

# Matching and Confidence\*

MEHMET YIGIT GURDAL  
Bogazici University

ELISABETTA LENI  
University of Essex

FRIEDERIKE MENGEL †  
University of Essex  
and Lund University

May 5, 2020

## Abstract

We study how matching affects confidence. Our lab experiment allows us to identify the effect of being matched with others of either similar or dissimilar performance (assortative or disassortative matching) on people's confidence in their own ability. Across a variety of tasks we find that assortative matching does not have a substantial nor statistically significant effect on confidence compared to a control group with random matching. By contrast, disassortative matching has a negative effect on confidence on average that is driven by the bottom half of performers. This group becomes substantially less confident compared to random matching. We discuss potential mechanisms and implications of this result.

*Keywords: Matching, Belief Formation, Confidence.*

*JEL Classification Numbers: C90, C91, D90.*

---

\*We thank the British Academy (Newton Mobility Grant NMG2R2-100082) for valuable financial support.

†Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ. Department of Economics, Lund University, SE-220 07 Lund (SE); *e-mail*: fr.mengel@gmail.com

# 1 Introduction

A crucial aspect of institutional design, relevant for workplace organization, schools and the higher educations sector alike, is how to match people to achieve the best outcomes. Outcomes that have received a lot of attention in the literature include the academic performance of students (Feld & Zoelitz, 2017; Hanushek & Woessmann, 2006; Sacerdote, 2001), the performance of workers and work teams (Bandiera, Barankay, & Rasul, 2010; Jackson & Bruegmann, 2009; Mas & Moretti, 2009) as well as cooperation and pro-social behaviour in groups (Branas-Garza et al., 2010; Currarini & Mengel, 2016; Fershtman & Gneezy, 2001; Grimm & Mengel, 2009). One outcome that has received much less attention in this context is confidence. This is despite the fact that confidence has been shown to be an important outcome explaining gender differences in competitiveness and leadership among young adults (Alan & Ertac, 2019; Alan, Ertac, Kubilay, & Loranth, 2019), intergenerational income mobility (Blanden, Gregg, & Macmillan, 2007) and academic performance of students (Golsteyn, Non, & Zoelitz, 2019) among other things.<sup>1</sup>

In this paper we focus on how matching affects confidence. We design a lab experiment that allows us to identify the effect of being matched with others of either similar or dissimilar performance (assortative or disassortative matching) on people’s confidence in their own ability. Across a variety of tasks we find that assortative matching does not have a substantial nor statistically significant effect on confidence compared to a control group with random matching. By contrast, disassortative matching has a negative effect on confidence on average that is driven by the bottom half of performers. This group becomes substantially less confident compared to random matching. However they become also more accurate, i.e. less overconfident, with disassortative compared to random matching. We also find that participants react more strongly to negative than to positive feedback on average, but there is some heterogeneity across tasks with this finding.

These results are relevant for educational tracking within schools, matching students across schools, matching workers in teams and the selection of peers more generally (Golsteyn et al., 2019). They are particularly relevant in contexts, such as education, where confidence is considered an important outcome (Anderson, Brion, Moore, & Kennedy, 2012; Ytterberg et al., 1998). To the extent that our results have external validity in the specific domain considered, they suggest actionable consequences for institutional design.

The paper is organized as follows. Section 2 discusses related literature. Section 3 contains details of our experimental design. Section 4 contains our main results and Section 5 concludes. A series of Online Appendices contain experimental instructions, screenshots and additional tables and figures.

## 2 Literature

Our paper contributes to two so far largely disjoint strands of literature. First we contribute to an active literature on how matching affects a variety of different outcomes (usually not including confidence). Second we contribute to literature aimed at understanding the various ways in which (over-) confident beliefs arise in a dynamic setting.

---

<sup>1</sup>Heckman, Stixrud, and Urzua (2006) highlight the importance of non-cognitive skills more generally.

There is a substantial literature on how matching affects outcomes including academic performance of students (Feld & Zoelitz, 2017; Hanushek & Woessmann, 2006; Sacerdote, 2001), the performance of workers and work teams (Bandiera et al., 2010; Jackson & Bruegmann, 2009; Mas & Moretti, 2009) as well as cooperation and pro-social behaviour in groups (Branas-Garza et al., 2010; Currarini & Mengel, 2016; Fershtman & Gneezy, 2001; Grimm & Mengel, 2009). Apart from the different outcomes studied, this literature also differs by whether matching is exogenous or endogenous and it includes a large literature on peer effects. Our paper is more closely related to papers where matching, as in our case, is exogenous. Those include Feld and Zoelitz (2017) who show that university student’s performance depends on the performance of other students in their class in a non-monotonic way or Mas and Moretti (2009) who show that the productivity of supermarket cashiers depends on which other cashiers work on the same shift. Hanushek and Woessmann (2006) show that early educational tracking in schools increases inequality. Their analysis also suggests that early tracking might reduce mean performance.<sup>2</sup> Sacerdote (2014) reviews some of the large literature on peer effects.

Our main contribution to this literature is to consider confidence as a key outcome variable. There are only very few papers on matching that include confidence as an outcome. Those can mostly be found in the peer effects literature (see for example Antonio, 2004). These papers differ from ours in that they consider the influence of endogenously selected friendship groups as opposed to assortative or disassortative matching based on performance.

Our paper also contributes to literature aimed at understanding the various ways in which (over-) confident beliefs come about. In standard economic models, beliefs matter only through their instrumental value in decision making. Recent theoretical work has relaxed this assumption and assumed that people could gain direct utility by holding optimistic beliefs on qualities that are relevant for the self (Koeszegi, 2006). When beliefs entail a direct utility, belief updating can depart from the Bayesian benchmark towards optimistic updating in a self-serving way.

There is a growing experimental literature on asymmetric updating, and the results are mixed. Mobius (2011) and Mobius, Niederle, Niehaus, and Rosenblat (2007) find that subjects who receive positive feedback in an IQ test revise their beliefs significantly more than those who receive negative feedback. Eil and Rao (2011) also find that subjects asymmetrically update their beliefs on intelligence, and physical attractiveness. They adhere quite closely to the Bayesian benchmark in case of positive signals, but they discount or ignore the signal when it is negative. In Zimmermann (2019), subjects perform an IQ test and receive feedback. He finds little evidence for asymmetry in the short run, but subjects recall negative feedback with lower accuracy one month after receiving the feedback. Sharot, Korn, and Dolan (2011) find that subjects updated their beliefs more in response to information that was better than expected than to information that was worse than expected. Other authors find evidence of asymmetric updating in the opposite direction - negative signals weighted more than positive ones. Ertac (2011) studies belief updating across tasks with different degrees of self-relevance. The results of her study indicate that subjects attribute more weight to negative signals than positive ones in the self-relevant context but not in the neutral one. Coutts (2019) also examines whether updating differs across ego-relevant and neutral contexts. His results show that negative signals receive

---

<sup>2</sup>Tracking usually includes assortative matching, but also a variety of other measures (differing teaching materials etc). Hence, studying the effect of tracking does not usually identify the pure effect of matching.

more weight than positive ones but these deviations do not differ across contexts. Our main contribution to this literature is to study how matching affects belief updating and in particular confidence and accuracy of beliefs across tasks with different degrees of ego-relevance and prior strength.

### 3 Design

In this section we describe the experimental design and procedures. Our experiment consists of a  $3 \times 3$  design where we vary the type of task and the type of matching. In all treatments participants go through the following stages: introduction, experimental task and belief updating. We will describe these stages in turn.

**Introduction stage** In this stage, we show participants information about the task they will be asked to complete in the next stage, and describe how their earnings from the task stage are going to be calculated. At the end of this introduction, participants are asked to answer two questions regarding their prior beliefs. In particular, they are asked to state (i) what they believe the average score is going to be in that session and (ii) where they believe they rank among the participants of that session.<sup>3</sup>

**Task stage** In the **task stage**, participants perform either one of the following tasks (i) observing pairs of contemporary paintings on the screen and guessing which painting received a higher price at an auction (**ART**), (ii) solving the Raven matrices task (**IQ**) or (iii) observing a pair of footballers on the screen and guessing which one of them scored more goals during a specified season (**FOOT**). We now describe details of each task.

The **ART** task consists in comparing pairs of paintings sold at a real auction.<sup>4</sup> The pictures of the paintings, the title of the work of art, and the author appear on the computer screen and participants are asked to indicate which painting was sold at the higher price. The task involves the comparison of 15 different pairs of paintings. The score is calculated as the total number of correct answers across the 15 pairs.

