Ma, Mingye; Riener, Gerhard; Xu, Youzong

Working Paper

Evaluating Yourself and Your Peers

IZA Discussion Papers, No. 17267

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Ma, Mingye; Riener, Gerhard; Xu, Youzong (2024) : Evaluating Yourself and Your Peers, IZA Discussion Papers, No. 17267, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/305709

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17267

Evaluating Yourself and Your Peers

Mingye Ma
Gerhard Riener
Youzong Xu

SEPTEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17267

Evaluating Yourself and Your Peers

Mingye Ma
University of Southampton

Gerhard Riener
University of Southampton and IZA

Youzong Xu
University of Nottingham Ningbo China

SEPTEMBER 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Evaluating Yourself and Your Peers

We explore the role of self- and peer evaluations in education, with a particular emphasis on gender differences. We construct a model of (self-)deception to predict outcomes for scenarios with and without self-evaluation. By using unique data from a first-year economics class at a Sino-UK university, we examine how students assess their own and their peers' contributions to group projects under varying self-assessment conditions. Our findings reveal a significant self-serving bias across both genders, though with subtle distinctions. Women, despite greater societal recognition, exhibit smaller self-social evaluation gaps (SSEG). The variation in abstention rates between treatments is mainly attributed to lowerperforming males. These findings indicate that the possibility of self-assessment influences rating behavior, potentially exacerbating gender disparities and affecting gender equity.

JEL Classification: D01, D91, I23, C93

Keywords: higher education, incentives, field experiment, peer evaluation, gender

Corresponding author:
Gerhard Riener
Department of Economics
University of Southampton
University Road
Southampton SO17 1BJ
Great Britain

E-mail: gerhard.riener@gmail.com

Nemo judex in causa sua.

Corpus Juris Civilis

I. Introduction

In complex societies, much of the production and learning is done collaboratively within teams. Consequently, educational tasks that promote teamwork are a vital tool for success in a society; therefore, group work has been integrated into higher education and is high on the agenda of employability and skills (Winterbotham et al., 2018). However, the assessment of group work presents multiple challenges, particularly in educational environments where collaborative efforts are typically short-lived. A prevalent method for evaluating such group tasks involves a combination of self- and/or peer-assessments over individual contributions (360-degree feedback, DeNisi and Kluger (2000)). Although widely used, peers lack the professional grading skills that teachers possess, and there are strategic motives that bias the evaluation and hence the grade, as these simple and common evaluation mechanisms are not incentive-compatible (Edwards and Ewen, 1996; Coates, 1998). Therefore, informative assessments rely on honest reporting, and previous research has shown that honesty differs between particular subgroups of society (Dreber and Johannesson, 2008; Houser, Vetter and Winter, 2012). This then prompts an essential question: How do personal traits such as gender relate to these assessments, and what are the effects when methods include self-evaluation alongside peer evaluation?

n a theoretical framework featuring students who are solely motivated by maximizing their grades and have the opportunity for self-evaluation, our findings indicate that complete collusion emerges as an equilibrium. In this scenario, all group members unanimously assign themselves the highest possible grade. This outcome extends the classical principle that "no one should be a judge in their own case" (Nemo judex in causa sua) to a group setting, where the potential oversight by other group members is insufficient to disrupt the collusive equilibrium. In addition, when self-evaluation is prohibited, grade-maximizing students are incentivized to rate their peers as low as possible to secure their own relative ranking. Consequently, in a setting where students act purely out of self-interest, these ratings do not provide any meaningful insight into the abilities or contributions of group members.

However, based on evidence that humans have some intrinsic (positive) cost of lying to others (Abeler, Nosenzo and Raymond, 2019) and potentially to themselves, we develop a simple model with lying costs and the potential of moral flexibility to show when self-evaluation leads to partially truth-revealing outcomes, when the costs of lying and self-deception are sufficiently high. Moreover, when self-deception is present, in Condition No Self, where subjects cannot rate themselves, students have incentives to assign lower ratings to all recipients com-

pared to Condition Self, where self-rating is possible. We then simulate our model, varying the cost of lying, to observe how the distribution of honest, (standard) partial lie, (standard) maximal lie, and self-deception changes.

We evaluated these predictions using data from student group assessments that varied the availability of self-rating over three academic years. We have a sample of 2,386 students from a first-year economics course at a Sino-UK university in China with approximately 800 primarily business school students annually. During three academic years, from 2016/17 to 2018/19, we implemented and adjusted a peer review system to evaluate group projects centered on economic data analysis. This system required students to rate their peers' contributions to the group project and affected their individual grades through a "contribution parameter". This parameter could be assigned solely to its group peers (Condition No Self) or also to themselves (Condition Self), depending on the treatment condition, and contributed to the final grade of the course. In addition, the course included individual exams that aim to evaluate overall understanding and provide individual performance metrics.

In situations where individuals have the opportunity to rate themselves, there is a tendency for them to give themselves higher ratings than their peers, illustrating the presence of self-serving bias and a noticeable gap between self-evaluation and social evaluation, as evidenced by a test of first-order stochastic dominance (Deb and Renou, 2022). Furthermore, we observe a clustering effect at the highest possible rating point, indicative of collusion, along with a significant incidence of abstention. In particular, in the condition Condition Self, women were more likely to assign themselves lower ratings while receiving higher ratings from their peers. However, in the Condition No Self scenario, women did not receive more favorable ratings. We suggest that this discrepancy may partly be due to differences in abstention rates between the two conditions.

In addition, we find a difference in abstention between the rating conditions that is primarily due to male students with weaker academic performance. A plausible explanation for this behavior is that these students may place less importance on their academic performance.

We contribute to several strands of the literature. First, we expand the recent literature on incentives in group performance and evaluation in non-routine tasks and incentive provision (Prendergast, 1999). Closely related are Ramm, Tjotta and Torsvik (2013), Englmaier et al. (2024), who examine the role of monetary incentives in nonroutine group tasks, and the recent contribution by Morgan, Neckermann and Sisak (2021), who explicitly look at the effect of peer-evaluations without self-evaluation, on performance. and found that overall group performance was not affected by peer evaluation. Participants reported higher motivation, worked longer, and communicated more. Moreover, they found a shift in behavior towards impressing team members and higher work effort leading to more timeouts and incomplete solutions.

Bohl (1996) Second, we extend theoretical considerations on lying aversion and

moral wiggle room on the aspect of lying over oneself or others

Third, we extend the literature on the self-social evaluation gap by adding a strategic component to the self-enhancement bias Sedikides and Gregg (2008); Krueger, Heck and Asendorpf (2017), using social consensus as a reference value (Larrick, Mannes and Soll, 2024; Vazire and Carlson, 2011).

In addition, we contribute to the literature on gender differences in evaluation and promotion. The closest to our paper are the experiments by Exley and Kessler (2022) who find a significant gender gap in self-evaluations on math and science tasks, with women rating their performance less favorably than equally performing men, indicating a gap in self-promotion. This gender gap persists even in private settings without promotional incentives and is consistent across various environments, including among more than 10,000 middle and high school students. Interestingly, the gender gap in self-evaluations does not appear in verbal ability assessments, suggesting that it is less likely in female-typed domains, and our research contributes in this direction.

The paper proceeds as follows. First, we describe the baseline game and deduce a hypothesis for this game. This will be followed by a description of the institutional environment and the design, the analysis, and a discussion of the implications of these findings.

II. Research design

Institutional setup

We use academic records, including student performance and behavior, in a first-year course at a Sino-UK university in China to test our arguments, since one of its assessments, called the *assignment*, which has a peer review, provides a good context that fits our study. The data we use are from three consecutive academic years, 2016/17 to 2018/19, during which the rules for peer review varied every year, which allows us to compare students' behavior under different peer review mechanisms. In addition, this course was taught and marked by the same instructor in these three years, so there was no consistency issue in marking or instructor effects over these years.

This course can accommodate up to 800 students and this cap of student number has been reached almost every year. It provides a basic understanding of economic data and some basic tools for data processing and analysis. Being delivered in the second semester of each academic year, it is one of the first professional courses that first-year students can take: In the first semester, first-year students take only language-related courses and some nonprofessional ones.

The course had two assessments, the assignment and the final exam, before the 2017/18 academic year and a third assessment, the midterm, since 2017/18. All of these exams were closed-book, assessing students' understanding of economic data and the data analysis tool. The assignment involves a group project consisting of a written report, a presentation, and a peer review. The group project assessed

the students' ability to apply what they learned to analyze real-world economic data, and peer review was used to punish free-riding and enhance collaboration.

