that overestimation may not be a simple indication of a lack of information.

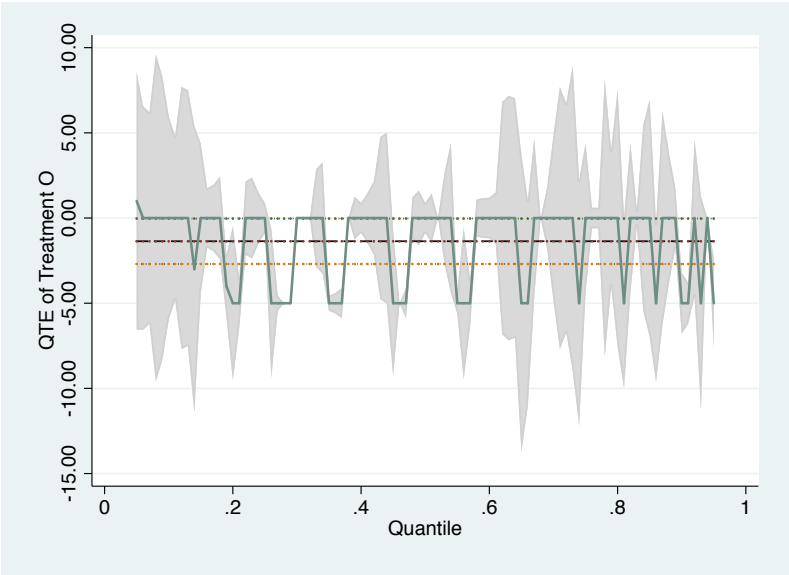
4.3 Relative performance

Next, we discuss the results for how participants evaluate their relative performances. We find that participants exhibit overplacement.[23] However, overconfidence is reduced and even disappears when they were informed of their own scores or the distribution of the scores of opponents. Interestingly, the correction effect is smaller when both their own scores and the distribution of opponents' scores were presented.

Table 8 summarizes the estimation results of the regressions in which the dependent variable

[23]However, this result may be an artifact of participants' attitudes toward the experiment as discussed in Section 5. We observe underconfidence if we select participants who appear to have engaged in the task seriously. Moore and Healy (2008) argue that the direction of confidence bias depends on whether participants are asked to evaluate their absolute or relative performance. Their results indicate that for difficult tasks, people tend to be overconfident (underconfident) about their absolute (relative) performance. As our task may be regarded as difficult, the presented results become consistent with theirs once we take the seriousness of participants into account.

Figure 4: QTE of Treatment O for absolute performance in the first session



Note: This figure plots the QTE function of Treatment O on the difference between the reported and actual scores in the first session. A 95% point-wise bootstrapped confidence band is the shaded area. The dashed line represents the average treatment effect and the dotted lines correspond to the 95% confidence interval for the average treatment effect.

is the difference between the predicted winning probability and true winning probability. Column (1) provides the estimates of the average treatment effects for the first round. Without any treatment, participants are overconfident and they report on average 5.028 points larger than the true winning probability. However, the difference decreases over rounds to 4.321 (second round), 4.271 (third round), and 4.008 (fourth round). Providing complete information thus has a small effect and is not statistically significant (see the coefficient of Treatment P). Under Treatment S, where participants observe their own scores, they become less overconfident. Indeed, the average deviation from the true winning probability becomes 0.054 ($= 5.028 - 4.974$), which indicates almost neutral confidence. Treatment O also makes participants less confident. The average deviation from the true score is 0.769 ($= 5.028 - 4.259$), which is close to the neutral confidence level. In Column (2), we find that female participants, older participants, and those with high scores are less confident. On the contrary, those with better financial knowledge and higher education are more overconfident. Columns (3)–(8) summarize the results from the analysis of the second to fourth rounds in the first session. The results from those rounds are qualitatively similar to those from the first round. Treatments S and O make participants less overconfident, whereas we do not find clear evidence that Treatment P has such an effect.

Table 9 presents the reported probabilities of successive wins. Treatment S has a relatively large and statistically significant effect and makes participants less confident. However, while Treatment O has a sizable effect, it is statistically significant only for the third and fourth rounds. The effect of Treatment P is small and not significant. The degree of overplacement, represented by the intercept, is observed to be very high. We suspect that participants may not fully understand that the probability of successive wins must be less than the probability of a single win. Indeed,

Table 8: Estimation results. Prediction error about relative performance in the first session

Dependent variable: Reported winning probability minus true probability								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Round	1st	1st	2nd	2nd	3rd	3rd	4th	4th
Treatment P	-0.560	-2.036**	0.151	-1.212	0.866	-0.466	1.315	-0.131
	(1.080)	(0.719)	(1.086)	(0.738)	(1.127)	(0.773)	(1.146)	(0.810)
Treatment S	-4.974***	-6.624***	-4.274**	-5.819***	-4.600**	-6.051***	-3.448*	-5.070***
	(1.444)	(0.916)	(1.455)	(0.956)	(1.495)	(1.005)	(1.548)	(1.064)
Treatment O	-4.259**	-4.405***	-4.133*	-4.151***	-4.267**	-4.275***	-4.610**	-4.589***
	(1.575)	(0.978)	(1.605)	(1.023)	(1.655)	(1.101)	(1.646)	(1.128)
Female		-2.090***		-1.672*		-1.655*		-1.956**
		(0.628)		(0.649)		(0.685)		(0.716)
Age		-0.069**		-0.090***		-0.096***		-0.064*
		(0.022)		(0.023)		(0.024)		(0.025)
Competitiveness		0.014		0.047		0.020		0.051
		(0.034)		(0.035)		(0.037)		(0.039)
Score		-2.071***		-2.055***		-2.082***		-2.064***
		(0.033)		(0.033)		(0.035)		(0.036)
Finance Score		0.388*		0.517**		0.298		0.100
		(0.162)		(0.166)		(0.173)		(0.183)
Schooling		-0.560*		-0.509*		-0.369		-0.485
		(0.234)		(0.240)		(0.256)		(0.268)
Constant	5.028***	80.147***	4.321***	77.712***	4.271***	80.192***	4.008***	78.250***
	(0.822)	(2.205)	(0.811)	(2.293)	(0.834)	(2.437)	(0.837)	(2.563)
$\bar{R}^2$	0.004	0.580	0.003	0.560	0.004	0.543	0.004	0.517
# of obs	4210	4154	4210	4154	4210	4154	4210	4154

Note: Estimated by OLS. Heteroskedasticity robust standard errors are presented within parentheses below the coefficients. See Table 4 for the variable definition.

the reported probability of successive wins by some subjects is higher than that of a single win.