The **IQ** task consists in Raven’s progressive matrices task, which is a 60-item test used in measuring abstract reasoning and regarded as a non-verbal estimate of fluid intelligence. For this task, one of the test questions is used as an example during the Introduction stage, and the task involves the 59 remaining questions. All of the questions on the Raven’s test consist of visual geometric design with a missing piece. The test taker is given six to eight choices to pick from and fill in the missing piece. Participants have 10 minutes to complete as many questions as they can. The score is calculated as the total number of correct answers given within the time limit.

The **FOOT** task consists in comparing the number of goals scored by two football players of the Turkish Super League. The pictures of the players, and the team they played in the season 2017-2018 appear on the computer screen and participants are asked to indicate which

---

<sup>3</sup>Each experimental session had 16 participants. See below.

<sup>4</sup>Impressionist & Modern Art Evening Sale conducted by Christie’s London on February 2, 2016.

player scored more goals in that season. The task involves the comparison of 15 different pairs of players. The score is calculated as the total number of correct answers across the 15 pairs.

The different tasks were chosen to be able to detect if main results were driven by task-specific properties. They primarily present variation along two dimensions: (i) ego-relevance (Coutts, 2019; Ertac, 2011) and (ii) the strength of the prior about one’s rank. Our assumption is that - on average - **ART** has both low ego-relevance and a weak prior. **IQ** has high ego-relevance and a strong prior and **FOOT** has high ego-relevance for some, and low for others and a strong prior for both groups.

**Belief revision stage** In the **belief revision stage**, participants are first shown their score and then asked a series of questions regarding their rank among the 16 participants in a session. Here, participants observe 8 subgroups (rank 1 or 2, rank 3 or 4,...rank 15 or 16) and for each subgroup, they specify the probability that their actual rank falls in that subgroup. In Step 2, participants receive information on the score of another participant from the session; in the assortative matching condition (*Assortative*), the score is that of a participant who is ranked similarly to them. Specifically participants with rank 1-8 are shown the score of participant ranked +1 (1 observes 2, 2 observes 3, ... , 8 observes 9) while participants with rank 9-16 are shown the score of participant ranked -1 (9 observes 8, 10 observes 9, ... , 16 observes 15). In the disassortative matching treatment (*Disassortative*), the score is that of a participant who is ranked differently to them. Specifically, participants ranked 1 to 8 are shown the score of participant ranked 16 and participants ranked 9 to 16 observe the participant ranked first. In the random condition (*Random*), they are shown the score of a randomly selected participant. Participants are then asked to state their beliefs regarding their ranking, by specifying the probabilities for 8 subgroups as in the previous step.

In subsequent steps, we disclose pieces of information while keeping track of how participants update their beliefs on their ranking. In Step 3 we reveal to each participant in which half of the distribution they performed and we ask them to guess in which one of the four remaining subgroups they think they performed. In Step 4 we reveal in which quarter of the score distribution they rank and we ask to guess in which of the two subgroups of that specific quarter they think they performed. Finally (Step 5), we tell participants in which subgroup they performed, and we ask them to guess their exact ranking. After this last step the true rank of the participant is revealed.

		Task		
		ART	IQ	FOOT
Matching	Random (C)	64	64	64
	Assortative (AT)	64	64	64
	Disassortative (DAT)	63	64	64

Table 1: Number of observations. One participant had to be dropped in ART-DAT as he had already participated in a prior session.

**Payments** The experiment is incentivized based on the performance in the **Task stage** and the accuracy of one of the guesses in the **Belief revision stage**. For the **Task stage**, participants are paid 1 TL for each correct answer in the ART task, 0.5 TL for each correct answer

in the IQ task and 2 TL for each correct answer in the FOOT task.<sup>5</sup> For the **Belief revision stage**, one of the 5 steps where participants stated their beliefs was randomly selected and participants were paid based on their accuracy for this step. In particular, we used a logarithmic rule and the earnings are calculated according to the formula  $32 + 32 \log(p)$ . Here,  $p$  is the number we find by dividing the probability that the participant assigns to the subgroup that contains her actual ranking. Since  $p \in [0, 1]$ ,  $\log(p)$  is a zero or negative number, the earnings here could be at most 32 TL from this part.<sup>6</sup> While novel, the scoring rule we used here is a proper scoring rule, that is, it incentivizes the participants to state their true beliefs. To see this, suppose there are two categories, *high* & *low*, and the participant’s belief that her rank is *high* is  $b$ . When her reported belief is  $p$ , her expected payoff becomes  $32(1 + b \log(p) + (1 - b) \log(1 - p))$ . Given  $b$ , this is a concave function of  $p$  and the first order conditions for maximization implies  $(1 - p)b + p(1 - b) = 0$ , hence  $p = b$ . The rule and payoff consequences were explained in detail in the Experimental Instructions (see Appendix B).

**Questionnaire** At the end of the experiment, participants are asked to complete a post-experimental questionnaire eliciting a number of demographics as well as measures of risk attitude and trust. The full list of questions can be found in Appendix C.

**Other Details** The experiment was conducted at the experimental lab at Bogazici University between November 2018 (ART task) and May 2019 (FOOT task). 575 students participated in our experiment (64 per treatment).<sup>7</sup> Ethical approval was obtained in March 2018 by the Faculty Ethics Committee of the University of Essex (under Annex B). The experiment was programmed using the software z-tree (Fischbacher, 2007) and we used ORSEE to recruit subjects (Greiner, 2004).

## 4 Main Results

In this section we present our main results. We ask how matching affects confidence (Section 4.1), the accuracy of guesses (Section 4.2) and when it leads to affects over- or under-confidence (Section 4.3). Section 4.4 is dedicated to studying the dynamics of confidence over time and in Section 4.5 we discuss potential mechanisms.

### 4.1 Confidence

The left panel of Figure 1 shows participants’ average guessed rank after treatment, i.e. after they observed the score of their match. The figure shows that in all treatments participants are on average somewhat overconfident. The average guessed rank is below the actual mean of 8.5. People seem more confident in their ability in terms of the IQ and FOOT tasks compared to the

---

<sup>5</sup>We chose to pay a lower amount for each correct answer in the IQ task due to higher number of questions in that task. On the other hand, the FOOT task pays a relatively higher amount because these sessions were conducted later and there was a substantial drop in the value of Turkish Lira during the six month period between the first and last session.

<sup>6</sup>If the participant stated a belief of  $p \leq 0.1$  for the subgroup where her rank is, then her earnings were rounded to 0. This was mainly to prevent negative earnings from this stage and was specified in the instructions, as well.

<sup>7</sup>One participant had to be dropped ex post as it was found out that he had participated twice in the experiment.

ART task. It should also be noted, though, that there is a substantial amount of heterogeneity in these guesses as illustrated by Appendix Figure E.2.

	<i>Assortative</i>	<i>Random</i>	<i>Disassortative</i>		<i>Assortative</i>	<i>Random</i>	<i>Disassortative</i>
ART	8.12	8.26	8.90	ART	1.68	1.54	1.23
IQ	6.38	6.74	7.00	IQ	1.76	1.86	1.59
FOOT	7.01	6.62	7.64	FOOT	1.56	1.60	1.54

Gessed Rank. Absolute Error.

Figure 1: Left: Average guessed rank after treatment. Right: Absolute Errors (absolute difference between guessed and actual rank after treatment).

	<i>Gessed Rank</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
	Entire Sample		Rank $\leq 9$		Rank $\geq 8$	
<i>Assortative</i>	-0.183 (0.383)	-0.198 (0.377)	-0.121 (0.433)	-0.122 (0.413)	-0.117 (0.417)	-0.105 (0.430)
<i>Disassortative</i>	0.560 (0.424)	0.560 (0.414)	-0.049 (0.423)	-0.057 (0.428)	1.277** (0.477)	1.289** (0.476)
Constant	10.26*** (1.194)	10.67*** (1.241)	8.632*** (1.587)	9.251*** (1.809)	8.663*** (1.569)	8.667*** (1.558)
Controls	YES	YES <sup>+</sup>	YES	YES <sup>+</sup>	YES	YES <sup>+</sup>
Observations	575	575	287	287	288	288
R-squared	0.086	0.099	0.130	0.152	0.079	0.080

Robust standard errors in parentheses  
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 2: OLS regression of Gessed Rank on treatment dummies. The small set of controls includes age, gender, whether the participant lives with their parents and whether they have siblings. The large set additionally controls for measures of risk aversion and trust, whether they are economics students and linearly for how many friends they have.