Every year, students taking this course formed groups of either 5 or six students for the assignment. The group formation process, conducted online, was voluntary: The instructor set groups with five or six open slots on a teaching platform, and the students chose which group to join by themselves. Once all open slots in a group have been filled, no more students could join this group. Each student could only join one group. Only students who did not choose their groups on time would be randomly assigned to groups with open slots by the instructor, whereas such cases were very rare every year. No student could drop out of this course and hence their group once after the deadline for group choice.

Each group had to write and present a report analyzing real-world data using the empirical skills they learned in this course. After having submitted their report, students (voluntarily) participated in the peer review to give grades to their group members as evaluations of their contributions to the group project. These grades were used to compute the *contribution parameter* that affected the student's scores for the assignment: A student's grade in the peer review is the arithmetic mean of the grades they received in the peer review, and their contribution parameter is the ratio between their grade and the highest one of their group members' grades in the peer review. For example, if a student's grade in peer review was 85 and the highest grade in their group was 90, then the contribution parameter for this student was $85 / 90 = 0.95$.

The peer review was conducted on the same online teaching platform used for group formation. The students had ten days to do the Peer Review, and they could do it at any time and location during the ten days. They neither needed to sit together to do the peer review nor told others the grades they gave. The grades a student gave in the peer review were private, known to them and the instructor. They did not know what grades they received or who in their group participated or did not participate in peer review.

Students could abstain from peer review. If they participated in the peer review, they needed to give grades to all the required members: Before 2018/19, students participating in peer review must give grades to all members of their group, including themselves. Since 2018/19, they were only able to give grades to their groupmates but not to themselves. The platform prevented a student from submitting his grades for peer review if he gave grades to some, but not all, of the required group members.

The students knew the rules of how peer review was conducted as well as other rules for the assignment since the beginning of this course every year: In the first week of this course, the instructor announced the assignment instructions online and explained them briefly in the first lecture. Examples of how the contribution parameters are calculated were also given in the instructions.

More details on the marking scheme of the assignment and the rules of computing the contribution parameter can be found below. μ_i represents student i's

mark for the assignment and c_i stands for their contribution parameter. m_g, r_g, and p_g denote the group mark for the group project, the mark for the report, and the mark for presentation, respectively.

2016/17

In the academic year 2016/17, the two assessments, the assignment and the final exam, contributed 30% and 70% to the final mark of a student of this course, respectively. That is, $m_g = r_g \times 0.7 + p_g \times 0.3$.

The marking scheme for the assignment was the following. The group mark for the group project is the weighted sum of the marks for the report and the presentation. A student's mark for this assignment equals the group mark for the project times their contribution parameter. Then, the student i's mark for the assignment is calculated using the following formula:

$$\mu_i = m_g \times c_i = \big(r_g \times 0.7 + p_g \times 0.3\big) \times c_i.$$

In cases where none of the group members participated in peer review (i.e., when all members abstained from providing peer evaluations), each individual in the group would automatically receive a contribution parameter of 1.

2017/18 and 2018/19

Beginning in the 2017/18 academic year, a new component, *midterm exam*, was introduced as part of the course assessment structure. The midterm exam accounts for 10% of a student's final grade, while the *assignment* and the final exam contribute 20% and 70%, respectively. This adjustment aimed to provide a more balanced evaluation of students' knowledge and performance throughout the course duration. The weighting scheme for the group project component was adjusted, with the report contributing 60% and the presentation contributing 40% to the overall group mark, defined as:

$$m_g = r_g \times 0.6 + p_g \times 0.4.$$

Additionally, the marking scheme for the assignment was revised to include an individual mark component (π_i), as mandated by the exam board. Consequently, the final mark for student i on the assignment was calculated using the formula:

$$\mu_i = m_g \times c_i \times 0.6 + \pi_i \times 0.4,$$

where the individual mark (π_i) was determined based on the specific contributions of each student to the group project. To ensure transparency, each group was required to submit a cover sheet detailing the contributions of each member, and all members signed the document to affirm their agreement with the reported contributions.

In both the 2017/18 and 2018/19 academic years, the rules for calculating the contribution parameter (c_i), as well as those governing the peer review process and group formation for assignment, remained largely consistent with those established in 2016/17, except for the following modifications:

- In 2016/17 and 2017/18, students were required to assign grades to all group members, including themselves, as part of the peer review. However, in 2018/19, the peer review rules were revised to prohibit self-grading, allowing students to grade only their peers.
- The instructions for the assignment in 2017/18 and 2018/19 explicitly stated that if no member of a group participated in the peer review, each member of that group would receive a contribution parameter of 1. This clause, although applied in practice by the instructor in 2016/17, was not explicitly mentioned in the instructions for that year.
- In 2018/19, it was clarified in the assignment instructions that if only one member of a group completed the peer review, that student would receive a contribution parameter of 1, even though they could not assign a grade to themselves. This provision was not included in the instructions for the assignment in 2016/17 or 2017/18.

Condition Self and Condition No Self

For ease of expression, we refer to the case in which agents must give evaluation to both themselves and their groupmates in a peer review as Condition Self and that in which agents only give evaluations to their groupmates but not themselves as Condition No Self in the rest of this paper. Then, the peer reviews for 2016/17 and 2017/18 belong to Condition Self and in 2018/18 belong to Condition No Self.

III. Behavioural considerations

Suppose we assume that students are grade maximizers. In that case, we show how a rational agent should behave in Condition Self and then discuss why deviations from the Nash prediction can be expected, given evidence from the previous literature. We conclude this section by discussing the predictions for Condition No Self, which largely mirror those for Condition Self.

A. Purely selfish, grade maximizing agents

Consider a group of N players in a simultaneous rating game, where $r_{i,j}$ is the rating provided by i to j. Therefore, the set of strategies is $R_i = \{r_{i,j} \in \mathbf{Z} | 0 \leq r_{i,j} \leq 10\} \ \forall \ i$. Let $\sigma_j = \sum_{i=1}^{i=N} r_{i,j}$ be the total rating subject j receives. Let $\bar{\sigma}$ be the maxima of the set $\{\sigma_j\}$. The individual payout is then the ratio of her

aggregated rating to the maximum level of rating in her group of N players. In the baseline model, we assume that agents care only about their own grade, which can be considered a material payoff.

Let $\sigma_j = \frac{\sum_{i \in I_j} r_{i,j}}{\#I_j}$ denote the arithmetic mean of the ratings subject j receives in peer review. Here, I_j denotes the set of subject j's group members who participate in peer review and give ratings to j and $\#I_j$ denotes the number of subjects in I_j. j is also in I_j if and only if j participates in the peer review and they are allowed to rate themselves in the peer review (Condition Self). It should be noted that if subjects are allowed to rate themselves, then for any two subjects, j and k, in the same group, $\#I_j = \#I_k$, regardless of whether both j and k participate or abstain, or only one of them participate in, peer review. Instead, if subjects are not allowed to rate themselves in peer review, then $\#I_j = \#I_k + 1$ if j abstains while k participates in the peer review. The payoff for the player is determined by their relative rating π_j.

$$\pi_j = \frac{\sigma_j}{\bar{\sigma}} \tag{1}$$

We restrict our attention to pure-strategy Nash equilibria. As shown in the following propositions, many pure-strategy NE exist.

LEMMA III.1: *Abstention is always a weakly dominated strategy in both the Condition Self and the Condition No Self cases.*

The proof can be found in Appendix A.A1.

PROPOSITION III.2: *A pure strategy profile r^* is a Nash equilibrium as long as $\pi_j = 1\ \forall j$.*

Proof skipped due to simplicity.

Indeed, the above proposition shows that any strategy profile leads to complete collusion (i.e., every group member receives the maximum payoff, and (1) is always a Nash equilibrium. Furthermore, by definition, all of these NEs are efficient and admissible.

PROPOSITION III.3: *For $N > 2$, there exists a pure strategy Nash equilibrium in which $\pi_j < 1$ for some j.*

PROOF:

Existence can be shown with an example. One such equilibrium is that all $N-1$ players rate themselves and all others except player i, 10, and all rate player i with 0. For the player i, she rates all the others 0 and herself 10. In such a case, all $N-1$ players have a payoff of 1, and player i has a payoff smaller than 1.

The above example speaks to the case of collusion within subgroups. In such a scenario, some agents in equilibrium may receive partial pay-offs. The following assumption rules out subgroup collusion and simplifies the analysis.