We also investigate the heterogeneity in the treatment effects by examining the quantile function and QTEs. Figure 5 plots the quantile function of the difference between the reported and actual winning probabilities for the control group. The level of overplacement is zero at slightly below the median. As in the case of absolute performance, subjects have different degrees of overplacement, ranging from about -80 to 100. Figure 6 plots the QTE functions of Treatments P, S, and O with respect to the control. Again, we interpret these QTE functions under the rank preservation assumption. Treatment P has a positive effect slightly below the median and a negative effect above that. Because there is no overplacement at slightly below the median, Treatment P adjusts people's beliefs in the right direction. While we do not observe a statistically significant average treatment effect for Treatment P, the QTE function indicates that Treatment P affects beliefs in the expected manner. Because it has both positive and negative effects depending on the quantile, the average treatment effect is estimated to be small. Treatment S also has a decreasing QTE function; however, it crosses zero at around the 20% quantile, indicating that it makes some underconfident people more underconfident. Treatment O has a negative effect universally. It alleviates the bias of overconfident subjects but exaggerates the bias of underconfident subjects.

We also investigate how participants update their beliefs about their winning probability over rounds. Table 10 summarizes the results. For each round, we focus on the observations whose

Table 9: Estimation results. Prediction error about successive wins in the first session

Dependent variable: Reported successively winning probability minus true probability			
	(1)	(2)	(3)
Rounds	1st and 2nd	2nd and 3rd	3rd and 4th
Treatment P	0.985	1.923	1.932
	(1.090)	(1.104)	(1.136)
Treatment S	-3.440*	-2.952*	-3.348*
	(1.450)	(1.464)	(1.509)
Treatment O	-2.431	-3.081	-3.290*
	(1.569)	(1.582)	(1.619)
Constant	17.779***	17.281***	17.508***
	(0.844)	(0.842)	(0.856)
R^2	0.002	0.003	0.004
# of obs	4210	4210	4210

Note: Estimated by OLS. Heteroskedasticity robust standard errors are presented within parentheses below the coefficients. See Table 4 for the variable definition.

posterior means under Bayesian updating are within (0,100). The coefficient of, for example, update12 is one under the Bayesian update. We find that in all the rounds, the coefficients are less than one, while they are also different from zero. The updates are directionally consistent but conservative. We also examine if participants update their beliefs differently after winning and after losing. While the coefficients of the interaction terms are not statistically significant, those of the terms involving the win dummies are jointly significant, indicating asymmetric belief updating. Mobius et al. (2014) report that people react more to good news than to bad news; yet, we do not find clear evidence for this. Although the effect of winning is observed from the third and fourth rounds, the coefficient of the interaction term is negative, which indicates that participants react more to bad news in contrast to the result of Mobius et al. (2014). Barron (2017) argues that such asymmetry is not observed in the domain of financial decision-making.

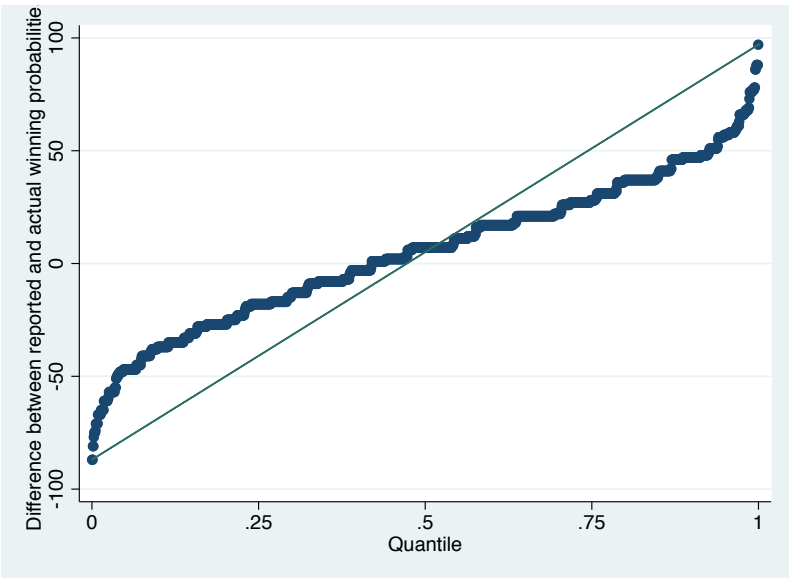
4.4 The second session

The results of the second session are similar to those of the first session; Appendices B and C present the results. In particular, Table B.2 summarizes the results on absolute performance for the second session, Table B.3 summarizes the results on relative performance for the second session; Figures C.1–C.4 plot the quantile functions and QTEs functions; Table B.4 summarizes the results for successive wins; and Table B.5 summarizes the results about belief updating. Table B.5, where we examine the updating pattern, may show a different result. Here, the coefficients of the winning dummies are large but the standard errors are also large, preventing us from drawing a strong conclusion.

4.5 Effects of the first session in the second session

Lastly, we investigate the treatment effects in the first session on the behavior in the second session to examine whether information has any long-term impact on participants' behavior. The results indicate that the effect may not last long.

Figure 5: Quantile of the difference between the reported and actual winning probabilities for the control group in the first session

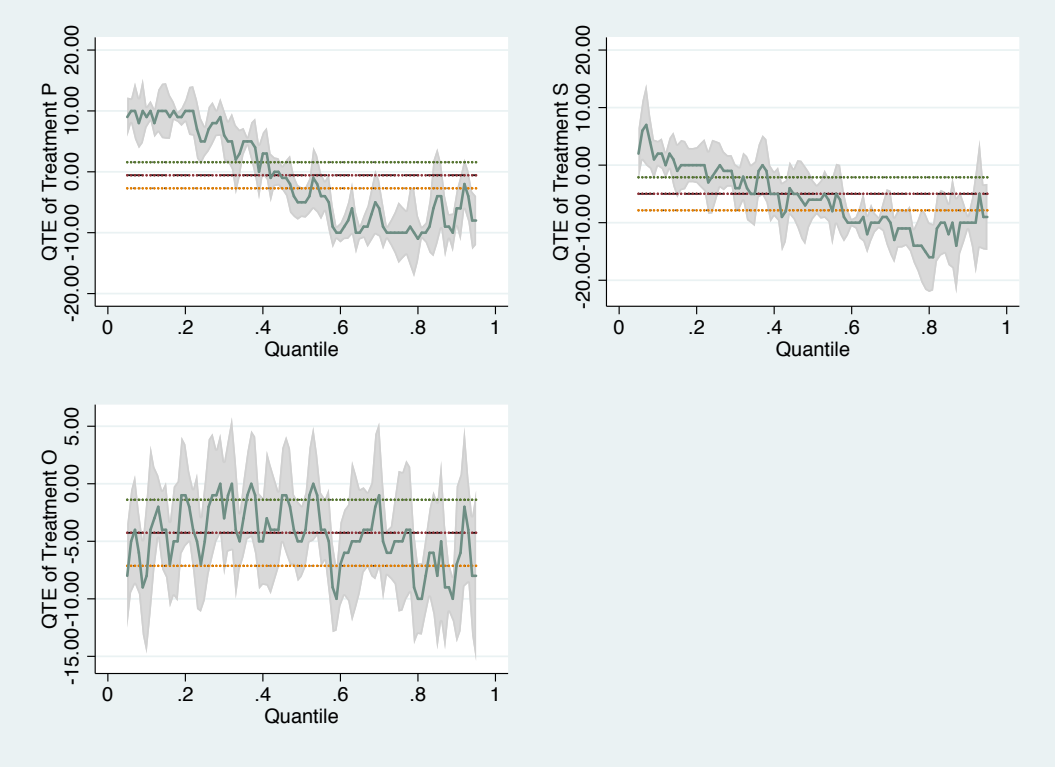


Note: This figure plots the quantile curve of the difference between the reported and actual winning probabilities for the control group in the first session. The quantile curve crosses zero at the 42% quantile. The right-hand side of the 42% quantile is above zero and this corresponds to overconfidence. The left-hand side is below zero, corresponding to underconfidence.