To understand how confidence is affected by matching we compare our different treatments. Table 2 shows regression results showing how confidence is affected by the treatment. For the entire sample (columns (1) and (2)) we do not see a statistically significant average treatment effect. While on average confidence tends to increase in the *Assortative* treatments (i.e. the guessed rank decreases) and decrease in the *Disassortative* treatments (i.e. the guessed rank increases), both of these effects are very imprecisely estimated. More interesting is how matching affects confidence differentially for those in the upper and lower half of the distribution. For those with a good (i.e. low) rank (columns (3) and (4)) the effect of matching on their guessed rank is close to zero and statistically insignificant. By contrast those with a high rank (columns (5) and (6)) suffer a substantial decrease in confidence (of more than one position) in *Disassortative* compared to the control condition. Disassortative matching hence seems to have a strong effect on those in bottom half of distribution. Their confidence is, as expected, lowered. The effect persists after receiving one more piece of information about one's rank but disappears after further pieces of information are revealed (Appendix Table D.1).

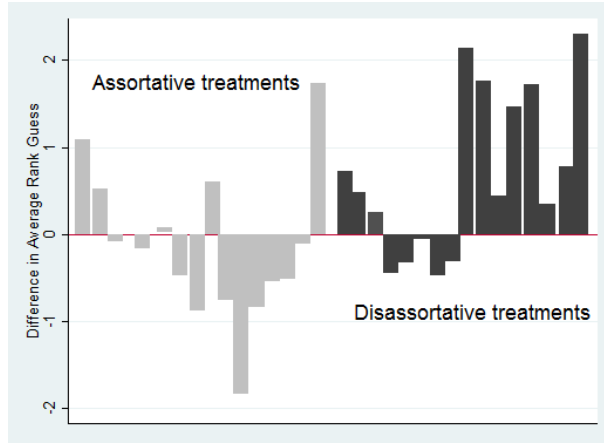


Figure 2: Difference between average rank guess in *Assortative* treatments (light gray bars) and control as well as between *Disassortative* treatments (dark gray bars) and control depending on participant’s rank (1-16). Positive numbers indicate that participants are *less* confident than in the control, i.e. indicate a bigger rank and negative numbers that they are *more* confident, i.e. indicate a smaller rank.

**Heterogeneity** Figure 2 shows this effect across the distribution of ranks. The figure shows that disassortative matching almost never leads to substantial increases in confidence across this distribution, while in *Assortative* confidence tends to increase for those in the bottom half of the distribution. These treatment effects do not differ substantially neither by task nor gender (Appendix Tables D.2 and D.3).

## 4.2 Accuracy

We have seen that worse-performing participants become less confident with disassortative matching, but do they become more accurate? The right panel of Figure 1 shows averages of the variable “absolute error” defined as the absolute difference between the guessed and the actual rank of a participant. The panel shows that participants guesses are fairly accurate on average. In the control group participants guesses differ on average between 1.5 to 1.8 ranks (depending on the task) from their true rank. They seem to make somewhat smaller errors under disassortative matching while there doesn’t seem to be a consistent difference between the control group and the *Assortative* treatment.

Table 3 shows regression results using absolute error as endogenous variable. The table shows that there is virtually no difference between the *Assortative* treatment and the control group neither for the high, nor for the low performers. In the *Disassortative* treatment by contrast participants make substantially smaller errors (by around 10%), i.e. the difference between guessed and actual ranks is smaller in this treatment. The table shows that this effect is driven by those with lower performance (columns (5)-(6)) for whom we observe an about 27% decrease compared to the control group.

**Heterogeneity** Splitting the sample by task reveals that the effects is driven by participants in the ART and IQ tasks (Appendix Table D.6) with the effect size being substantially smaller in the FOOT task (F-test,  $p = 0.067$ ). Worse performing women become more accurate in the *Disassortative* treatment (compared to men), while well performing men, but not women, become less accurate in this treatment (Appendix Table D.5). Neither of these differences is



	<i>Absolute Error</i>					
	(1) Entire Sample	(2)	(3) Rank $\leq 9$		(5)	(6) Rank $\geq 8$
<i>Assortative</i>	-0.020 (0.091)	-0.024 (0.093)	0.015 (0.105)	0.013 (0.095)	-0.039 (0.201)	-0.038 (0.208)
<i>Disassortative</i>	-0.241** (0.098)	-0.238** (0.105)	0.144 (0.097)	0.143 (0.085)	-0.602*** (0.197)	-0.595*** (0.208)
Constant	2.308*** (0.733)	2.367*** (0.728)	1.345** (0.609)	1.158* (0.604)	2.157** (0.904)	2.199** (0.945)
Controls	YES	YES <sup>+</sup>	YES	YES <sup>+</sup>	YES	YES <sup>+</sup>
Observations	575	575	287	287	288	288
R-squared	0.014	0.029	0.033	0.063	0.061	0.068

Robust standard errors in parentheses  
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 3: OLS regression of absolute error (absolute difference between guessed and actual rank) on treatment dummies. The small set of controls includes age, gender, whether the participant lives with their parents and whether they have siblings. The large set additionally controls for measures of risk aversion and trust, whether they are economics students and linearly for how many friends they have.

statistically significant, though (F-test,  $p > 0.197$ ).

We have seen that worse performers do not only become less confident, but also more accurate in the *Disassortative* treatments. This suggests that there was over-confidence in this group to start with. In the next subsection we will study effects on over- and under- confidence explicitly.

### 4.3 Over- or Under- confidence?

To study overconfidence we focus on two outcome variables: (i) the average amount by which participants underestimate their rank and (ii) the share of participants who believe they are ranked better than they actually are.

Table 4 shows results for the first measure. The table shows again that it is mostly the *Disassortative* treatment which has a significant treatment effect. Again the effect seems to operate mostly on the worse performers who, as anticipated, become less overconfident.

	<i>Overconfidence</i>					
	(1) Entire Sample	(2)	(3) Rank $\leq 9$		(5)	(6) Rank $\geq 8$
<i>Assortative</i>	-0.008 (0.182)	-0.013 (0.190)	0.017 (0.212)	0.008 (0.212)	0.036 (0.215)	0.039 (0.223)
<i>Disassortative</i>	-0.345* (0.202)	-0.346 (0.208)	0.008 (0.207)	0.001 (0.209)	-0.676*** (0.244)	-0.672** (0.255)
Constant	1.980** (0.973)	2.057* (1.038)	-0.802 (0.916)	-0.916 (0.956)	2.442*** (0.879)	2.418** (0.937)
Controls	YES	YES <sup>+</sup>	YES	YES <sup>+</sup>	YES	YES <sup>+</sup>
Observations	575	575	287	287	288	288
R-squared	0.015	0.025	0.052	0.064	0.068	0.074

Robust standard errors in parentheses  
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 4: OLS regression of overconfidence (difference between actual and guessed rank) on treatment dummies. The small set of controls includes age, gender, whether the participant lives with their parents and whether they have siblings. The large set additionally controls for measures of risk aversion and trust, whether they are economics students and linearly for how many friends they have.

We then turn our attention to the share of participants who believe they are ranked better

than they actually are, i.e. the share of overconfident participants. Table 5 shows regressions where the endogenous variable is a dummy indicating whether a participant is overconfident. In line with the previous results we find that the share of overconfident participants decreases in the *Disassortative* treatment in the group of the worse performers. With this measure we also see for the first time an effect of the *Assortative* treatment. Among the worse performers the share of overconfident participants increases in this case. The effect size, however, is smaller (about half of the effect size of the *Disassortative* treatment).

	<i>Overconfidence Dummy</i>					
	(1) Entire Sample	(2)	(3) Rank $\leq$ 9	(4)	(5) Rank $\geq$ 8	(6)
<i>Assortative</i>	-0.010 (0.036)	-0.013 (0.037)	-0.059 (0.059)	-0.064 (0.060)	0.060** (0.026)	0.060** (0.027)
<i>Disassortative</i>	-0.048 (0.051)	-0.049 (0.051)	0.024 (0.065)	0.020 (0.065)	-0.109** (0.053)	-0.112** (0.055)
Constant	0.748** (0.308)	0.724** (0.318)	0.173 (0.408)	0.141 (0.416)	0.820*** (0.200)	0.746*** (0.221)
Controls	YES	YES <sup>+</sup>	YES	YES <sup>+</sup>	YES	YES <sup>+</sup>
Observations	575	575	287	287	288	288
R-squared	0.010	0.024	0.020	0.036	0.069	0.080

Robust standard errors in parentheses  
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 5: OLS regression of dummy indicating whether a participant is overconfident on treatment dummies. The small set of controls includes age, gender, whether the participant lives with their parents and whether they have siblings. The large set additionally controls for measures of risk aversion and trust, whether they are economics students and linearly for how many friends they have.