ASSUMPTION III.4: *The strategies for all agents are non-discriminating, meaning that agents give the same ratings to all players other than themselves. Mathematically, for each individual i,* $r_{i,j} = r_{i,j}\ \forall j \neq i$.

This assumption precludes the possibility of collusion among subgroups, which is a reasonable expectation if individuals are solely concerned with maximizing their own material gains.

PROPOSITION III.5: *If the strategies for all agents are non-discriminating, then* $\pi_j = 1\ \forall j$ *for all pure strategy Nash equilibria.*

PROOF:

By contradiction. Assume $\exists$ a Nash equilibrium r^* in which $\pi_j < 1$. Then her best response must be that $r_{j,j} = 10$ and $r_{j,i} = 0\ \forall\ i \neq j$. Given the non-discriminating property, $\sigma_{i \neq j} \leq \sigma_j$.[1] Thus $\sigma_j = \bar{\sigma}$, contradiction.

Then we apply several popular refinements to the set of equilibria. The only strategy that survives all three of the following criteria: strong Nash equilibrium, iterative elimination of weakly dominated strategies, and stability is $r_{j,j} = 10$ and $r_{j,i} = 0\ \forall\ i \neq j$. Therefore, the baseline model hypothesizes that agents will self-rate at the highest possible level while assigning the minimum possible ratings to their peers.

B. Rationales for deviations from grade maximizing behavior

Despite the baseline model, characterized by the predictions $s_{j,j} = 10$ and $s_{j,i} = 0$ for all $i \neq j$, achieving Pareto efficiency, there are multiple reasons for deviations from this prediction. First, in this context, there were moral appeals[2], both social and individual norms suggest that rating decisions should not be viewed solely through the lens of a strategic game. Recent experimental evidence suggests that the psychological and social costs associated with dishonesty often motivate people to tell the truth (see Abeler, Nosenzo and Raymond, 2019, for a meta-analysis). Furthermore, even among those who choose to deceive, partial lies are frequently observed (Gneezy, Kajackaite and Sobel, 2018).

Moreover, there are different types of misreports possible in this situation that may carry different moral values. There may be white lies, i.e., misreporting that benefits others and not oneself. Dishonestly assigning very low ratings to your peers' contributions does not qualify as a 'white lie' (Erat and Gneezy, 2010). Although assigning maximum ratings to peers could be interpreted as white lies, which are generally more acceptable on a social and individual level (Gneezy et al., 2017; Michailidou and Rotondi, 2019). Hence, we hypothesize that psychological lying costs or self-image concerns can dissuade individuals from drastically misrepresenting, especially in underreporting all group members' contributions.

[1]The non-discriminating property ensures that for any two agents i, j, the comparison of their aggregated rating σ_i and σ_j, depends only on $r_{i,i}, r_{i,j}, r_{j,i}, r_{j,j}$.

[2]The instruction (that can be found in Appendix A.A7) highlighting that the design aims at group enhancement and free-riding punishment, and one should give a rate based on actual contribution.

Although truth-telling preferences tend to shift the outcome away from the equilibria (Pareto-efficient realizations), the moral wiggle room literature points to a possibility of collusion with truth-telling preference (Dana, Weber and Kuang, 2007; Spiekermann and Weiss, 2016). In our particular case, moral flexibility arises from the subjectivity inherent in the evaluations, which excuses self-serving assessments. For instance, although aware that a teammate has contributed more, maximum ratings could still be awarded to both parties, justifying this with the claim "I am a generous marker and I think both of us deserve the highest rating." An alternative excuse could be "Given my lack of experience in evaluating contributions, equal ratings seem appropriate." Essentially, moral wiggle room reduces the dishonesty cost for certain deviations that can be excused.

An individual decision model with lying cost and moral flexibility

In this model of N players, the type of agent i' is characterized by a pair (S_i, θ_i), where S_i is a set of observed contributions in the eyes of agent i, and θ_i is her lying cost parameter. Let $s_{i,j}$ represent the observed contribution of player j, as perceived by player i. Without loss of generality, we assume that $s_{i,j}$ follows an i.i.d. discrete distribution F, with the probability mass function: $P(s_{i,j} = k) = p_k$, for $k = 0, 1, 2, ..., 10$. Thus, $S_i = \{s_{i,1}, s_{i,2}, ..., s_{i,N}\}$ comprises the collection of all observed contributions, including her own, with each player i observing contributions from N players (including herself), implying $|S_i| = N$ for all i.

The parameter $\theta_i \in [0, \infty)$ denotes the cost parameter for agent i to lie, where higher values indicate a higher cost associated with lying. $G_i(\cdot)$ is the cumulative distribution of θ_i, which is independently distributed with full support.

Then we specify the actions. An agent i gives a rating $r_{i,j}$ to player j, and $r_{i,j} \in \{\mathbb{Z} \mid 0 \leq x \leq 10\}$. Then an agent is honest if $r_{i,j} = s_{i,j} \forall\, i, j$, and any deviations that $r_{i,j} \neq s_{i,j}$ are considered as standard lying. In addition, the agent can surrender to moral flexibility by fabricating the observation $s_{i,j}$ to $\widetilde{s_{i,j}}$, where $\widetilde{s_{i,j}} = \widetilde{s_{i,j}}\ \forall\ j$. Put differently, the model defines moral wiggle room as a justification stemming from the agent's purported inability to accurately discern and evaluate the true contributions, which parallels the concept of information avoidance seen in excuse-based self-deception (Dana, Weber and Kuang, 2007; Jaroszewicz, Loewenstein and Benabou, 2024). To simplify, the model posits that the agent has the option to engage in either standard lying/honesty or excuse-based lying, but cannot adopt both strategies concurrently.

An agent's preference depends on three elements: the material payoff, a lying cost that varies with the size of the lie, and a psychological cost associated with excuse-based lying. Let U_i stand for the utility of agent i.

$$U_i = V(r_{i,j}) - (1 - I_i)C(s_{i,j}, r_{i,j}, \theta_i) - I_i\gamma \tag{2}$$

The function $V(\cdot)$ stands for the material payoff, which is solely dependent on the ratings. The variable I_i serves as an indicator variable, taking the value of 1 if the agent opts for excuse-based lying and 0 otherwise. The function $C(\cdot)$ measures the cost of standard lying. Finally, we use γ to represent the fixed cost of using moral wiggle room to fabricate the observation.

To provide a practical prediction, we limit the analysis to the following special functional forms. First, we characterize the material payoff function $V(\cdot)$.

$$V(r_{i,j}) = \sum_j (r_{i,i} - r_{i,j}) \tag{3}$$

The payoff function reflects a principal insight derived from our baseline prediction. The aggregated difference between self-ratings and ratings given to others captures the direct strategic incentive of the game, which is the individual's preference to attain a higher rank within the group.

Then, we specify the standard lying cost as a function of the size of the lies.

$$C(s_{i,j}, r_{i,j}, \theta_i) = \sum_j [(|r_{i,j} - s_{i,j}| + 1)^{\theta_i} - 1] \tag{4}$$

This cost function exhibits the following characteristics: For agents characterized by $\theta = 0$, no cost is associated with lies of any magnitude. For those with $\theta > 0$, the lying cost escalates to the size of the lie. Furthermore, maintaining honesty is cost-free for agents across all values of θ.

Finally, a fixed cost γ measures the cost of self-deception when agents choose to falsify $s_{i,j}$ to $\widetilde{s_{i,j}}$.

In sum, decision making is characterized by the expression (5), with the uniform misrepresenting constraint: If $I_i = 1$, then $s_{i,j} = \widetilde{s_{i,j}} = c \ \forall \ j$, where $\widetilde{s_{i,j}}$ is an integer agent that can choose between 0 and 10. In addition, we present the mutually exclusive assumption: If $I_i = 1$, then $r_{i,j} = \widetilde{s_{i,j}} \ \forall \ j$. These two assumptions lead to zero material payoff that if $I_i = 1$, $V(r_{i,j}) = \sum_j (\widetilde{s_{i,i}} - \widetilde{s_{i,j}}) = 0$.

$$\max_{r_{i,j}, I_i} (1 - I_i) \sum_j (r_{i,i} - r_{i,j}) - (1 - I_i) \sum_j [(|r_{i,j} - s_{i,j}| + 1)^{\theta_i} - 1] - I_i \gamma \tag{5}$$

The decision-making process ultimately aligns with a binary comparison. Agents weigh the maximum payoff from standard lying (including honesty) against the benefits of self-deception. The model leads to the following predictions.[3]

[3] A detailed version can be found in Appendix A.A2.