Table 11 summarizes the results for absolute performance. The information provided in the first session appears to have some effect but it is not large. If participants received complete information and carried out the tasks with that information in the first session, they were less confident in the second session even when no information was provided (see the coefficients of Group I-V). The effect is smaller than that of Group I-I, and statistical significance disappears once we control for the score in the second session. If participants received complete information at the end of the first session and did not engage in the task with any information (see the coefficients of Group IV-V), we find a similar effect size to that of Group I-V but this is not statistically significant. Therefore, we fail to reject the null hypothesis that the effects of Treatment P in the second session are equal irrespective of the treatments in the first session. Similarly, we fail to reject the null hypothesis that the control treatment in the second session is equal across treatments in the first session (see the F-tests). The tasks in the two sessions were identical. Nonetheless, we find little evidence of long-lasting effects.

Table 12 summarizes the results for relative performance. While we seldom find statistically significant results, the overall pattern is similar to that for absolute performance. The effect of Group I-I may look larger than those of the other treatments. However, the differences are moderate and standard errors are large. These results indicate that the treatments in the first session may not have a lasting effect.

Figure 6: QTE for relative performance in the first session



Note: This figure plots the QTE functions of Treatments P (upper left), S (upper right), and O (lower left) on the difference between the reported and actual scores in the first session. 95% point-wise bootstrapped confidence bands are the shaded areas. The dashed lines represent the average treatment effects and dotted lines correspond to the 95% confidence interval for the average treatment effect.

5 Robustness checks

In this section, we discuss two main issues regarding the robustness of the main results. The first is that some participants may have engaged in the tasks less seriously. The second issue is attrition in the second session.

We gathered participants and conducted experiments using an online survey platform. While the platform enabled us to collect a large number of demographically heterogeneous participants at a relatively low cost, the main concern was that participants may not engage in the experimental task seriously because it is difficult to monitor them. In an RCT, the effect of treatments should be estimated without bias. However, the baseline level of confidence may be affected by the presence of non-serious participants. Therefore, we rerun our analysis by constructing a subsample that contains participants who appear to have engaged in the experiment seriously. The details of our sample selection strategy and the results are presented in the supplemental materials. Although we obtain similar results in terms of the effects of the treatments, different results are observed regarding the degree of overconfidence. The main difference is that while overplacement is observed without the sample selection, we find underplacement after the sample selection. Another important finding is that the estimated effects of the treatments are larger after the sample selection.

Table 10: Estimation results. Belief updating patterns in the first session

Dependent variable: Reported winning probability						
	(1)	(2)	(3)	(4)	(5)	(6)
Round	1st to 2nd	1st to 2nd	2nd to 3rd	2nd to 3rd	3rd to 4th	3rd to 4th
update12	0.319***	0.371***				
	(0.020)	(0.038)				
win1		-4.301				
		(3.674)				
win1 × update12		0.005				
		(0.062)				
update23			0.320***	0.403***		
			(0.022)	(0.039)		
win2				0.056		
				(3.795)		
win2 × update23				-0.077		
				(0.065)		
update34					0.341***	0.470***
					(0.021)	(0.040)
win3						0.701
						(3.849)
win3 × update34						-0.130*
						(0.066)
Constant	30.127***	29.516***	30.410***	28.950***	29.026***	26.794***
	(1.082)	(1.326)	(1.121)	(1.314)	(1.109)	(1.301)
F-test						
H_0: All terms		3.504		3.524		8.479
involving "win" =0		(0.030)		(0.030)		(0.000)
$\bar{R}^2$	0.184	0.188	0.176	0.181	0.213	0.225
# of obs	1237	1237	1086	1086	1048	1048

Note: Estimated by OLS. Heteroskedasticity robust standard errors are presented within parentheses below the coefficients and p-values are presented within parenthesis below the F statistic. See Table 4 for the variable definition.

We also find that subjects expressed a strong degree of underconfidence (both underestimation and underplacement) in the second session when they were exposed to complete information in the first session. In particular, a particularly large effect was observed when subjects obtained complete information at the end of the first session but did not receive any information in the second session (Group I-V).

We also examine the attrition problem. There is little evidence that the observed characteristics affected the decision to participate in the second session. On the contrary, behaviors in the first session might be related to attrition. To examine whether the attrition problem affects the results for the second section, we re-estimate the models using Heckit and inverse probability weighting and find that the results are similar to those presented in the main text (see the supplemental materials).

Table 11: Estimation results. The effects of interventions in the first session on the second session

Dependent variable: Reported score minus true score in the second session	(1)	(2)	(3)
Group I-I	-5.583***	-5.590***	-4.909***
	(1.284)	(1.291)	(1.117)
Group I-V	-2.667*	-2.630*	-1.625
	(1.322)	(1.327)	(1.165)
Group II-I	-4.556***	-4.497**	-5.282***
	(1.363)	(1.368)	(1.189)
Group III-V	-1.427	-1.158	-1.775
	(1.378)	(1.394)	(1.228)
Group IV-V	-2.336	-2.221	-2.011
	(1.448)	(1.450)	(1.213)
Group V-I	-3.888**	-3.745**	-4.038***
	(1.418)	(1.426)	(1.222)
Female		-2.616***	-1.973**
		(0.737)	(0.635)
Age		0.017	-0.001
		(0.026)	(0.022)
Competitiveness		0.164***	0.140***
		(0.039)	(0.035)
Score		-0.046	0.145***
		(0.033)	(0.030)
Finance Score		-0.288	0.354*
		(0.185)	(0.160)
Schooling		-0.056	-0.154
		(0.276)	(0.236)
Score_2			-0.950***
			(0.028)
Constant	13.381***	10.990***	36.014***
	(0.988)	(2.642)	(2.459)
F test			
H_0: All Group *-I dummies	0.893	1.054	0.590
have the same effect	(0.409)	(0.349)	(0.554)
H_0: All Group *-V dummies	1.534	1.501	1.120
=0	(0.204)	(0.212)	(0.340)
R^2	0.006	0.016	0.276
# of obs	3000	2960	2960

Note: Estimated by OLS. Heteroskedasticity robust standard errors are presented within parentheses below the coefficients and p-values are presented within parenthesis below the F statistic. See Table 4 for the variable definition.