To summarize, matching seems to predominantly affect worse performers. Disassortative matching reduces their confidence and the share of overconfident participants in this group and, as a consequence, makes their perceptions of their own rank more accurate. Assortative matching slightly increases the share of overconfident participants in this group. Whenever we see a treatment effect the direction of effect is in line with what we would expect from rational (Bayesian) learning. What is somewhat puzzling, though, is why it is predominantly the *Disassortative* treatment that affects learning and why assortative matching does not seem to differ much from random matching in terms of its effects on beliefs. Section 4.5 will focus on the mechanisms behind our results to shed some light on these questions. Before we do so, we study in some more detail the dynamics of belief revision across the different steps of our experiment.

#### 4.4 Dynamics

The feedback provided in our experiment during the different steps of the belief revision task is highly diversified across participants, due to differences in their scores, their own ranking in the main task and the scores of other participants in their session. In this section, we introduce a normalized measure of confidence for the different steps of the belief revision task. This will allow us to compare the different steps of the revision process and study its dynamics. We start by noting that during all instances where beliefs are elicited, participants specified beliefs for an even number categories. Based on the construction used in Zimmermann (2019), we define normalized confidence as the difference between the probabilities assigned to the upper and

lower half of the ranks that these categories represent. For example, a participant whose rank is 5, will learn at step 3 of the belief revision task, that her rank is somewhere between 1 to 8. Here, she is also asked to specify her beliefs for being in one of the 4 subgroups (being ranked 1st or 2nd, 3rd or 4th, 5th or 6th and 7th or 8th). In this instance, normalized confidence is measured as probabilities assigned to first two of these categories (being ranked 1st or 2nd or being ranked 3rd or 4th) minus the probabilities assigned to last two categories (being ranked 5th or 6th or being ranked 7th or 8th). By construction normalized confidence takes a value between -100 and 100 when beliefs are represented as percentages. This measure allows us to compare confidence across the different steps of belief revision in our experiment and to study its dynamics.

Next, using a regression analysis we set out to analyze the determinants of this confidence at different stages of the belief elicitation process. Our results are presented in Table 6 where confidence is regressed on a series of dummy variables. The analysis provides us with the following insights. In Step 1 of the belief revision task, the participant observes only her own score, and not surprisingly this score has a substantial and highly significant effect on confidence, as those scoring lower than average ( $\text{Score} < \text{mean} = 1$ ) have lower normalized confidence levels. While not incentivized, participant’s prior on the average score from the task and their prior about their own rank in the task also have significant and persistent effects on confidence. In particular, we observe higher normalized confidence for those who have lower than average expectations for the average score in the associated task ( $\text{Prior for average score} < \text{mean} = 1$ ), and lower confidence for those who have higher than average, i.e. worse, expectations for their ranking ( $\text{Prior for personal rank} > \text{mean} = 1$ ). These effects are also not surprising, since an expectation of a higher average score among participants or a worse personal rank would both mitigate personal confidence.

In Step 2 of the belief revision task, where participants learn the score of another participant, we observe a significant negative effect of *Disassortative* treatment for the participants ranked in the bottom 50% ( $\text{Rank} > 8$ ) of the session, as evidenced by the coefficient for “*Disassortative*  $\times$   $\text{Rank} > 8$ ”. This effect is also significant and has the same sign and similar magnitude at Step 3 of the belief revision task, but not at Steps 4 and 5. On the other hand, the coefficient for “ $\text{Rank} > 8$ ” is positive and significant at Step 3. This observation has an intuitive explanation. At this step, a participant learns whether she is ranked in the bottom half or the upper half of the distribution. Upon learning this news, subjects with moderate beliefs would assign higher weights to categories close to the middle of the overall distribution. This means subjects learning that they are in the lower 50 % would assign a higher belief to being at the 3rd quartile compared to being at the 4th quartile. On the other hand, subjects learning that they are in the upper 50% would assign a higher belief to being at the 2nd quartile compared to being at the 1st quartile. These in turn would imply that normalized confidence at Step 3 would be higher for subjects ranked at the bottom 50%. The nature of tasks also seem to affect normalized confidence. In particular, the coefficients for IQ and FOOT are both positive and significant. During the initial stages of belief elicitation, normalized confidence seems to be highest in the IQ task compared to the FOOT and ART task (0.43 vs. 0.64 and 0.63, respectively). As subjects move to the latter steps, the effect sizes become smaller and the coefficients for IQ task and FOOT become more similar. Overall, hence, these results show again that matching has the strongest effect for

	<i>Normalized Confidence</i>				
	(1) Step 1	(2) Step 2	(3) Step 3	(4) Step 4	(5) Step 5
Score < mean	-46.67*** (5.941)	-36.36*** (7.387)	-34.08*** (6.213)	-2.184 (4.312)	3.838 (5.973)
Prior for average score < mean	15.18*** (5.244)	17.89*** (4.547)	16.62*** (3.485)	13.92*** (3.980)	9.141** (3.581)
Prior for personal rank > mean	-38.61*** (4.364)	-25.55*** (4.116)	-28.77*** (4.690)	-20.03*** (4.207)	-17.85*** (4.251)
IQ	41.66*** (6.073)	38.41*** (5.011)	30.72*** (4.717)	15.53*** (4.580)	9.386*** (3.324)
FOOT	13.97** (6.338)	23.69*** (4.725)	25.61*** (4.071)	12.14*** (4.334)	15.91*** (4.567)
<i>Assortative</i>		-7.187 (6.578)	1.843 (7.537)	-18.61** (6.893)	-5.007 (5.943)
<i>Disassortative</i>		-2.913 (7.191)	7.593 (7.538)	-10.19 (6.922)	-14.43*** (5.120)
Rank > 8		-28.87*** (9.468)	68.01*** (7.078)	14.82 (9.621)	4.009 (9.983)
<i>Assortative</i> × Rank > 8		13.72 (10.72)	-3.507 (11.76)	16.00 (13.63)	11.06 (10.22)
<i>Disassortative</i> × Rank > 8		-24.70** (11.12)	-30.52*** (11.06)	-8.032 (10.57)	10.50 (9.729)
Constant	55.18*** (5.721)	58.21*** (6.600)	0.0482 (6.984)	19.36*** (5.021)	11.72** (5.039)
Observations	575	575	575	575	575
R-squared	0.329	0.414	0.215	0.097	0.079

Robust standard errors in parentheses  
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 6: Normalized Confidence regressed on treatment and other dummies across the five steps of belief revision. Standard errors clustered at the session level.

those in the bottom half of the distribution. They also show that the effect is strongest at Step 2, where the treatment takes place.

## 4.5 Mechanisms

In this section we dig deeper into our data to gain some insight into the behavioural mechanisms underlying these patterns. Figure 3 shows by how much participants' rank guess changes on average (at Step 2) depending on the difference between their score and the score of the participant they observe (their match). The top left panel shows the entire sample. The figure shows that on average participants become more pessimistic about their rank (i.e. increase their guess) if they have a lower score than their match and become more optimistic (i.e. decrease their guess) if they have a higher score than their match. If they have the same score they become more pessimistic. This figure masks a considerable amount of heterogeneity across tasks. In the ART task (top right panel) and the IQ task (bottom left panel) the pattern is as described above. It can also be seen that participants in these tasks react much more strongly to negative feedback (having a lower score than the match) than to positive feedback. If we accept that the IQ task carries more ego-relevance than the ART task, then we can conclude that ego-relevance is not a crucial mechanism behind this, as both tasks show a very similar pattern. The FOOT task (bottom right panel) shows a different pattern, though. Here participants seem to become slightly more confident after feedback was received irrespective of whether feedback was positive or negative. This could be because participants have - on average - a more pessimistic prior in this task. For the remainder of this section we will aggregate the three tasks.