PREDICTION 1: *The baseline game strategy is adopted dishonestly with a positive probability.*

The intuition of Prediction 1 is straightforward. For those with a lying cost parameter θ smaller than a threshold, lying to the maximum trumps other alternative strategies.

PREDICTION 2: *Honest ratings different from the baseline game strategy are proposed with a positive probability.*

On the other hand, for those players with θ greater than a threshold, being honest is the best standard lying strategy. Among these cases, some also dominate self-deception.

PREDICTION 3: *Dishonest ratings different from the baseline game strategy (partial lies) are used with a positive probability.*

Our model predicts that partial lies may be optimal when θ falls within a certain range.

PREDICTION 4: *The occurrence of dishonest egalitarian ratings is probable (i.e., self-deceptions are possibly observed).*

In addition to the predictions that emerge directly from the model's functional form, we also propose two further behavioral predictions to account for other behavioural motives.

ADD. PREDICTION 1: *When self-deception is present, agents tend to assign the highest ratings to all recipients.*

Although the functional form of the model does not explicitly distinguish between egalitarian ratings resulting from moral flexibility, we posit that assigning the maximum ratings to all participants represents the most appealing option. Primarily, when calculating the rankings, providing the maximum egalitarian ratings has the most significant impact on the outcomes.[4] In addition, according to the literature on expressive voting, such rating decisions could be self-expressive, and therefore people might prefer higher procedure ratings (Greene and Nelson, 2002).

ADD. PREDICTION 2: *There may be gender differences in the ratings, possibly due to variations in the perceived costs of dishonesty between genders.*

[4]In other words, among all egalitarian ratings, the maximum one has a weakly stronger effect driving the final outcome towards an egalitarian outcome. Here is an illustrative example. Suppose in a group of two agents, A and B. Let A choose the selfish strategy of $(10, 0)$, and B choose the egalitarian ratings (x, x). Then the final payoff for B positively correlates with the value of x. When $x = 0$, B receives a final score of $\frac{0+0}{10+0} = 0$. When $x = 10$, B receives a final score of $\frac{0+10}{10+10} = \frac{1}{2}$.

Given the central role of the lying cost parameter in our predictions, we anticipate observing gender differences. Specifically, as research indicates that men are generally more prone to lying than women (see meta-analysis by Capraro (2018)), we expect distinct lying cost parameters between genders, and thus different rating behaviors.

C. *Predictions for* Condition No Self

The prediction of the baseline model for Condition No Self follows the same theoretical insight. In terms of the propositions, we could establish identical results. Specifically, Pareto-efficient outcomes are equilibria. There exist other equilibria which are not Pareto-efficient, but once we introduce the non-discriminatory assumption, then we find that all pure-strategy Nash equilibria lead to Pareto-efficient results. Applying similar refinement principles also demonstrates that each agent is likely to assign the minimum possible rating to others for strategic benefits.

Behaviorally, however, the utility model shifts from focusing on the differential between self-ratings and others' ratings to a preference for the lowest feasible ratings for others, driven by strategic positioning within the group. The dynamics of standard lying and self-deception motives remain consistent with those in the Condition Self scenario, where agents encounter variable and fixed costs associated with misreporting and misrepresenting observations, respectively. Mathematically, with a comparable functional form, the utility representation is as follows:

$$\max_{r_{i,j},I_i} \sum_{j\neq i}(-r_{i,j}) - (1-I_i)\sum_{j\neq i}[(|r_{i,j}-s_{i,j}|+1)^{\theta_i}-1] - I_i\gamma \tag{6}$$

Although the standard predictions in the Condition No Self are replicated—specifically, standard lying and honesty are probable, partial lies are present, and self-deception about the observation is also possible—the main difference lies in the following prediction when comparing the expressions (5) and (6):

Differential prediction

For Condition Self, the material payoff is the aggregated differences between the ratings for themselves and others, while for Condition No Self, this difference no longer exists. Both utility functions share the same intuition; without lying costs, the agent has strategic incentives to rate the other group members as low as possible. On the contrary, differences in the two preference representations offer very different behavioral insights once self-deception-based egalitarian ratings are in place. For Condition Self, Prediction 3 presents tie-breaking reasons for the highest ratings among all egalitarian ratings. In contrast, for

CONDITION NO SELF, the utility function itself points to the opposite result, which is that the agents prefer to provide the minimum ratings to others.

ADD. PREDICTION 3: *When self-deception is possible, in CONDITION NO SELF, agents have incentives to assign lower ratings to all recipients compared to CONDITION SELF.*

D. Simulation

To characterize our theoretical prediction, we conducted a simulation focusing on the impact of the cost of standard lying, θ, the observation (initial distribution of the contributions) $s_{i,j}$ and the fixed cost of excused-lying γ. We classify the agents' actions into four categories: honest, (standard) partial lie, (standard) maximal lie, and self-deception. The simulation results show how the optimal action evolves depending on the parameters in the model.

Setup (CONDITION SELF baseline)

The simulation proceeds as follows: For each value of the cost parameter θ, the simulation is run 1,000 times. In each iteration, the observed signals s_{ij} for all 6 members of the group are generated randomly from a predefined uniform distribution. The model evaluates four lying strategies: excuse-based lying, honest reporting, maximal lying, and optimized lying.

The utility of each strategy is calculated based on the utility function (5) that considers the benefit of the reported ratings and the cost of lying. The strategy with the highest utility is selected as the best strategy for each iteration. The simulation tracks the self-rating, the average rating for others, and the category of the best strategy for each iteration.

Results and visualisation

The results of the simulations are visualised in two parts:

1) An area plot showing the percentage outcomes for each strategy as θ varies.

2) Scatter plots that display the relationship between self-ratings ($r_{i,i}$) and the average rating given to others ($r_{i,j}$, $j \neq i$).

The results in Figure 1 confirm our expectations and the four theoretical predictions. First, when the lying cost is low, many players adopt the baseline game strategy (maximal lie). Second, honest actions become more prevalent as the cost of lying increases. Third, partial lies are optimal for some agents with a medium lying cost. Fourth, self-deception actions emerge as standard lying becoming more costly. In Appendix A.A3, we include additional simulation results detailing the impact of $s_{i,j}$ and γ. The simulation results are consistent with our expectations about our model.

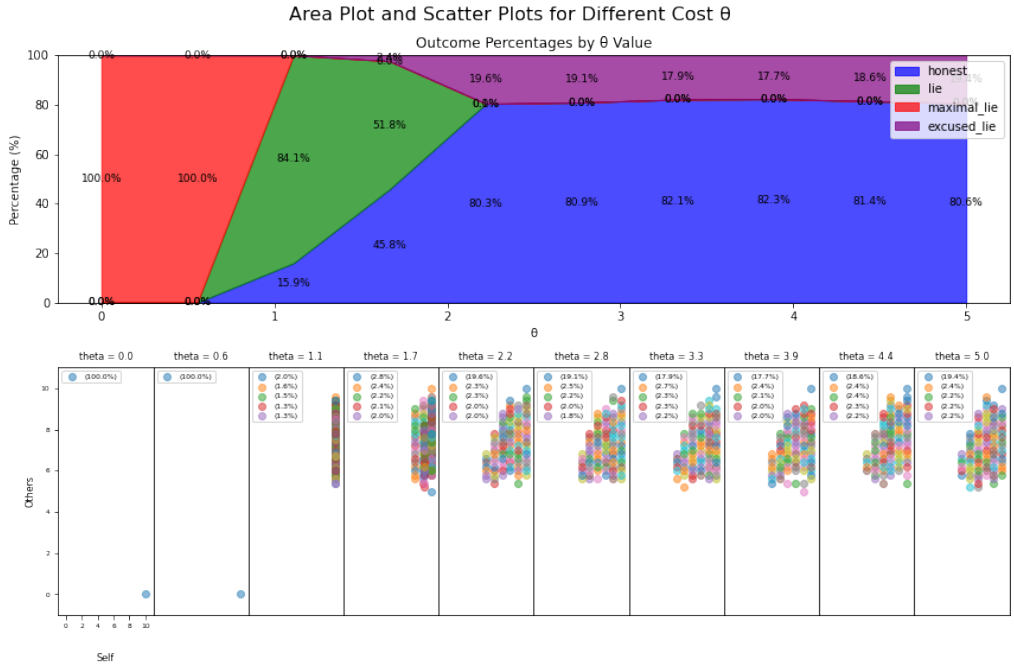


FIGURE 1. AREA AND SCATTER PLOTS FOR THE SIMULATION.