Table 12: Estimation results. The effects of interventions in the first session on the second session

Dependent variable: Reported winning probability minus true probability in the second session							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Round(s)	1st	2nd	3rd	4th	1st and 2nd	2nd and 3rd	3rd and 4th
Group I-I	-4.011*	-3.283	-2.933	-3.524	-1.787	-2.289	-0.883
	(1.918)	(1.948)	(2.063)	(2.115)	(1.953)	(1.984)	(2.015)
Group I-V	-2.040	-1.281	-2.602	-2.240	-1.430	-1.251	-0.729
	(2.066)	(2.106)	(2.179)	(2.254)	(2.145)	(2.163)	(2.162)
Group II-I	-3.127	-0.474	0.287	-0.588	0.363	0.536	3.045
	(1.995)	(2.059)	(2.191)	(2.217)	(2.021)	(2.035)	(2.070)
Group III-V	1.444	1.105	0.438	0.246	1.840	0.805	1.733
	(2.181)	(2.190)	(2.287)	(2.365)	(2.230)	(2.222)	(2.255)
Group IV-V	-1.677	-1.587	-1.696	-2.163	-0.935	-2.151	-0.457
	(2.274)	(2.309)	(2.388)	(2.468)	(2.343)	(2.326)	(2.366)
Group V-I	-3.141	-2.877	-1.128	-2.363	-0.860	-0.953	0.799
	(2.030)	(2.081)	(2.207)	(2.285)	(2.085)	(2.111)	(2.166)
Constant	5.379***	4.104**	3.613*	4.315*	17.446***	17.569***	15.934***
	(1.537)	(1.575)	(1.650)	(1.698)	(1.580)	(1.584)	(1.584)
F test							
H_0: All Group *-I	0.178	1.404	1.468	1.191	0.796	1.306	2.323
dummies have	(0.837)	(0.246)	(0.230)	(0.304)	(0.451)	(0.271)	(0.098)
the same effect							
H_0: All Group *-V	1.132	0.644	0.872	0.677	0.864	0.661	0.486
dummies =0	(0.335)	(0.587)	(0.455)	(0.566)	(0.459)	(0.576)	(0.692)
R^2	0.002	0.000	-0.000	-0.000	-0.000	-0.001	-0.000
# of obs	3000	3000	3000	3000	3000	3000	3000

Note: Estimated by OLS. Heteroskedasticity robust standard errors are presented within parentheses below the coefficients and p-values are presented within parenthesis below the F statistic. See Table 4 for the variable definition.

6 Discussion

This section discusses the theoretical implications of our findings. We argue that our results can be rationalized by the conjecture that information provision has two roles: it raises the accuracy of beliefs (although imperfectly) and it reduces confidence about one's own performance even when it is consistent with prior beliefs. We can make such information processing consistent with limited attention theory but it is still speculative. This is not fully consistent with confirmation bias.

First, two kinds of beliefs play roles in our experiments: participants' beliefs about their own performance and about others' performances. Their evaluation of their own absolute performance depends solely on their belief about their own performance. On the contrary, the evaluation of relative performance depends on both types of beliefs. Different types of information affect these two beliefs differently. Our experimental design thus enables us to separate the effect of information on these two beliefs.

In summary, our results indicate that the beliefs about others' performances change little after receiving information because they are as expected. Information about their own score raises the accuracy of their beliefs about their own performance, although the correction is far from perfect. In additional to these effects, information provision appears to reduce the level of confidence even if the information is as expected.

Our results indicate that information about the distribution of others' scores lowers the belief about one's own performance (i.e., it alleviates overestimation). As shown in Tables 7 and B.2, Treatment O lowered the level of overestimation. In the first session, the degree of overestimation in the control group was 10.821 points. By contrast, overestimation under Treatment O was 1.881 lower after controlling for individual attributes. As such, 17.38% of the overestimation was corrected toward neutral-confidence. Moreover, as shown in Table 8, Treatment O generated lower reported probabilities of winning. Treatment O reduced the overplacement level by 4.405 after controlling for individual attributes; that is, 87.61% of overplacement was corrected to almost neutral-confidence. Figures 4 and 6 show that the effect is homogeneous across subjects.

On the contrary, information about others' scores may not affect the beliefs about others' scores greatly. Because Treatment O yielded lower predictions of their own scores and lower reported probabilities of winning, even if information affects the beliefs about others' scores, the effect on the beliefs of others' scores should be smaller than the effect on the belief about one's own score. While it is somewhat anecdotal, subjects stated that the distribution of scores is "as expected" when it was shown.[24] It is thus sensible to interpret this result as that information on the scores of others may not change the beliefs about others' scores but does affect the evaluation of one's own score and this raises the concern that one's score may be worse.

The presented results indicate that providing information about one's own score raises the

[24]Having been provided with a score distribution table of the test performance of others before predicting their own scores, respondents were asked to choose an answer on a five-point scale: 1 = other participants' scores are very high; 2 = other participants' scores are fairly high; 3 = would not say that other people's scores are either high or low; 4 = other participants' scores are fairly low; and 5 = other participants' scores are very low. The average evaluation by Group I (Treatment O in the absolute evaluation stage) was 2.96, with a standard deviation of 0.870, while the average evaluation by Group III (Treatment O in the relative evaluation stage) was 2.91, with a standard deviation of 0.922. In other words, when provided with the test performances of others as reference information, participants felt the scores were neither high nor low, but as expected. Moreover, Group IV's (Treatment P at the end of the first session) evaluation of the information given after the event was 2.96, with a standard deviation of 0.870. As such, even when provided with reference information on the test performance scores of others after finishing a set of tasks, participant evaluation did not change greatly.

accuracy of a participant's belief about his or her own score. Figure 6 shows that while the effect of information is in the right direction for most people (most underconfident (overconfident) people initially upgrade (downgrade) their beliefs about their relative performance), those who are slightly underconfident also lower their beliefs. There are two possible explanations for this result. One is that information about one's own score indeed lowers the belief about one's own score markedly. The other is that information about one's own score shifts the beliefs about others' scores upward. Although it is difficult to separate these effects, we surmise that information about one's own score is unlikely to shift the beliefs about others' scores given that the initial beliefs about others' scores turn out to be accurate, as discussed above. Moreover, participants appear to have felt that their scores were not as high as expected.[25] We conjecture that providing information about one's own score has two effects: it raises the accuracy of the belief and it lowers the belief about one's own score.

It is thus possible to explain our results in a unified framework by considering the two roles of information provision. The first is standard: information provision raises the accuracy of beliefs. The second is non-standard: it reduces confidence about one's own performance even if the information is as expected and is not directly about one's own performance. Information about the score distribution does not change the beliefs about others' scores because it is as expected, but it nonetheless reduces confidence merely because some information is provided. Although information about one's own score improves the accuracy of the beliefs about one's own score, it may cause underconfidence because information provision also reduces confidence. These results should help us better understand the mechanisms and sources of overconfidence.