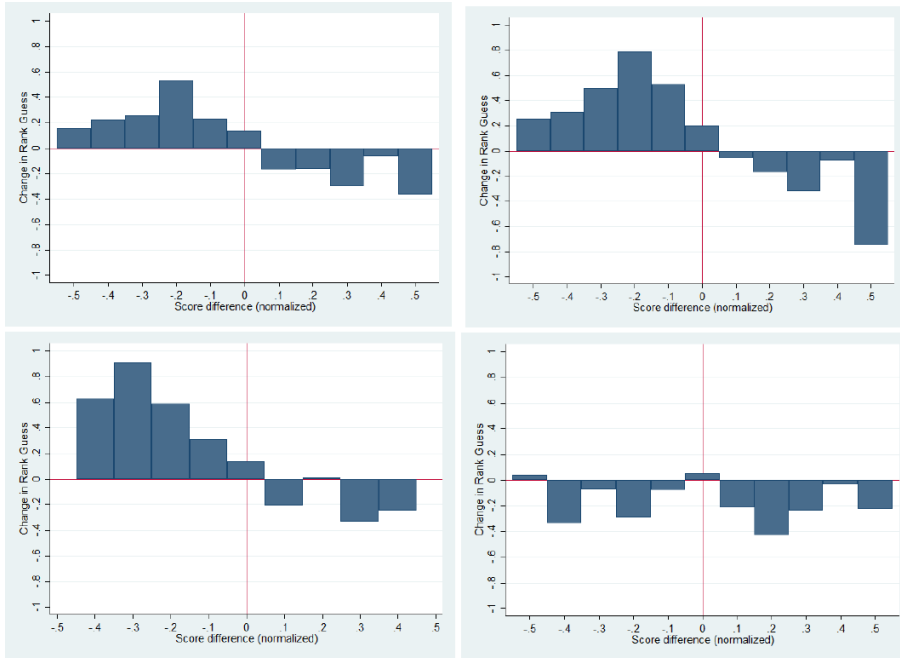


Figure 3: Change in guessed rank depending on difference between own and other participant’s score. Top left: entire sample; Top right: ART task; Bottom left: IQ task; Bottom right: FOOT task.

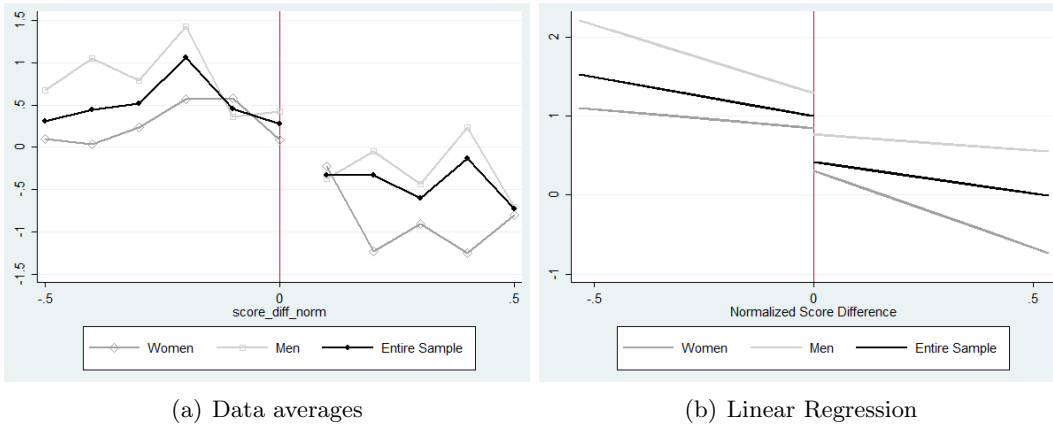


Figure 4: Change in Guessed Rank depending on the difference between own and other participant’s score and on whether that difference is negative or positive. Gray line with diamond markers show women, square markers men and the black line the entire sample. The left panel shows raw data averages and the right panel predicted values from the regression in Table 7 including controls for age, gender, task, siblings and housing situation.

Table 7 shows regression results where we regress the change in confidence on the observed score difference and an indicator for whether the difference is positive. The table shows that the indicator matters even when score difference is included in the regression (columns (3)-(4)). There is a discontinuous jump in participant’s reaction to feedback once it changes sign. Figure 4 illustrates the regression results (Panel (b)) and also shows the raw data averages (Panel (a)). The figure illustrates that confidence changes not as much for positive feedback as it does for negative feedback.<sup>8</sup> It also illustrates the discontinuous change at zero.

As there is a large psychological literature on gender differences in reacting to feedback and

<sup>8</sup>Note that there is a level shift between panels (a) and (b) which is due to the inclusion of controls in Panel (b).

	(1)	(2)	(3)	(4)
	<i>Change in Guessed Rank</i>			
$\Delta_{Score}$	-1.907*** (0.400)		-0.919 (0.568)	-0.990 (0.735)
pos		-0.866*** (0.144)	-0.572** (0.210)	-0.587*** (0.214)
$\Delta_{Score} \times pos$				0.191 (1.047)
Constant	1.007 (0.944)	1.104 (0.902)	1.020 (0.926)	1.007 (0.945)
Observations	576	576	576	576
R-squared	0.074	0.080	0.085	0.085

Robust standard errors in parentheses  
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 7: Change in Guessed Rank regressed on depending on difference between own and other participant’s score ( $\Delta_{Score}$ ), a dummy indicating whether this difference is positive (pos) and the interaction between the two. Linear controls included for age and number of siblings as well as gender dummy, indicator for housing situation and task fixed effects.

attribution, which shows that (i) women tend to react more strongly to feedback (for a review see Roberts, 1991)<sup>9</sup> and (ii) that women are more likely to blame their ability for their failures (see for instance Dweck, Davidson, Nelson, & Enna., 1978; Dweck & Reppucci, 1973; Nicholls, 1975) we also split the figures by gender. The gray lines in Figure 4 show the change in confidence after feedback separately for women and men. In our experiment, women react more strongly to feedback, but only if it is positive feedback. The effect is just outside of statistical significance, though ( $p = 0.1016$ ). There is no (statistically significant) gender difference in the reaction to negative feedback.

Overall these results show that the direction of feedback (positive vs negative) seems more important than the extent, i.e. how different the score of the other person is from one’s own score. They also show that people react more when feedback is negative. The latter fact can also explain why disassortative matching has a stronger treatment effect. While both assortative and disassortative matching have an equal share of participants exposed to positive and negative feedback, they differ in how strongly negative or positive the feedback is.<sup>10</sup> The fact that participants - on average - react more strongly to negative feedback means that treatment effects will be stronger under disassortative matching.

## 5 Conclusion

We conducted a lab experiment to study how matching affects confidence. Across a variety of tasks we find that assortative matching does not have a substantial nor statistically significant effect on confidence compared to a control group with random matching. By contrast, disassor-

<sup>9</sup>Interestingly, there are also experimental papers which find that women tend to update their beliefs less strongly than men in response to feedback (Albrecht, Von Essen, Parys, & Szech, 2013; Buser, Gerhards, & van der Weele., 2018; Coutts, 2019; Mobius, 2011). Some authors point to gender differences which depend on the valence - good or bad - of the feedback received. Ertac (2011) finds that women that completed a verbal task interpret positive feedback more conservatively than men while no gender difference is found for negative feedback. Berlin and Dargnies (2016) show that women update more pessimistically than their male counterparts after receiving negative feedback but not after positive feedback.

<sup>10</sup>Under assortative matching many participants also see the same score as their own, i.e. receive “neutral” or “weakly positive” feedback.

tative matching has a negative effect on confidence on average that is driven by the bottom half of performers. This group becomes substantially less confident compared to random matching. However they become also more accurate, i.e. less overconfident, with disassortative compared to random matching.

These are important findings that should be taken into account when designing policies like tracking in schools or matching peers at work. There are, however, also several caveats and open questions for future research. One important question regards the trade-off between overconfidence and accuracy. It seems obvious that there are advantages to holding accurate beliefs, however some also argue that there are benefits from being overconfident (Anderson et al., 2012; Kahneman & Lovallo, 1993; Murphy et al., 2015; Radzevick & Moore, 2011). Hence, which level of (over-) confidence to target is not immediately obvious. Furthermore we should remember that confidence is only one outcome affected by matching and obviously the effect on other outcomes has to be taken into account.