Note: θ ranges from $[0, 5]$, $\gamma = 10$ and $s_{i,j} \sim U(5, 10)$. For the scatter plots, the legend details the five most frequent occurrences.

E. Collusion

Although our simple behavioral model effectively complements the baseline game-theoretic analysis, exploring an approach without non-standard behavioral assumptions that deviate from pure payoff maximization could provide additional insight. Despite the private and anonymous mechanism of our rating system, collusion could still influence the outcomes. The structure of payoffs, where all equilibria are Pareto-efficient and individually optimal, naturally discourages deviation among colluding groups if agents prioritize their payoffs. The field experiment setting, characterized by frequent communications between players 'outside the game', supports findings from the literature that suggest that communication enhances collusion in games (Fischer and Normann, 2019). In response, our subsequent analysis will include results that both omit and assume the presence of collusion to demonstrate the range of potential outcomes.

IV. Results

A. An overview of the data

A total of 2,376 students formed 415 groups. In terms of size, 318 groups of six (76.6%), 83 groups of 5 (20.0%) and 14 groups of four or less (3.4%).[5] Group formation happened after the possibility of changing course, that is, after the third week, all students who initially were in a group stayed there until the end of the course.

The first outcome measure that we explore is related to how individuals evaluate themselves and others in a strategic environment. Understanding the evaluation behavior in our data can provide insight into understanding workplace evaluations in practice. Hence, we explore whether there is a gender difference.

Table 1—Summary Statistics: individuals

	Condition Self				Condition No Self	
Dataset	D16-17		D17-18		D18-19	
	M	**F**	**M**	**F**	**M**	**F**
N	253	541	217	570	248	547
% Econ	74.5	64.8	71.0	61.6	74.6	65.4
% Management	25.3	34.0	28.6	36.7	25.0	34.4
Exam	72.3	78.4	61.5	67.0	60.0	65.0
Self-rate	9.83	9.69	9.80	9.79	N/A	N/A
Other-rate	9.37	8.95	9.48	9.31	9.45	9.30
Rated (excluding self)	8.59	9.14	9.03	9.38	9.11	9.39
% Abstention	26.1	10.4	16.1	11.1	7.7	7.7

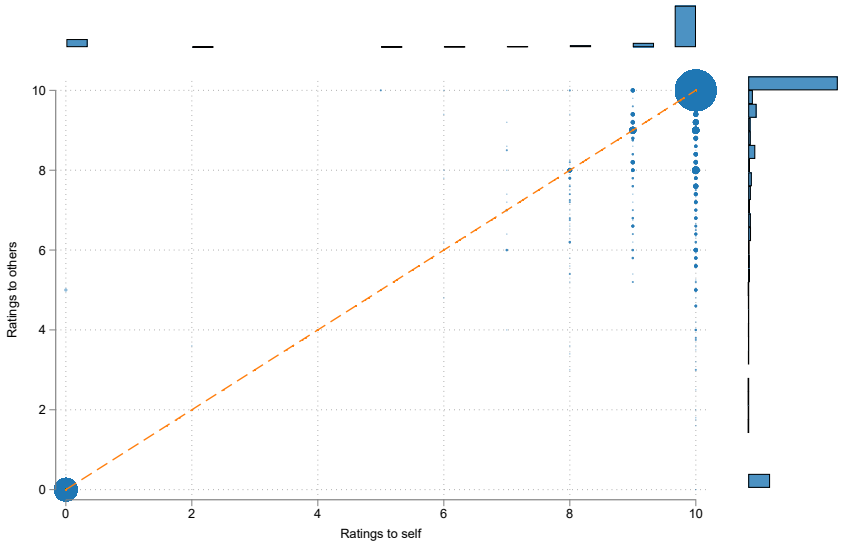
Note: Self-rate is the rating one gives to oneself. Other-rate is the average of one's ratings to her peers. Rated (excluding self) is the average rating one received from all her peers. For the academic year D16-17, we exclusively documented the final exam scores of the course participants. However, for subsequent years D17-18 and D18-19, the variable *Exam* encompasses a weighted mean calculation derived from both one midterm and one final exam. The weighting assigned replicates the actual importance in their overall course evaluation, and is 12.5% for the midterm and 87.5% for the final exam.

The primary variables of interest are the self-assessment ratings and the peer evaluation ratings. How are social and self-evaluations influenced? We anticipate that strategic motivations and the adherence to social norms play a crucial role. On the one hand, individuals may have incentives to exaggerate their own contributions. However, there are also lying costs associated with misrepresenting the actual contributions of all group members.

[5] A total of thirteen students who did not attend the final exam are excluded.

Figure 2 illustrates the distribution of these ratings.[6] The horizontal axis represents self-assessment ratings, while the vertical axis denotes the average ratings given to group members excluding oneself. The distribution of individual rating choices is represented using bubbles and histograms. A dashed line serves as a reference line: in the absence of bias towards oneself, we anticipate observations to exhibit symmetry around this line. In particular, our visual examination contradicts this symmetry, with more observations falling below the 45-degree line, indicating that individuals tend to rate themselves higher than their peers. Furthermore, our data cluster around the potential cooperative point (10,10), along with a notable incidence of abstention at the origin (0,0).[7]

FIGURE 2. RATINGS TO SELF AND OTHERS



Note: The diameter of the bubbles is directly proportional to the observation frequency. The dashed line represents the line of equality at a 45-degree angle. The histograms displayed along the x- and y-axes illustrate the distribution of observations along each axis. The point at coordinates (0, 0) indicates individuals who did not respond.

Table IV.A outlines the characteristics of the groups. In particular, between 23. 6% and 32. 1% of all groups exhibit zero variance, which means that all members received the highest ratings. Further analysis reveals that complete abstention by group members is not a significant factor, since only 5 out of 112

[6]For a gender comparison version of the figure, see Figure A7 in the appendix. The primary result is that both genders exhibit a similar self-serving pattern in their rating behavior.

[7]All observations located at the origin correspond to instances of abstention. No participant intentionally provided a rating of (0,0).

zero-variance groups resulted from collective abstention. Instead, the majority of these groups consistently rated all members the highest (only four zero-variance groups provided ratings below 10).

Our model proposes three explanations for zero-variance groups. First, collusion may be a factor, as identical rating actions among members can result in efficiency, incentivizing collusion. Second, self-deception can lead to uniformly high ratings. Third, zero variance could accurately reflect the actual high contributions perceived as such by all members. Lastly, a combination of these factors could be responsible for the observed zero variance. Therefore, fully characterizing zero-variance groups is challenging. What follows is that we cannot identify collusive groups with certainty and exclude them when examining the determinants of rating behaviors. Instead, we include all groups with zero variance, providing a conservative estimate and a comprehensive overview. Additionally, we replicate several key results excluding zero-variance groups in the appendix for comparison.

Table 2—Summary statistics: Groups

	Condition Self		Condition No Self
	D16-17	D17-18	D18-19
N	138	137	140
Size	5.75	5.74	5.68
% Female-Majority	66.0	77.4	75.0
% Female-Only	15.2	24.1	22.1
% Zero-Variance	25.4	32.1	23.6

Note: Female-majority groups contain strictly more than 50% of female members. Zero-variance groups are those where all group members receive the same aggregated ratings.

B. Regression model

We use a regression model to statistically examine whether individuals exhibit self-serving bias and/or truth-telling tendencies. The identification strategy is straightforward: we employ individual exam performance as a proxy for academic ability and attitude. We *assume* that the latent contribution to group projects is positively correlated with academic performance, as academically stronger students are more likely to contribute more to collaborative projects (Espey, 2022). Therefore, a positive relationship between the evaluation and the academic indicator suggests that the evaluation reflects, at least in part, the actual contribution. To assess self-serving bias, we introduce a binary variable $Self$, anticipating a positive and statistically significant coefficient.

$$Y_{i,j} = \alpha + \beta Self + \gamma X + \epsilon$$

Here, $Y_{i,j}$ represents the ratings provided by subject i to subject j, while $Self$

denotes the dummy variable for self-assessment. X includes a vector of control variables, such as the gender and exam performance of the raters and receivers.