The results of this experiment suggest that limited attention plays some role in overconfidence. However, it cannot explain the observed phenomena fully. Even when complete information on one's own performance and that of others is provided, overconfidence is not completely corrected, suggesting the presence of limited attention. Partial information has a stronger effect than complete information. We may thus speculate that providing information partially may be more effective for improving attention. However, we cannot rule out other possible explanations. Further, confirmation bias may also be unable to explain our findings. While some of the results are consistent with confirmation bias, we found no evidence that people react to favorable information. Moreover, slightly overconfident people become underconfident in the experiment, which is hard for confirmation bias to explain. We reject the simplistic view that overconfidence derives from a lack of information and can thus be eliminated by information provision. In addition, we do not find evidence that known behavioral bias can explain overconfidence. Rather, overconfidence arises from a complex interaction of cognitive processes.

[25] Having been provided with their own test performance score as reference information before predicting the probability of scoring higher than an opponent, respondents were asked to choose an answer on a five-point scale: 1 = their own score is very high; 2 = their own score is fairly high; 3 = would not say that their own score is either high or low; 4 = their own score is fairly low; and 5 = their own score is very low. The average evaluation by Group I (Treatment P in the relative evaluation stage) was 3.55, with a standard deviation of 1.086, while the average evaluation by Group II (Treatment S in the relative evaluation stage) was 3.67, with a standard deviation of 1.051. In other words, when provided with their own test performance scores as reference information, participants felt their own test performances were lower than expected. The average evaluation by Group IV (Treatment N in the relative evaluation stage and Treatment P at the end of the first session) after the event was 3.55, with a standard deviation of 1.035. Thus, after finishing a set of tasks, the evaluation did not change greatly even when provided with one's own test performance scores as reference information.

7 Conclusion

This paper presents the results of an RCT of the effect of information provision on overconfidence. While research on overconfidence is widespread, few studies have used RCTs. However, given the growing recognition of their importance, we argue that this study's illustration of the impact of information on overconfidence is an important contribution to the literature. Moreover, we recruited a large number of participants from the general public that reflected the composition of the working age population in Japan better than with laboratory-based experiments. While some sample selection bias might have arisen from their participation in an online test, we believe that this method nonetheless enabled us to investigate behavior closer to that of ordinary people than if we had used university students from our faculty.

These results can guide how to model the effect of information. In this study, we argue that information provision not only raises the accuracy of beliefs, but also reduces confidence even if it is as expected. In particular, when provided with information on other people's test performances, overestimation was corrected despite judging the scores of others to be as expected. In other words, even when information about other people is more or less correct, the provision of information on others affects self-evaluation and the effect is homogeneous.

This study has some limitations. As the respondents answered questions through an online test with no supervision, it was hard to gauge the extent to which they participated in the test seriously. While we investigated this problem in Section 5 and found most of our results to indeed be robust, we may still need to be cautious when drawing conclusions. In particular, caution should be exercised with regard to the degree of overplacement because the result could be affected by how seriously participants treated the exercise. On the contrary, because an RCT was used, the intervention effect can be considered to have been properly estimated.

Appendix

A Mathematical appendix

A.1 Incentive scheme

We used an incentive-compatible mechanism to elicit participants' beliefs about their own scores and their relative performance. The method is proposed by Grether (1981), Holt (2007), and Karni (2007) as well as used in Mobius et al. (2014) under the name of the crossover mechanism.

Under this mechanism, it is optimal for participants to report the mean of their beliefs under a weak condition. For example, this mechanism is incentive compatible without risk neutrality and even expected utility maximization is not needed.[26] Here, we provide a proof for the case of expected utility maximization. See Karni (2007), among others, for a proof in more general settings. This method elicits the mean of the belief about $p \in [0, 1]$, which is the probability of the event occurring. Let $\Pi(p)$ denote the belief. Here, we assume for simplicity that $\Pi(p)$ exhibits

[26] The method proposed by Allen (1987) and binarized scoring rule of Hossain and Okui (2013) are also incentive compatible without risk neutrality (and without expected utility maximization). However, these methods may create a hedging problem by providing an incentive to obtain a score from the quizzes that simplifies the prediction. Mobius et al. (2014) point out this problem for the quadratic scoring rule. Because the methods of Allen (1987) and Hossain and Okui (2013) extend the quadratic scoring rule, they also share the same problem.

density $\pi(p)$. We normalize the utility of no award to zero and the utility of receiving an award is denoted as u. Let $y \sim U[0,1]$. Expected utility under this scheme when a subject reports x is

$$\begin{aligned}
&E\left(\mathbf{1}\{y > x\}\frac{y}{100}u + \mathbf{1}\{y \leq x\}pu\right) \\
&= \left(\frac{1 + x/100}{2}(1 - x/100) + (x/100)E(p)\right)u \\
&= \left(\frac{1 - (x/100)^2}{2} + (x/100)E(p)\right)u.
\end{aligned}$$

It is easy to see that $x = 100E(p) = 100\int p d\Pi(p)$ maximizes expected utility.

A.2 Belief updating

Our experimental design enables us to examine how beliefs are updated and whether the observed belief updating is consistent with Bayesian updating. The key factor is that we elicit not only the probability of winning but also the probability of successive wins.

First, we examine the change in the belief after a win. Let p be the probability of winning, which is the parameter in our case, and $\pi(p)$ be the prior over p. Bayes' law implies that the posterior follows

$$\pi(p|W) = \frac{p\pi(p)}{\int p'\pi(p')dp'},$$

where W denotes the event of winning. Thus, the posterior mean is

$$E(p|W) = \int p\pi(p|W)dp = \frac{\int p^2\pi(p)dp}{\int p'\pi(p')dp'} = \frac{E(p^2)}{E(p)}.$$

$E(p^2)$ is the mean of the belief of successive wins. When participants lose (this event is denoted as L), the posterior mean under Bayesian updating is

$$E(p|L) = 1 - \frac{E((1-p)^2)}{E(1-p)} = 1 - \frac{1 - 2E(p) + E(p^2)}{1 - E(p)} = \frac{E(p) - E(p^2)}{1 - E(p)}.$$

These results imply that the posterior mean under Bayesian updating can be computed by eliciting the probabilities of both a single win and successive wins.

B Additional tables

Table B.1: Balance test results for the second session

Dependent variable	Female	Age	Competitiveness	Score	Finance Score	Schooling
Group I-I	-0.002	0.088	-0.566	0.419	-0.060	-0.115
	(0.034)	(0.920)	(0.618)	(0.730)	(0.136)	(0.811)
Group I-V	-0.000	-0.070	-0.644	0.771	-0.014	-0.265
	(0.034)	(0.925)	(0.618)	(0.722)	(0.135)	(0.791)
Group II-I	-0.003	-0.078	-0.528	-1.022	-0.013	-0.396
	(0.036)	(0.976)	(0.646)	(0.764)	(0.143)	(0.814)
Group III-V	-0.001	-0.142	-1.159	-1.026	-0.170	0.318
	(0.036)	(0.974)	(0.655)	(0.761)	(0.144)	(0.906)
Group IV-V	-0.000	-0.003	-0.339	0.280	0.160	-0.832
	(0.037)	(0.993)	(0.683)	(0.803)	(0.147)	(0.769)
Group V-I	-0.003	-0.029	0.168	-0.413	0.027	-0.907
	(0.037)	(1.000)	(0.696)	(0.790)	(0.148)	(0.769)
Constant	0.501***	44.477***	35.688***	34.533***	4.216***	4.675***
	(0.026)	(0.697)	(0.471)	(0.550)	(0.102)	(0.626)
F test						
H_0: All group	0.002	0.013	0.862	1.926	0.904	0.700
dummies =0	(1.000)	(1.000)	(0.522)	(0.073)	(0.491)	(0.650)
R^2	-0.002	-0.002	-0.000	0.002	-0.000	-0.001
# of obs	3000	3000	3000	3000	3000	2960

Note: Estimated by OLS. Heteroskedasticity robust standard errors are presented within parentheses below the coefficients and p-values are presented within parenthesis below the F statistics. See Table 4 for the variable definition.