## References

- Alan, S., & Ertac, S. (2019). Mitigating the gender gap in willingness to compete. *Journal of the European Economic Association*, *in press*.
- Alan, S., Ertac, S., Kubilay, E., & Loranth, G. (2019). Understanding gender differences in leadership. *Economic Journal*, *in press*.
- Albrecht, K., Von Essen, E., Parys, J., & Szech, N. (2013). Updating, self-confidence, and discrimination. . *European Economic Review*, *60*, 144-169.
- Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*(4), 718-735..
- Antonio, A. (2004). The influence of friendship groups on intellectual self-confidence and educational aspirations in college. *The Journal of Higher Education*, *75*(4), 446-471.
- Bandiera, O., Barankay, I., & Rasul, I. (2010). Social incentives in the workplace. *Review of Economic Studies*, *77*, 417-458.
- Berlin, N., & Dargnies, M. P. (2016). Gender differences in reactions to feedback and willingness to compete. . *Journal of Economic Behavior & Organization*, *130*, 320-336.
- Blanden, J., Gregg, P., & Macmillan, L. (2007). Accounting for intergenerational income persistence: Noncognitive skills, ability and education. *Economic Journal*, *117*, C43-C60.
- Branas-Garza, P., Cobo-Reyes, R., Espinosa, M., Kovarik, J., Jimenez, N., & Ponti, G. (2010). Altruism and social integration. *Games and Economic Behavior*, *69*(2), 249-257.
- Buser, T., Gerhards, L., & van der Weele., J. (2018). “responsiveness to feedback as a personal trait.”. *Journal of Risk and Uncertainty* *56* (2): 165-92..
- Coutts, A. (2019). Good news and bad news are still news: experimental evidence on belief updating. *Experimental Economics*, *22*(2), 369-395.
- Currarini, S., & Mengel, F. (2016). Identity, homophily and in-group bias. *European Economic Review*, *90*, 40-55.
- Dweck, C. S., Davidson, W., Nelson, S., & Enna., B. (1978). “sex differences in learned helplessness: ii. the contingencies of evaluative feedback in the classroom and iii. an experimental analysis.” . *Developmental Psychology* *14* (3): 268-76.
- Dweck, C. S., & Reppucci, N. D. (1973). Learned helplessness and reinforcement responsibility in children. *Journal of Personality and Social Psychology*, *25*,109-116.
- Eil, D., & Rao, J. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114-138.
- Ertac, S. (2011). Does self-relevance affect information processing? experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior and Organization*, *80*(3), 532-545.
- Feld, J., & Zoelitz, U. (2017). Understanding peer effects - on the nature, estimation and channels of peer effects. *Journal of Labor Economics*, *35*(2).
- Fershtman, C., & Gneezy, U. (2001). Discrimination in a segmented society: An experimental approach. *The Quarterly Journal of Economics*, *116*(1), 351-377.

- Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171-178.
- Golsteyn, B., Non, A., & Zoelitz, U. (2019). The impact of peer personality on academic achievement. *mimeo*.
- Greiner, B. (2004). An online recruitment system for economic experiments. In K. Kremer & V. Macho (Eds.), *Forschung und wissenschaftliches rechnen*. GDWG Bericht 63.
- Grimm, V., & Mengel, F. (2009). Cooperation in viscous populations - experimental evidence. *Games and Economic Behavior*, 66(1), 202-220.
- Hanushek, E. A., & Woessmann, L. (2006). Does educational tracking affect performance and inequality? differences-in-differences evidence across countries. *The Economic Journal*, C63-C76.
- Heckman, J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3), 411-482.
- Jackson, C., & Bruegmann, E. (2009). Teaching students and teaching each other: the importance of peer learning for teachers. *American Economic Journal: Applied Economics*, 1(4), 85-108.
- Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk and risk taking. *Management Science*, 39, 17-31.
- Koeszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4), 673-707.
- Mas, A., & Moretti, E. (2009). Peers at work. *American Economic Review*, 99(1), 112-145.
- Mobius, M. (2011). Managing self-confidence: theory and experimental evidence. *NBER working paper 17014*.
- Mobius, M., Niederle, M., Niehaus, P., & Rosenblat, T. (2007). Gender differences in incorporating performance feedback. *Mimeo, Harvard University*.
- Murphy, S. C., von Hippel, W., Dubbs, S. L., Angilleta Jr, M. J., Wilson, R. S., Trivers, R., & Barlow, F. K. (2015). The role of overconfidence in romantic desirability and competition. *Personality and Social Psychology Bulletin*.
- Nicholls, J. G. (1975). Causal attributions and other achievement-related cognitions: Effects of task outcome, attainment value and sex. *Journal of Personality and Social Psychology*, 31, 379-389.
- Radzevick, J. R., & Moore, D. (2011). Competing to be certain (but wrong): Social pressure and overprecision in judgment. *Management Science*, 57(1), 93-106.
- Roberts, T. A. (1991). Gender and the influence of evaluations on self-assessments in achievement settings. *Psychological bulletin*, 109(2), 297.
- Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics*, 116(2), 681-704.
- Sacerdote, B. (2014). Experimental and quasi-experimental analysis of peer effects: Two steps forward? *Annual Review of Economics*, 6(1), 253-272.
- Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature neuroscience*, 14(11), 1475.
- Ytterberg, S., Harris, I., Allen, S., Anderson, D., Kofron, P., Kvasnicka, J., ... Moller, J. (1998). Clinical confidence and skills of medical students: use of an osce to enhance confidence in clinical skills. *Academic Medicine*, 73(10), S103-105.
- Zimmermann, F. (2019). The dynamics of motivated beliefs. *American Economic Review*, 110(2), 337-361.



# Online Appendix for “Matching and Confidence”

M.Y. Gurdal, E. Leni, F. Mengel

For Online Publication

## Contents

<b>A Appendix: Sample</b>	<b>2</b>
<b>B Appendix: Instructions</b>	<b>3</b>
<b>C Appendix: Questionnaire</b>	<b>8</b>
<b>D Appendix: Additional Tables</b>	<b>9</b>
<b>E Appendix: Additional Figures</b>	<b>12</b>

## A Appendix: Sample

We perform a balancing check for our treatments. We regress the observables collected in the post-experimental questionnaire on treatment dummies. The systematic presence of significant coefficients in the regression would reveal differences in the pool of participants of *Assortative* and *Disassortative* treatments compared to the random treatment.

We consider the following dependent variables: in column (1) age, in (2) gender (male = 1), in (3) living arrangement of subject (0= dorm; 1= with family; 2= with friends; 3= alone), in (4) number of siblings, (5) risk attitude measured by the question “How willing are you to take risks in general?” (10=high; 1=low), (6) trust measured by the question “Would you say that most people can be trusted?” (1=yes), (7) number of economics classes taken, (8) number of friends among participants in the session.

Table A.1 presents the results of the balancing check. Only two coefficients are significant - *Assortative* when regressed on age and *Disassortative* when regressed on siblings - indicating that participants do not systematically differ across treatments. The data support the hypothesis of random assignment to treatments.

Dependent variable:	(1) Age	(2) Male	(3) Living	(4) Siblings	(5) Risk	(6) Trust	(7) Economics	(8) Friends
<i>Assortative</i>	-0.474*** (0.159)	-0.062 (0.050)	0.052 (0.010)	-0.042 (0.098)	-0.125 (0.195)	-0.005 (0.037)	-0.130 (0.149)	0.083 (0.090)
<i>Disassortative</i>	-0.252 (0.160)	-0.033 (0.050)	-0.021 (0.010)	0.180* (0.098)	-0.034 (0.195)	0.006 (0.037)	-0.046 (0.150)	0.108 (0.090)
Observations	575	575	575	575	575	575	575	575
R-squared	0.015	0.003	0.001	0.010	0.001	0.000	0.001	0.003

Standard errors in parentheses  
 \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table A.1: Balancing check. Regression of observables on treatment dummies.

## B Appendix: Instructions

All participants in a treatment received the same instructions. We handed out written instructions at the beginning of each session. In B.0.1 and B.0.2, we report written instructions and screenshots for the ART task. For the FOOT and IQ task, the instructions were modified according to the specificity of each task. We provide further details on the FOOT task in B.0.3.

The original instructions were in Turkish. Here, we provide the English translation.

### B.0.1 Written instructions - ART task

#### General Information

Welcome! You are about to participate in an experiment on decision-making. If you follow the instructions carefully, you can earn a significant amount of money based on your choices.

This instruction set is for your private use only. You cannot communicate with anyone during the experiment. If you have any questions, please raise your hand. Then we'll come and answer your questions. Violation of this rule requires that we immediately exclude you from the experiment.

With the decisions you will make in the experiment, you will earn a profit. Below, we will explain the details of this. All your decisions will be handled confidentially, both during and after the experiment. This means that none of the other participants will know the decisions you make.