Table 3—Determinants of Rating Behaviour

Dependent: Rate	Condition Self						Condition No Self
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	D16-17	D16-17	D17-18	D17-18	Combined	Combined	D18-19
Self	0.631***	0.756***	0.425***	0.555***	0.527***	0.659***	
	(0.0706)	(0.134)	(0.0581)	(0.132)	(0.0458)	(0.0939)	
MarkerFemale	-0.299***	-0.445***	-0.179*	-0.327**	-0.241***	-0.394***	-0.228**
	(0.0946)	(0.126)	(0.0990)	(0.135)	(0.0686)	(0.0922)	(0.0963)
ReceiverFemale	0.355***	0.384***	0.244**	0.273**	0.303***	0.333***	0.0991
	(0.0938)	(0.109)	(0.0998)	(0.115)	(0.0682)	(0.0790)	(0.104)
MarkerExam	-0.00925**	-0.00926**	-0.0103***	-0.0103***	-0.00989***	-0.00990***	-0.00858**
	(0.00368)	(0.00367)	(0.00376)	(0.00376)	(0.00264)	(0.00264)	(0.00392)
ReceiverExam	0.0180***	0.0180***	0.0183***	0.0183***	0.0182***	0.0181***	0.0261***
	(0.00345)	(0.00345)	(0.00379)	(0.00379)	(0.00258)	(0.00258)	(0.00538)
Self × MarkerFemale		0.174		0.177		0.182*	
		(0.151)		(0.153)		(0.107)	
D17-18					0.369***	0.368***	
					(0.116)	(0.116)	
Constant	8.290***	8.252***	8.750***	8.709***	8.346***	8.306***	8.294***
	(0.366)	(0.382)	(0.314)	(0.333)	(0.265)	(0.277)	(0.436)
Observations	3925	3925	3991	3991	7916	7916	3469

Note: The standard errors are presented in parentheses, and the significance levels are * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Multilevel mixed-effects models with robust standard errors, incorporating random intercepts for both groups and individual subjects. The categorical variables D17-18 and D18-19 represent different datasets, with D16-17 as the reference category. Self is a dummy variable that indicates whether the rating is a self-assessment. MarkerFemale is a dummy variable equal to one if the marker is female. ReceiverFemale is a dummy variable equal to one if the receiver of the rating is female. MarkerExam stands for the marker's exam performance. ReceiverExam stands for the receiver's exam performance. Self × MarkerFemale denotes the interaction between Self and MarkerFemale.

Table 3 is structured to highlight the effects of key variables, such as self-assessment, gender of the marker and receiver, and academic performance on ratings across different treatments and datasets. Subjects were significantly more inclined to rate themselves higher, corroborating our visual observations. This tendency towards self-serving is consistent across all models presented. Moreover, both men and women exhibit self-interested rating behaviors, with only subtle indications of variance in intensity (the interaction term Self × MarkerFemale is marginally significant in the combined dataset, column (6)). Another consistent finding is the correlation between individual academic performance and ratings. As expected, ratees with higher scores tend to receive more favorable ratings. Similarly, individual academic performance also relates to how students rate their peers, with academically stronger students tending to rate others less favorable. Assuming a positive correlation between individual academic performance and contributions to group activities, a plausible explanation for this phenomenon is that those who contributed more judged their peers more rigorously.[8]

[8]This finding aligns with previous experimental evidence from public goods games, indicating that

Another observation is that in the context of Condition Self, women were more likely to assign lower ratings while receiving higher ratings from their peers. In contrast, within the scenario Condition No Self, women did not receive more favorable ratings. We posit that this disparity can be partially attributed to differences in abstaining between the two treatments. The latter section on abstention provides a detailed discussion.

C. Nonparametric analysis

We apply an alternative nonparametric test to depict the presence of self-serving bias and the disparity between self-evaluation and social evaluation (Deb and Renou, 2022). A visual representation is shown in Figure 3. The test asserts that, under the assumption of equal means, discrimination or bias exists if one distribution stochastically dominates the other. This confirms our result that subjects consistently rate themselves higher than others. The nonparametric test (Linton, Maasoumi and Whang (2005) test) robustly confirms this first-order stochastic dominance. Moreover, in conjunction with the regression results, this underscores a significant self-serving pattern in rating behavior.

Of particular interest is the gender disparity in social recognition. Females consistently receive higher social recognition across all three independent samples, surpassing their male counterparts at all levels. This observation is further highlighted by the tendency of both men and women to rate their female peers more favorably than their male peers.

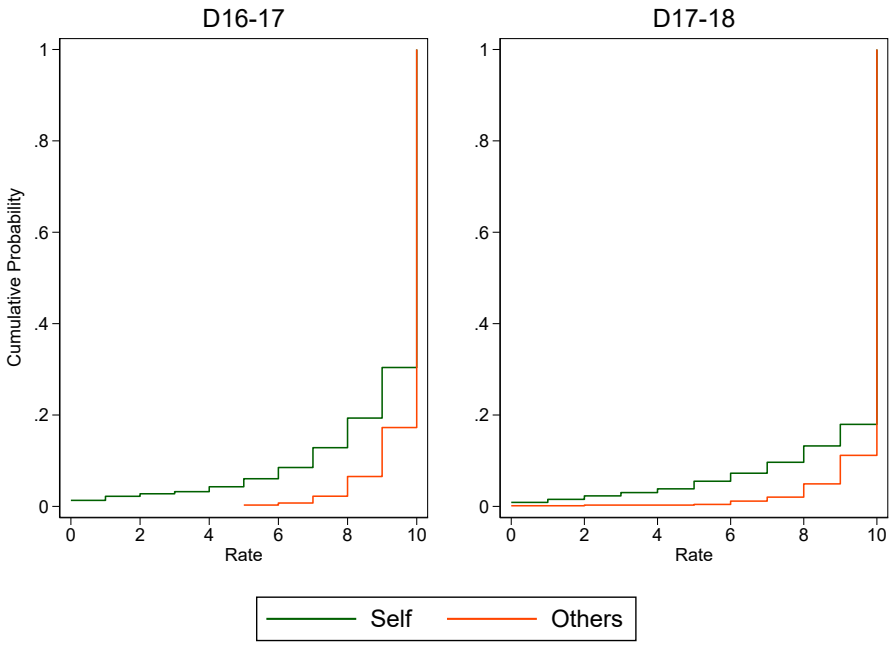
Two competing hypotheses emerge to explain the apparent discrimination against male students. The first postulates taste-based discrimination against male students, while the second suggests that unobserved characteristics (e.g., communication and social skills), in addition to controlled differences such as academic performance, contribute to the higher social recognition of females. We approach these hypotheses with caution, tentatively rejecting the former and providing a more nuanced explanation in the discussion section.

FINDING 1: *(Self-serving bias). a) Both men and women rate themselves significantly higher than they rate their peers. However, the extent of this self-serving bias is moderate, and individuals almost never adopt a purely selfish rating strategy. b) Male students rate themselves slightly higher than female students. However, male students also rate others higher than female students. Overall, the degree of self-serving bias is similar between male and female students.*

FINDING 2: *(predictors of ratings). a) Academic performance strongly influences rating behavior. Academically stronger raters tend to rate others lower, whereas academically stronger raters receive higher ratings. b) Gender also predicts ratings on top of academic performance. Female raters generally rate others lower, while female ratees receive higher ratings from raters of both genders.*

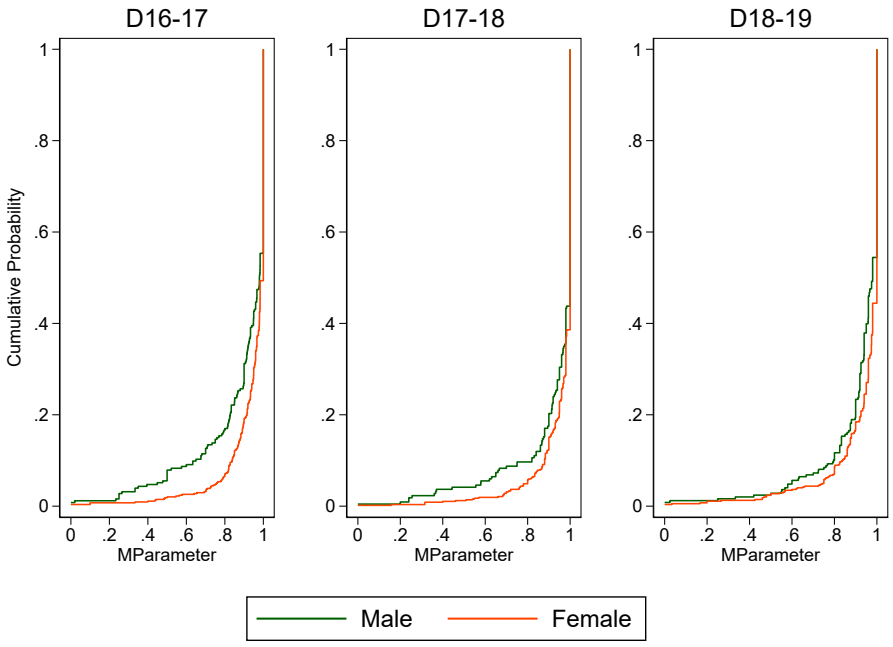
participants often use their own contributions as benchmarks for making punishment decisions(Carpenter and Matthews, 2009).