Table B.2: Estimation results. Prediction error about absolute performance in the second session

Dependent variable: Reported score minus true score in the second session				
	(1)	(2)	(3)	(4)
Treatment O	-3.078***	-3.083***	-3.424***	-2.409
	(0.716)	(0.714)	(0.614)	(2.349)
Female		-2.455***	-1.908**	-1.575
		(0.712)	(0.636)	(0.858)
Age			0.002	0.002
			(0.022)	(0.022)
Competitiveness			0.144***	0.144***
			(0.035)	(0.035)
Score_2			-0.921***	-0.921***
			(0.028)	(0.028)
Finance Score			0.511**	0.566**
			(0.156)	(0.210)
Schooling			-0.208	-0.200
			(0.237)	(0.329)
Treatment O × Female				-0.797
				(1.277)
Treatment O × Finance Score				-0.135
				(0.310)
Treatment P × Schooling				-0.017
				(0.468)
Constant	11.683***	12.913***	37.814***	37.399***
	(0.483)	(0.594)	(2.198)	(2.359)
R^2	0.006	0.009	0.270	0.269
# of obs	3000	3000	2960	2960

Note: Estimated by OLS. Heteroskedasticity robust standard errors are presented below the coefficients (within parentheses). See Table 4 for the variable definition.

Table B.3: Estimation results. Prediction error about relative performance in the second session

Dependent variable: Reported winning probability minus true probability in the second session								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Round	1st	1st	2nd	2nd	3rd	3rd	4th	4th
Treatment P	-2.830**	-3.433***	-1.802	-2.432***	-0.325	-1.078	-1.151	-1.882*
	(1.047)	(0.709)	(1.061)	(0.724)	(1.120)	(0.775)	(1.150)	(0.810)
Female		-0.943		-0.454		-1.193		-1.440
		(0.730)		(0.741)		(0.794)		(0.838)
Age		-0.042		-0.066*		-0.088**		-0.061*
		(0.026)		(0.026)		(0.028)		(0.029)
Competitiveness		0.061		0.049		0.035		0.027
		(0.038)		(0.040)		(0.043)		(0.045)
Score_2		-2.025***		-2.025***		-2.088***		-2.117***
		(0.039)		(0.039)		(0.041)		(0.043)
Finance Score		0.322		0.134		-0.054		-0.009
		(0.181)		(0.179)		(0.193)		(0.202)
Schooling		-0.805**		-0.614*		-0.677*		-0.470
		(0.263)		(0.269)		(0.289)		(0.304)
Constant	4.724***	76.039***	3.622***	76.293***	2.533**	80.286***	3.197***	80.284***
	(0.764)	(2.610)	(0.768)	(2.680)	(0.792)	(2.812)	(0.822)	(2.992)
$\bar{R}^2$	0.002	0.571	0.001	0.560	-0.000	0.544	-0.000	0.527
# of obs	3000	2960	3000	2960	3000	2960	3000	2960

Note: Estimated by OLS. Heteroskedasticity robust standard errors are presented below the coefficients (within parentheses). See Table 4 for the variable definition.

Table B.4: Estimation results. Prediction error about successive wins in the second session

Dependent variable: Reported successively winning probability minus true probability in the second session			
	(1)	(2)	(3)
Rounds	1st and 2nd	2nd and 3rd	3rd and 4th
Treatment P	-0.643	-0.348	0.738
	(1.069)	(1.082)	(1.114)
Constant	17.243***	16.898***	16.029***
	(0.789)	(0.788)	(0.800)
$\bar{R}^2$	-0.000	-0.000	-0.000
# of obs	3000	3000	3000

Note: Estimated by OLS. Heteroskedasticity robust standard errors are presented below the coefficients (within parentheses). See Table 4 for the variable definition.

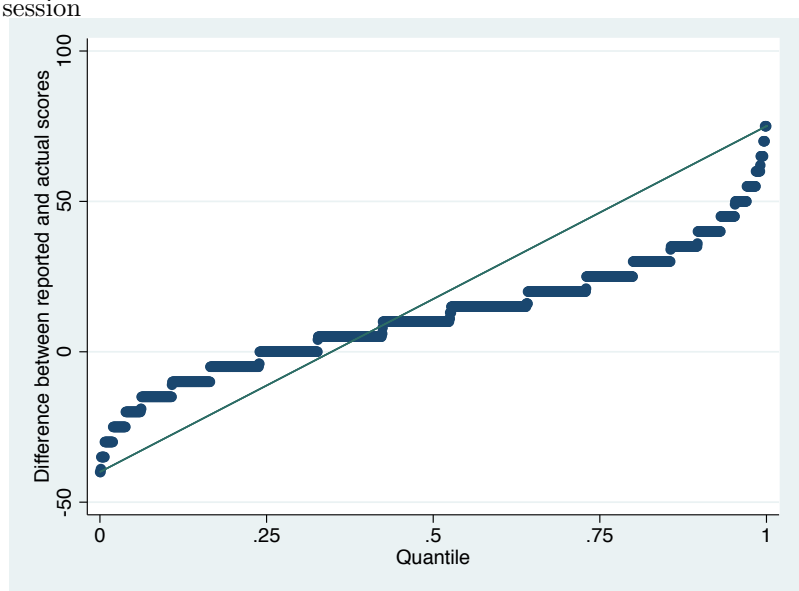
Table B.5: Estimation results. Belief updating patterns in the second session

Dependent variable: Reported winning probability in the second session						
	(1)	(2)	(3)	(4)	(5)	(6)
Round	1st to 2nd	1st to 2nd	2nd to 3rd	2nd to 3rd	3rd to 4th	3rd to 4th
update12_2	0.339***	0.378***				
	(0.020)	(0.036)				
win1_2		-1.824				
		(3.452)				
win1_2 × update12_2		-0.013				
		(0.059)				
update23_2			0.298***	0.394***		
			(0.023)	(0.040)		
win2_2				-8.209*		
				(3.876)		
win2_2 × update23_2				0.012		
				(0.065)		
update34_2					0.315***	0.487***
					(0.024)	(0.044)
win3_2						-6.183
						(3.828)
win3_2 × update34_2						-0.080
						(0.068)
Constant	28.386***	27.801***	31.188***	30.086***	29.885***	27.311***
	(1.082)	(1.299)	(1.217)	(1.427)	(1.279)	(1.508)
F test						
H_0: All terms		1.452		9.848		16.902
involving "win" =0		(0.235)		(0.000)		(0.000)
R^2	0.231	0.232	0.170	0.184	0.181	0.208
# of obs	1005	1005	907	907	844	844

Note: Estimated by OLS. Heteroskedasticity robust standard errors are presented below the coefficients (within parentheses) and p-values are presented below the F statistic (within parenthesis). See Table 4 for the variable definition.