#### Part 1:

In this first part of the experiment we ask you to answer the questions in a test. To perform the test, we will show paintings sold in an auction in February 2016. In addition to the paintings, we present the name of the painting and the painter. You will see something similar to the following images:



Pablo Picasso, Mandoline



Egon Schiele, Oesterreichisches Mädel

We then ask you to specify which one of these two paintings was sold at a higher price in this auction. To complete the task, you are asked to compare 15 different pairs of paintings. Your test score is calculated as the total number of correct answers in 15 pairs.

The maximum score you can get is 15 (if all of your answers are correct). The minimum score is 0 (if all of your answers are incorrect). You will earn 1 TL for each correct answer in this section.

**Part 2:**

We will rank the participants in this room based on your test scores in the previous section. If two participants get the same score, their rankings will be random. In this part of the experiment you are asked to answer some questions about the performance of the other participants in the experiment and the comparison of your performance vs. others performances.

In this section, we'll ask you questions about your beliefs regarding your ranking in the test. We'll ask you to do this a total of 5 times, and give you new information every time. As you know, there are 16 people in the experiment, and your rankings range from 1st to 16th.

Each time, you will be asked to guess how well you think you did the test compared to the rest of the students in the lab. You will do this by specifying your probability estimates for the ranking groups you see on the screen. We ask you to enter numbers from 0 to 100 for the probabilities for each ranking group. Please note that the odds allocated to the groups you see on the screen should always add up to 100%. For example:

Your rank	1. or 2.	3. or 4.	5. or 6.	7. or 8.	9. or 10.	11. or 12.	13. or 14.	15. or 16.	SUM
Probability	5%	10%	40%	10%	5%	0%	0%	0%	100%

If the total of the odds is not 100%, the computer will inform you and you will need to correct your answer before proceeding to the next stage.

**Payment**

As we explained above, we will ask you to specify the possibilities for your ranking 5 times in total. One of the 5 predictions you make will be chosen randomly and we will pay you based on the accuracy of that prediction.

Payment is made according to the formula  $32 + 32 \log(p)$ . Here,  $p$  is the number we find by dividing the probability that you assign to the category that contains your actual ranking. Note that  $\log(p)$  is a zero or negative number, and your payment will be between 0 and 32, depending on how accurate the prediction is.

**Example:** For example, if you scored the best score in the test, that is, if your ranking is 1 and if you correctly guess the probability of being 1st or 2nd as 100%, you earn the highest possible amount. Your payment is 32 TL (calculation:  $32 + 32 \times \log(1) = 32$ ).

Your rank	1. or 2.	3. or 4.	5. or 6.	7. or 8.	9. or 10.	11. or 12.	13. or 14.	15. or 16.	SUM
Probability	100%	0%	0%	0%	0%	0%	0%	0%	100%

However, if your ranking is 1, and you assign a probability lower than 100% for this category, your earnings will be reduced. This reduction is proportional to the inaccuracy of the probability assigned to the correct group.

For example, if your ranking is 1 and you set the probability of being ‘1st or 2nd’ to 90%, your payment will be reduced by 5% and will be  $32 + 32 \times \log(1.01) = 32 - 32 \times 0.05 = 30.54$  TL.

Your rank	1. or 2.	3. or 4.	5. or 6.	7. or 8.	9. or 10.	11. or 12.	13. or 14.	15. or 16.	SUM
Probability	90%	10%	0%	0%	0%	0%	0%	0%	100%

If your ranking is 1 and you set the probability of being ‘1st or 2nd’ to 80%, your payment will be reduced by 9.7% and will be  $32 + 32 \times \log(0.8) = 32 - 32 \times 0.097 = 28.89$  TL.

If your ranking is 1 and you set the probability of being ‘1st or 2nd’ to 70%, your payment will be reduced by 15.5% and will be  $32 + 32 \times \log(0.7) = 32 - 32 \times 0.155 = 27.04$  TL.

If your ranking is 1 and you set the probability of being ‘1st or 2nd’ to 60%, your payment will be reduced by 22.2% and will be  $32 + 32 \times \log(0.6) = 32 - 32 \times 0.222 = 24.9$  TL.

If your ranking is 1 and you set the probability of being ‘1st or 2nd’ to 50%, your payment will be reduced by 30.1% and will be  $32 + 32 \times \log(0.5) = 32 - 32 \times 0.301 = 22.37$  TL.

If your ranking is 1 and you set the probability of being ‘1st or 2nd’ to 40%, your payment will be reduced by 39.8% and will be  $32 + 32 \times \log(0.4) = 32 - 32 \times 0.398 = 19.27$  TL.

If your ranking is 1 and you set the probability of being ‘1st or 2nd’ to 30%, your payment will be reduced by 52.3% and will be  $32 + 32 \times \log(0.3) = 32 - 32 \times 0.523 = 15.27$  TL.

If your ranking is 1 and you set the probability of being ‘1st or 2nd’ to 20%, your payment will be reduced by 69.9% and will be  $32 + 32 \times \log(0.2) = 32 - 32 \times 0.699 = 15.27$  TL.

Finally, if your ranking is 1, if you set the probability of being 1 or 2 as 10% or less, your payment will be reduced to 0.

You should try to be as accurate as possible in your predictions. A good estimate will give you the best benefit.

At the end of the experiment, all your winnings will be paid in cash.

## B.0.2 Screenshots - ART task



Before you start the experiment, we would like to learn some of your thoughts about the test.

We realize that you probably have not performed a similar task before. Try to answer as accurately and honestly as possible.

Consider the group of 16 students who are currently performing the test in this room. What do you think will be the average score (minimum score 0 - maximum score 15) of the test?

How well do you think you can perform the test compared to the rest of the students? So, what do you think will be your ranking (1, 2, ..., 15, 16)?

OK

	
Otto Dix, Schwangerschaft	Karl Schmidt-Rottluff, Windsiger Tag
Which of the two pictures has been sold at a higher price? <input type="radio"/> Left <input type="radio"/> Right	
OK	

## B.0.3 FOOT task

In the FOOT task, we ask participants to compare 15 pairs of players for whom we provide names, team, and a close-up picture. Table B.1 reports the pairs of players that participants face during the experiment. The team refers to the Turkish League 2017/2018.

Taliska (Besiktas JK)	Serdar Aziz (Galatasaray SK)
Garry Rodrigues (Galatasaray SK)	Adriano Correia Claro (Besiktas JK)
Mathieu Valbuena (Fenerbahce SK)	Aziz Behich (Burnaspor)
Pepe (Besiktas JK)	Sofiane Feghouli (Galatasaray SK)
Mustafa Yumlu (Akhisarspor)	Maicon Pereira Roque (Galatasaray SK)
Hasan Ali Kaldirim (Fenerbahce SK)	Bafétimbi Gomis (Galatasaray SK)
Bogdan Stancu (Burnaspor)	Juraj Kucka (Trabzonspor)
Pablo Batalla (Burnaspor)	Deniz Kadah (Antalyanspor)
Dusko Tasic (Besiktas JK)	Ricardo Quaresma (Besiktas JK)
André Castro (Göztepe)	Serginho (Akhisarspor)
Younès Belhanda (Galatasaray SK)	Emre Akbaba (Alanyaspor)
Titi (Burnaspor)	Burak Yilmaz (Trabzonspor)
Martin Skrtel (Fenerbahce SK)	Mehmet Topal (Fenerbahce SK)
Ryan Babel (Besiktas JK)	Josef de Souza Dias (Fenerbahce SK)
Roman Neustädter (Fenerbahce SK)	Emmanuel Adebayor (Basaksehir)

Table B.1: Pairs of players in FOOT test

## C Appendix: Questionnaire

We list here the different variables elicited in our post-experimental questionnaire.