FIGURE 3. RATINGS TO SELF AND TO OTHERS



Note: The figure denotes the cumulative probability distribution for the ratings one gives to herself and to her peers.

FIGURE 4. RECEIVED PARAMETER BY GENDER



Note: The figure shows the cumulative probability distribution for the overall parameter (MParameter) one received, by gender.

D. self-social-evaluation gap

Table 4—Ratings received by self and others

	D16-17		D17-18		D18-19	
	M	F	M	F	M	F
Self	9.83	9.69***	9.80	9.79	NA	NA
Social	8.59	9.14***	9.03	9.38**	9.11	9.39***

Note: Ratings where abstention occurred have been excluded from the analysis. *Self* denotes the rating that an individual assigned to herself, while *Social* represents the average ratings received from peers, excluding the self-assigned rating. Mann-Whitney two-sample statistic significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The initial set of results suggests that gender significantly influences rating behavior. This section compares individuals' self-ratings with social recognition, with a particular emphasis on gender differences. Table 4 reveals notable patterns between genders. Specifically, women consistently receive higher social ratings than men across all datasets, although men tend to give higher self-evaluations (not statistically significant for D17-18). To quantify the discrepancy between self-perception and social recognition, we introduce a measure called the *self-social evaluation gap* (SSEG), defined as:

$$\textit{self-social-evaluation gap} = \textit{self-rating} - \textit{average social-ratings}$$

This index captures the extent to which an individual's self-rating diverges from the average ratings given by others.[9]

Given that academic performance and gender are significant predictors of social ratings, we present a scatter plot of SSEG (self-social-evaluation gap) against exam scores, segmented by gender. Figure 5 reveals several noteworthy aspects of the SSEG index.

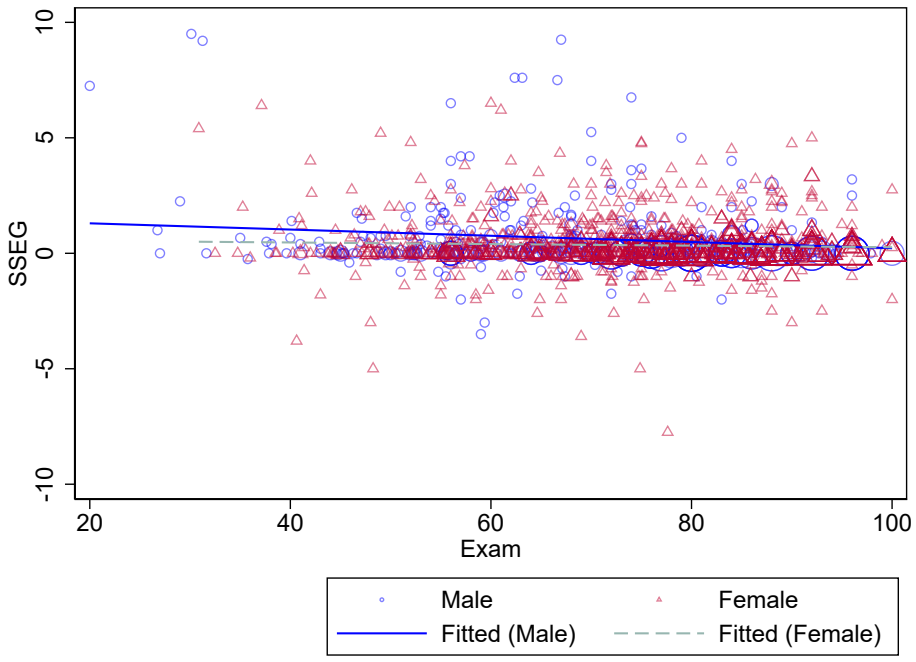
First, there is a general tendency for individuals to rate themselves higher than they are rated by others, as indicated by the higher concentration of points in the upper half of the figure. Second, a substantial number of individuals have an SSEG close to zero, suggesting that their self-assessments align with the evaluations they received from peers. This outcome is primarily influenced by groups without variance in ratings, where all members consistently awarded and received the highest possible scores.[10]

[9] Our SSEG measure involves more complex strategic interactions compared to the discrepancy scores commonly used in the psychology literature on self-enhancement. However, it still aligns with this body of research. For comprehensive reviews on self-enhancement bias and its measurement, see Sedikides and Gregg (2008); Krueger, Heck and Asendorpf (2017). Furthermore, for discussions on the rationale behind the use of social consensus as a reference value, see Larrick, Mannes and Soll (2024); Vazire and Carlson (2011).

[10] For results excluding these zero-variance groups, see Appendix .

Third, the data indicate that men tend to have a higher SSEG than women, a difference that is particularly pronounced among men with lower academic performance. The fitted lines for men and women illustrate this pattern. On average, males exhibit a higher SSEG, driven largely by those with weaker academic performance, and there is a negative correlation between SSEG and academic performance among males. In contrast, the relationship between SSEG and exam performance is much flatter for women, indicating less variation based on academic performance.

FIGURE 5. SSEG, ACADEMIC PERFORMANCE AND GENDER



Note: The size of the circles and triangles is proportional to the observation frequency.

The regression analysis corroborates several patterns observed in our preliminary visual examination. In particular, both male and female participants show positive SSEG, indicating a tendency to overstate their contributions relative to peer evaluations. Additionally, the data reveal that women generally report lower SSEG compared to men. The inverse relationship between exam performance and SSEG is statistically significant, which confirms that higher exam scores are associated with smaller SSEGs. However, the disparity in the slopes of the relationship between SSEG and exam performance across genders is not robustly significant, with only marginal significance detected in one of the samples.

FINDING 3: *(Self-Social Evaluation Gap) There is a significant gender differ-*

Table 5—Determinants of SSEG

Dependent: SSEG	(1) D16-17	(2) D16-17	(3) D17-18	(4) D17-18	(5) Combined	(6) Combined
Exam	-0.00733*	-0.0104	-0.0162***	-0.0323***	-0.00732***	-0.0136**
	(0.00379)	(0.00887)	(0.00468)	(0.0111)	(0.00279)	(0.00631)
Female	-0.381***	-0.748	-0.204	-1.467**	-0.313***	-0.927*
	(0.122)	(0.821)	(0.125)	(0.747)	(0.0915)	(0.491)
Female_Majority	0.240*	0.241*	0.00156	-0.00545	0.0938	0.0942
	(0.142)	(0.142)	(0.149)	(0.149)	(0.0984)	(0.0977)
Female × Exam		0.00489		0.0226*		0.00937
		(0.0103)		(0.0118)		(0.00681)
Constant	1.206***	1.431**	1.472***	2.341***	1.103***	1.504***
	(0.328)	(0.679)	(0.328)	(0.674)	(0.222)	(0.441)
Observations	667	667	685	685	1352	1352

Note: The standard errors are presented in parentheses, and the significance levels are * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Multilevel mixed-effects models with robust standard errors, incorporating random intercepts for individual subjects. Female_Majority is a dummy variable which equals one if the group contains strictly more female students.

ence in the Self-Social Evaluation Gap (SSEG), with females exhibiting a lower SSEG compared to males.

V. Discussion

A. Abstention

All previous analysis is based on data from those who submitted their ratings; however, it is important to note that the abstention rate varies between treatments, particularly among male participants. A key reason for examining abstention is the observed differences in attitudes towards members of the female group in the different treatments. Table 6 builds on the findings presented in Table 3 by disaggregating the data by gender. The results reveal two notable gender-related dynamics. First, the *ReceiverFemale* row indicates that both male and female raters tend to give more favorable ratings to female recipients in Condition Self; however, in Condition No Self, this trend holds only for female raters, as males provide statistically insignificant lower ratings to female recipients. Second, while the receiver's academic performance consistently influences the ratings provided by markers of both genders, the academic performance of male markers does not significantly predict their rating behavior. In contrast, academically stronger female markers tend to be more stringent in Condition Self, but this strictness does not extend to Condition No Self.