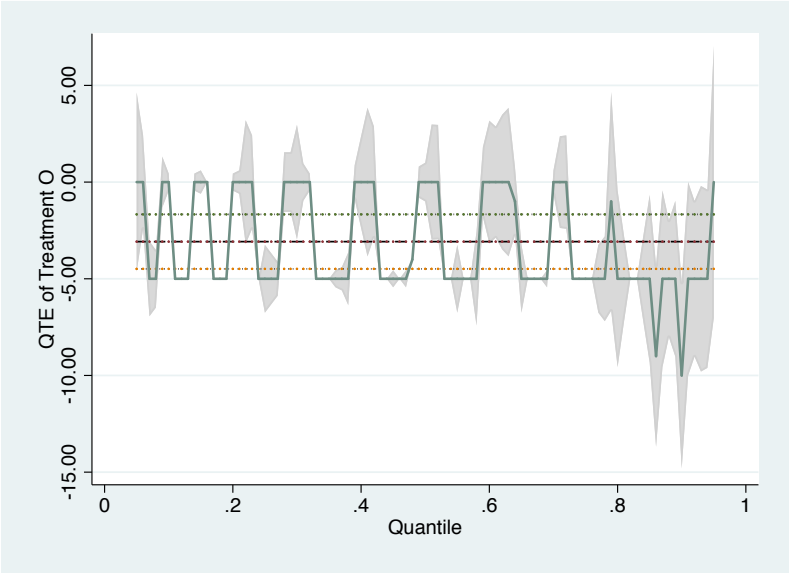
C Additional figures

Figure C.1: Quantile of the difference between the reported and actual scores for the control group in the second session



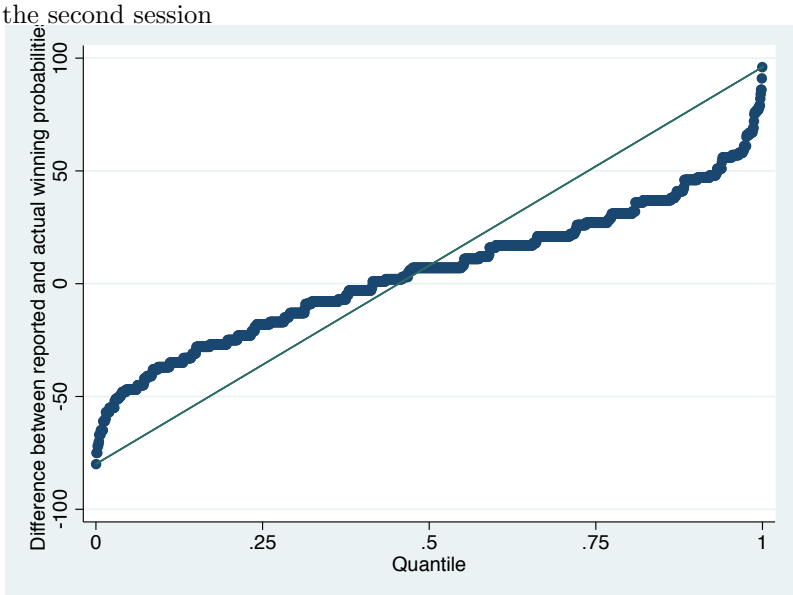
Note: This figure plots the quantile curve of the difference between the reported and actual scores for the control group in the second session. The quantile curve crosses zero at the 33% quantile. The right-hand side of the 33% quantile is above zero and this corresponds to overconfidence. The left-hand side is below zero, corresponding to underconfidence.

Figure C.2: QTE of Treatment O for absolute performance in the second session



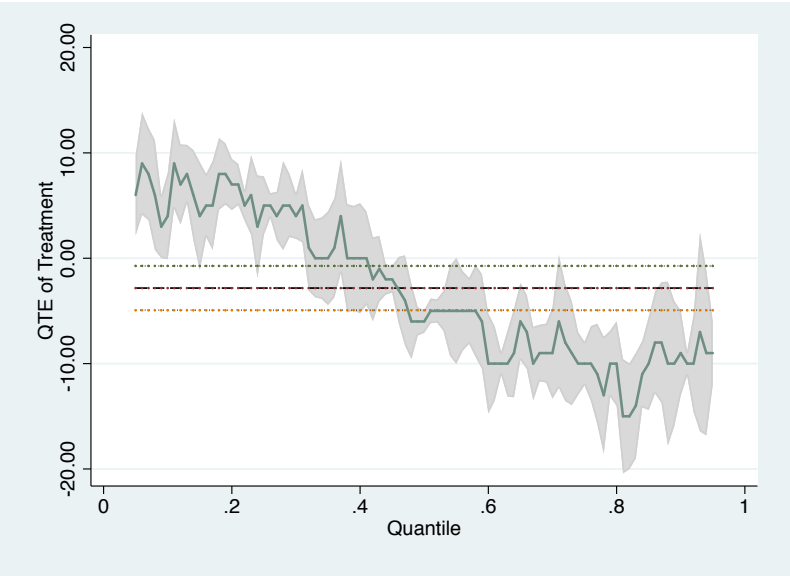
Note: This figure plots the QTE function of Treatment O on the difference between the reported and actual scores in the second session. A 95% point-wise bootstrapped confidence band is the shaded area. The dashed line represents the average treatment effect and the dotted lines correspond to the 95% confidence interval for the average treatment effect.

Figure C.3: Quantile of the difference between the reported and actual winning probabilities under the control in the second session



Note: This figure plots the quantile curve of the difference between the reported and actual winning probabilities for the control group in the second session. The quantile curve crosses zero at the 42% quantile. The right-hand side of the 42% quantile is above zero and this corresponds to overconfidence. The left-hand side is below zero, corresponding to underconfidence.

Figure C.4: QTE of Treatment P for relative performance in the second session



Note: This figure plots the QTE function of Treatment O on the difference between the reported probability and true probability in the second session. A 95% point-wise bootstrapped confidence band is the shaded area. The dashed line represents the average treatment effect and the dotted lines correspond to the 95% confidence interval for the average treatment effect.

References

Allen, Franklin (1987). "Discovering Personal Probabilities When Utility Functions are Unknown". In: *Management Science* 33, pp. 542–544.

Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy (2018). "Incentives in Experiments: A Theoretical Analysis". In: *Journal of Political Economy* 126.4, pp. 1472–1503.