- Age of the subject (in years).
- Sex of the subject (1=male, 0=female).
- Living: living arrangement for the subject (0=student housing, 1=with family, 2= with friends, 3=alone).
- Siblings: number of siblings of subject.
- Older siblings: number of siblings who are older than the subject.
- Trust: “Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?” (Be careful 0 ... 10 can be trusted)
- Risk: “How willing are you to take risks in general?” (0 lowest – 10 highest)
- Major: subject’s major (2=economics, 1=business, political science or international trade, 0=other).
- Econ: number of economics classes (censored at 4).
- Friends: number of people known in the session
- Rely: “How much can we trust the data coming from you in this experiment?” (0 lowest – 10 highest)



## D Appendix: Additional Tables

	<i>Persistence of Guessed Rank</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
	Step 3		Step 4		Step 5	
<i>Assortative</i>	-0.091 (0.232)	-0.058 (0.164)	0.094 (0.121)	-0.000 (0.0931)	-0.123 (0.171)	-0.157 (0.132)
<i>Disassortative</i>	-0.089 (0.227)	0.579*** (0.193)	0.062 (0.102)	0.149 (0.0937)	0.080 (0.118)	-0.125 (0.142)
Constant	4.708*** (0.930)	12.25*** (0.852)	6.240*** (1.594)	13.31*** (1.213)	12.60*** (3.559)	26.10*** (2.637)
Controls	YES	YES <sup>+</sup>	YES	YES <sup>+</sup>	YES	YES <sup>+</sup>
Observations	287	288	287	288	287	288
R-squared	0.068	0.136	0.054	0.006	0.028	0.008

Robust standard errors in parentheses  
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table D.1: OLS regression of confidence (guessed rank) on treatment dummies across the various steps of the belief revision process. Bottom half of performers only. The small set of controls includes age, gender, whether the participant lives with their parents and whether they have siblings. The large set additionally controls for measures of risk aversion and trust, whether they are economics students and linearly for how many friends they have.

	<i>Guessed Rank - split by gender</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
	All Ranks		Rank ≤ 9		Rank ≥ 8	
	F	M	F	M	F	M
<i>Assortative</i>	0.006 (0.536)	-0.290 (0.582)	0.177 (0.627)	-0.277 (0.576)	0.007 (0.578)	-0.265 (0.636)
<i>Disassortative</i>	0.808 (0.493)	0.444 (0.632)	-0.180 (0.703)	0.054 (0.534)	1.378*** (0.493)	1.208 (0.723)
Constant	10.34*** (1.993)	8.678*** (1.584)	11.07*** (3.166)	5.769** (2.150)	9.995*** (3.056)	7.320*** (1.991)
Controls	YES	YES	YES	YES	YES	YES
Observations	231	344	97	190	134	154
R-squared	0.029	0.012	0.028	0.006	0.076	0.057

Robust standard errors in parentheses  
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table D.2: OLS regression of confidence (guessed rank) on treatment dummies split by gender. The small set of controls includes age, whether the participant lives with their parents and whether they have siblings.

	<i>Guessed Rank - split by task</i>			
	(1) all	(2) ART	(3) IQ	(4) FOOT
<i>Assortative</i>	-0.183 (0.383)	-0.121 (0.361)	0.155 (0.578)	0.155 (0.578)
<i>Disassortative</i>	0.560 (0.424)	0.550 (0.481)	0.405 (0.763)	0.405 (0.763)
Constant	10.26*** (1.194)	9.902*** (1.620)	8.976*** (1.836)	8.976*** (1.836)
Controls	YES	YES	YES	YES
Observations	575	191	192	192
R-squared	0.086	0.037	0.330	0.330

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table D.3: OLS regression of confidence (guessed rank) on treatment dummies split by task. The small set of controls includes age, gender, whether the participant lives with their parents and whether they have siblings.

	<i>Change in Confidence</i>					
	(1) Entire Sample	(2)	(3) Rank ≤ 9	(4)	(5) Rank ≥ 8	(6)
<i>Assortative</i>	0.265 (0.255)	0.261 (0.257)	0.428** (0.178)	0.421** (0.174)	0.126 (0.392)	0.124 (0.389)
<i>Disassortative</i>	0.0720 (0.267)	0.0743 (0.268)	-0.173 (0.210)	-0.169 (0.210)	0.329 (0.416)	0.365 (0.423)
Constant	1.463 (0.885)	1.530* (0.849)	1.324 (0.842)	1.061 (0.929)	0.802 (1.630)	1.320 (1.497)
Controls	YES	YES+	YES	YES+	YES	YES+
Observations	575	575	287	287	288	288
R-squared	0.014	0.017	0.063	0.080	0.014	0.023

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table D.4: OLS regression of the Change in Confidence on treatment dummies. The small set of controls includes age, gender, whether the participant lives with their parents and whether they have siblings. The large set additionally controls for measures of risk aversion and trust, whether they are economics students and linearly for how many friends they have.

	<i>Absolute Error - split by gender</i>					
	(1) All Ranks		(3) Rank ≤ 9		(5) Rank ≥ 8	
	F	M	F	M	F	M
<i>Assortative</i>	-0.052 (0.194)	-0.001 (0.133)	0.158 (0.226)	-0.051 (0.113)	-0.141 (0.307)	0.061 (0.249)
<i>Disassortative</i>	-0.430* (0.212)	-0.134 (0.141)	0.053 (0.202)	0.205* (0.114)	-0.756*** (0.271)	-0.479* (0.279)
Constant	1.142 (1.080)	2.630*** (0.891)	0.514 (1.400)	1.247* (0.710)	1.805 (1.482)	2.556** (1.046)
Observations	231	345	97	191	134	154
R-squared	0.039	0.007	0.025	0.028	0.077	0.035

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table D.5: OLS regression of Absolute Error on treatment dummies split by gender. The small set of controls includes age, whether the participant lives with their parents and whether they have siblings.

	<i>Absolute Error - split by task</i>			
	(1)	(2)	(3)	(4)
	all	ART	IQ	FOOT
<i>Assortative</i>	-0.014 (0.092)	0.125 (0.110)	-0.152 (0.149)	-0.079 (0.174)
<i>Disassortative</i>	-0.245** (0.101)	-0.359*** (0.037)	-0.325 (0.185)	-0.080 (0.138)
Constant	2.299*** (0.732)	2.184** (0.743)	2.923* (1.475)	2.427* (1.329)
Observations	576	192	192	192
R-squared	0.013	0.036	0.024	0.005

Robust standard errors in parentheses  
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table D.6: OLS regression of Absolute Error on treatment dummies split by task. The small set of controls includes age, gender, whether the participant lives with their parents and whether they have siblings.

## E Appendix: Additional Figures

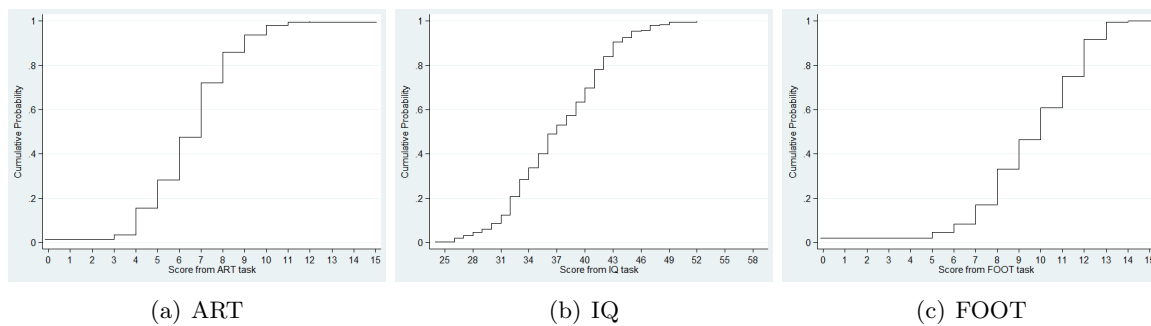


Figure E.1: Score distributions for the three tasks.

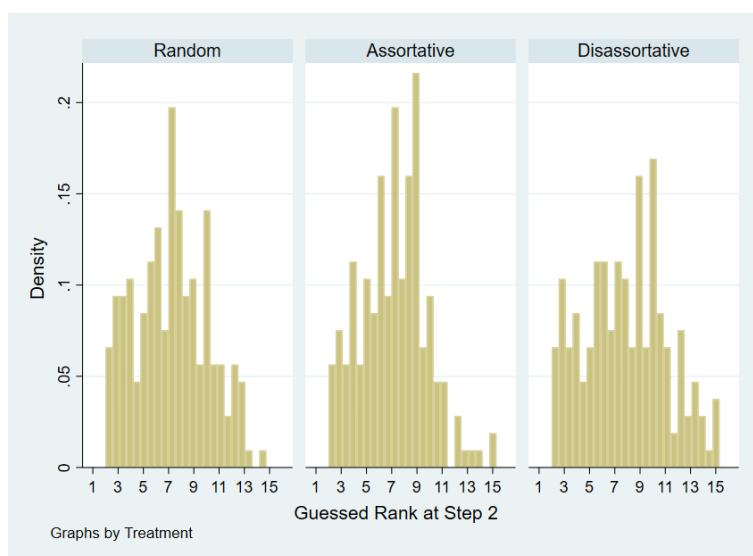


Figure E.2: Histogram of Average Gussed Rank at Step 2 of the Belief Revision Process.