Barber, Brad M. and Terrance Odean (2001). "Boys will be Boys: Gender, Overconfidence, and Common Stock Investment". In: *Quarterly journal of Economics*, pp. 261–292.

Barron, Kai (2017). "Belief Updating: Does the 'Good-news, Bad-news' Asymmetry Extend to Purely Financial Domains?" mimeo.

Benoît, Jean-Pierre and Juan Dubra (2011). "Apparent overconfidence". In: *Econometrica* 79.5, pp. 1591–1625.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman (2010). "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment". In: *Quarterly Journal of Economics* 125.1, pp. 263–306.

Bertrand, Marianne and Adair Morse (2011). "Information Disclosure, Cognitive Biases, and Payday Borrowing". In: *Journal of Finance* 66.6, pp. 1865–1893.

Brandts, Jordi, Valeska Groenert, and Christina Rott (2015). "The Impact of Advice on Women's and Men's Selection into Competition". In: *Management Science* 61.5, pp. 1018–1035.

Camerer, Colin and Dan Lovallo (1999). "Overconfidence and Excess Entry: An Experimental Approach". In: *American Economic Review* 89.1, pp. 306–318.

Caplin, Andrew (2016). "Measuring and Modeling Attention". In: *Annual Review of Economics* 8, pp. 379–403.

Charness, Gary, Aldo Rustichini, and Jeroen van de Ven (2018). "Self-confidence and Strategic Behavior". In: *Experimental Economics* 21, pp. 72–98.

Chetty, Raj, Adam Looney, and Kory Kroft (2009). "Salience and Taxation: Theory and Evidence". In: *American Economic Review* 99.4, pp. 1145–1177.

Coffman, Katherine Baldiga (2014). "Evidence on Self-stereotyping and the Contribution of Ideas". In: *Quarterly Journal of Economics* 129.4, pp. 1625–1660.

Deshmukh, Sanjay, Anand M. Goel, and Keith M. Howe (2013). "CEO Overconfidence and Dividend Policy". In: *Journal of Financial Intermediation* 22.3, pp. 440–463.

Eil, David and Justin M. Rao (2011). "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself". In: *American Economic Journal: Microeconomics* 3.2, pp. 114–138.

Enke, Benjamin and Florian Zimmermann (2019). "Correlation neglect in belief formation". In: *The Review of Economic Studies* 86.1, pp. 313–332.

Ertac, Seda (2011). "Does Self-relevance Affect Information Processing? Experimental Evidence on the Response to Performance and Non-performance Feedback". In: *Journal of Economic Behavior & Organization* 80.3, pp. 532–545.

Fan, Yanqin and Sang Soo Park (2012). "Confidence Intervals for the Quantile of Treatment Effects in Randomized Experiments". In: *Journal of Econometrics* 167.2, pp. 330–344.

Grether, David M. (1981). "Financial Incentive Effects and Individual Decision Making". Social Science Working Paper 401, California Institute of Technology.

Grossman, Zachary and David Owens (2012). "An Unlucky Feeling: Overconfidence and Noisy Feedback". In: *Journal of Economic Behavior & Organization* 84.2, pp. 510–524.

Handel, Benjamin R. and Jonathan T. Kolstad (2015). "Health Insurance for "Humans": Information Frictions, Plan Choice, and Consumer Welfare". In: *American Economic Review* 105.8, pp. 2449–2500.

Hoelzl, Erik and Aldo Rustichini (2005). "Overconfident: Do You Put Your Money On It?" In: *Economic Journal* 115.503, pp. 305–318.

Holt, Charles (2007). *Markets, Games & Strategic Behavior*. Boston: Pearson/Addison-Wesley.

Horino, Midori (1987). "Analysis and Reconsideration of the Concept of Achievement Motive". In: *Japanese Journal of Educational Psychology* 35.148–154. in Japanese.

Hossain, Tanjim and Ryo Okui (2013). "The Binarized Scoring Rule". In: *Review of Economic Studies* 80, pp. 984–1001.

Karni, Edi (2007). "A Mechanism for Eliciting Probabilities". In: *Econometrica* 77, pp. 603–606.

Korn, C. W., T. Sharot, H. Walter, H. R. Heekeren, and R. J. Dolan (2014). "Depression is Related to an Absence of Optimistically Biased Belief Updating about Future Life Events". In: *Psychological Medicine* 44.3, pp. 579–592.

Landier, Augustin and David Thesmar (2009). "Financial Contracting with Optimistic Entrepreneurs". In: *Review of Financial Studies* 22.1, pp. 117–150.

Lord, Charles G, Lee Ross, and Mark R. Lepper (1979). "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence". In: *Journal of Personality and Social Psychology* 37.11, pp. 2098–2109.

Malmendier, Ulrike and Geoffrey Tate (2005). "CEO Overconfidence and Corporate Investmen". In: *Journal of Finance* 60.6, pp. 2661–2700.

— (2008). "Who Makes Acquisitions? CEO Overconfidence and the Market's Reaction". In: *Journal of Financial Economics* 89.1, pp. 20–43.

— (2015). "Behavioral CEOs: The Role of Managerial Overconfidence". In: *Journal of Economic Perspectives* 29.4, pp. 37–60.

Malmendier, Ulrike, Geoffrey Tate, and Jon Yan (2011). "Overconfidence and Early–Life Experiences: The Effect of Managerial Traits on Corporate Financial Policies". In: *Journal of Finance* 66.5, pp. 1687–1733.

Mobius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat (2014). "Managing Self-confidence: Theory and Experimental Evidence". mimeo.

Moore, Don A. and Paul J. Healy (2008). "The Trouble with Overconfidence." In: *Psychological review* 115.2, p. 502.

Moore, Don A. and Derek Schatz (2017). "The Three Faces of Overconfidence". In: *Social and Personality Psychology Compass* 11.8, e12331.

Offerman, Theo, Joep Sonnemans, Gijs van de Kuilen, and Peter P. Wakker (2009). "A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes". In: *Review of Economic Studies* 76, pp. 1461–1489.

Ortoleva, Pietro and Erik Snowberg (2015). "Overconfidence in Political Behavior". In: *American Economic Review* 105.2, pp. 504–535.

Rabin, Matthew and Joel L Schrag (1999). "First Impressions Matter: A Model of Confirmatory Bias". In: *Quarterly Journal of Economics* 114.1, pp. 37–82.

Shah, Anuj K., Sendhil Mullainathan, and Eldar Shafir (2012). "Some Consequences of Having Too Little". In: *Science* 338.6107, pp. 682–685.

Stango, Victor and Jonathan Zinman (2015). "Borrowing High versus Borrowing Higher: Price Dispersion and Shopping Behavior in the U.S. Credit Card Market". In: *Review of Financial Studies* 29.4, pp. 979–1006.

Zhang, Hanzhe (2013). "Evolutionary Justifications for Non-Bayesian Beliefs". In: *Economics Letters* 121.2, pp. 198–201.

Zhang, Jie and Els Breugelmans (2012). "The impact of an item-based loyalty program on consumer purchase behavior". In: *Journal of Marketing Research* 49.1, pp. 50–65.