This observation prompts an important question: Why do male subjects demon-

Table 6—Determinants of Rating Behaviour, separated by gender

Dependent: Rate	Condition Self						Condition No Self	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	D16-17	D16-17	D17-18	D17-18	Combined	Combined	D18-19	D18-19
Marker gender:	M	F	M	F	M	F	M	F
Self	0.750***	0.578***	0.549***	0.376***	0.651***	0.475***		
	(0.153)	(0.0801)	(0.148)	(0.0678)	(0.106)	(0.0525)		
ReceiverFemale	0.401***	0.374***	0.268*	0.277**	0.337***	0.331***	-0.0225	0.177
	(0.123)	(0.137)	(0.137)	(0.127)	(0.0912)	(0.0935)	(0.196)	(0.112)
MarkerExam	0.00631	-0.0159***	-0.00850	-0.0108**	-0.000662	-0.0132***	-0.00900*	-0.00751
	(0.00667)	(0.00425)	(0.00546)	(0.00509)	(0.00444)	(0.00340)	(0.00487)	(0.00509)
ReceiverExam	0.0120***	0.0206***	0.0169***	0.0189***	0.0143***	0.0197***	0.0266***	0.0258***
	(0.00388)	(0.00405)	(0.00436)	(0.00428)	(0.00294)	(0.00297)	(0.00777)	(0.00652)
D17-18					0.253*	0.379***		
					(0.138)	(0.138)		
Constant	7.686***	8.321***	8.716***	8.553***	8.085***	8.224***	8.378***	7.977***
	(0.603)	(0.459)	(0.386)	(0.442)	(0.397)	(0.355)	(0.511)	(0.550)
Observations	1085	2840	1050	2941	2135	5781	1081	2388

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: The standard errors are presented in parentheses, and the significance levels are * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Multilevel mixed-effects models with robust standard errors, incorporating random intercepts for both groups and individual subjects. Self is a dummy variable that indicates whether the rating is a self-assessment. ReceiverFemale is a dummy variable equal to one if the receiver of the rating is female. MarkerExam stands for the marker's exam performance. ReceiverExam stands for the receiver's exam performance.

strate markedly different attitudes toward their female peers across these two treatments? One plausible explanation is that the variation in treatment directly affects subjects' rating behavior. Since individuals rate themselves in CONDITION SELF, their assessments of others may be influenced by the self-assessment they provide. [11]

Another subtle effect might be the dramatic reduction in the abstention rate among male subjects in CONDITION NO SELF. We argue that abstention reduction is potentially explained by self-assessment avoidance, especially for male subjects. One possible motive for avoidance of self-assessment is self-confidence management. In other words, by avoiding self-evaluation situations, people may find it easier to maintain a positive self-image. Existing evidence on overconfidence and self-evaluation conclusively indicates that men are more overconfident and more likely to adopt commitment tools to maintain confidence. Empirical and experimental evidence has indicated that those with low self-esteem or a negative prior are more likely to avoid feedback information (Fast, Burris and Bartel, 2014; Golman, Hagmann and Loewenstein, 2017). Therefore, we hypothesize that men, especially men with weaker academic performance, are those most likely to avoid self-assessment, and as a consequence, they are more likely to abstain when self-assessment is required (in CONDITION SELF).

As illustrated in Figure 6, the decrease in abstention between treatments is primarily driven by males with weaker academic performance. Students are categorized into four balanced groups based on their exam performance. In addition, both male and female students with poorer academic performance tend to abstain more frequently. There are several possible explanations for why weaker students behave differently. One plausible explanation is that these students may be less concerned about their academic performance and thus adopt this dominated strategy. Alternatively, these academically weaker students might be more strategic, as honest responses may seem less advantageous to them. Finally, the Dunning-Kruger effect could influence these results (Kruger and Dunning, 1999).

The regression results in Table 7 provide insights into the characteristics of those who abstain. First, the rate of abstaining correlates with academic performance, especially for female subjects. Students with higher scores are less likely to abstain. Second, we identify a very significant treatment effect when self-assessment is not required, male subjects abstain less frequently. This effect is not statistically significant for the female sample. Third, we find an overall difference between men and women in abstention, as women are less likely to abstain.

[11] According to the self-deception argument in our model, different rating behaviors are expected across these treatments. For example, an individual concerned with self-assessment may justify overrating themselves by also assigning high ratings to others, thereby signalling that they are applying generous, objective criteria.

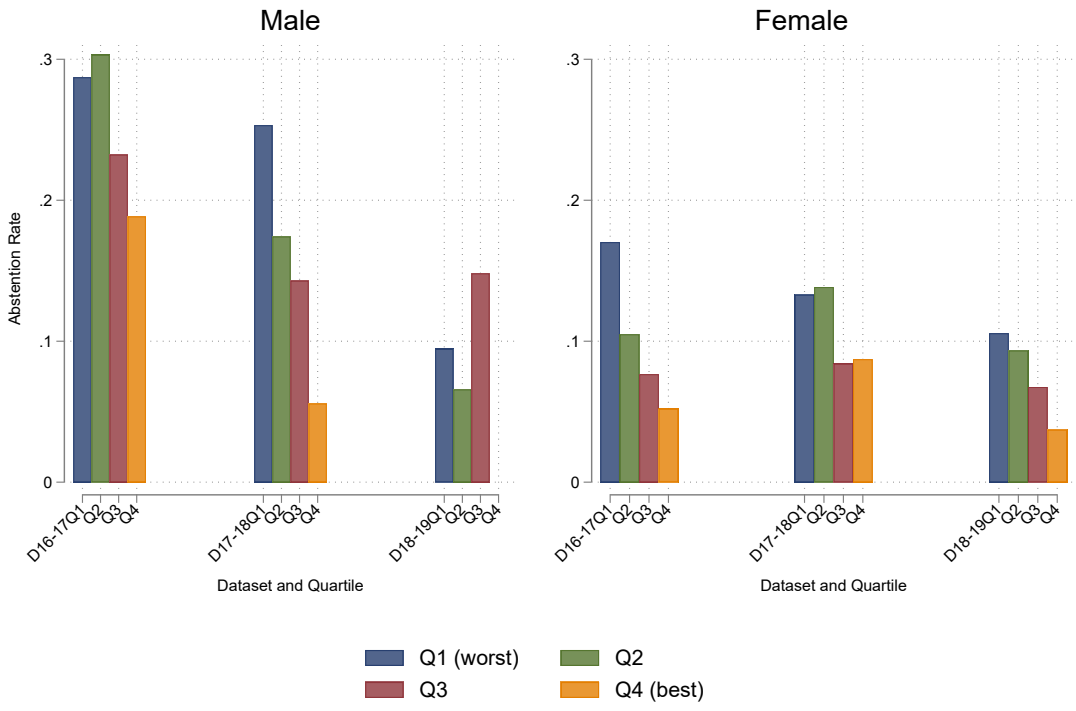


FIGURE 6. ABSTENTION RATE BY ACADEMIC PERFORMANCE

Table 7—Determinants of Abstention

Dependent: Abstention	Condition Self			Condition No Self	All
	(1)	(2)	(3)	(4)	(5)
	D16-17	D17-18	Combined	D18-19	All
Exam	-0.0394***	-0.0472***	-0.0425***	-0.0360***	-0.0370***
	(0.00952)	(0.0129)	(0.00772)	(0.00998)	(0.00602)
Female	-1.015***	-0.205	-0.675***	0.310	-0.693***
	(0.259)	(0.311)	(0.200)	(0.411)	(0.193)
Female_Majority	-0.672	-0.238	-0.491	-0.270	-0.494*
	(0.432)	(0.514)	(0.327)	(0.478)	(0.276)
Dataset			-0.779**		
			(0.314)		
Condition No Self					-1.758***
					(0.421)
Female × Condition No Self					1.061**
					(0.459)
Constant	1.523**	0.296	2.157**	-0.801	0.733
	(0.750)	(0.895)	(0.843)	(0.704)	(0.465)
Observations	794	787	1581	795	2376

Note: Standard errors in parentheses. The significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Why females receive higher ratings

Our findings show that people overrate themselves, but also take contributions into account. A puzzling observation is why women rate others lower, but are rated higher. We propose that it is because women contribute more to the group projects, and subjects partly follow the instructions and social norms to reward contribution.

We have the following supporting evidence to reach our conclusion. In summary, (1) the previous and evidence from the present study suggest that female students have better academic performance due to several desirable characteristics. (2) We find that the students in our sample do not strategically play a Nash strategy but engage in collusion and follow the instruction and social norm to reward the contribution. (3) Female students do not overrate themselves compared to their male counterparts. (4) We identify no evidence of gender discrimination against male students.

First, it is conclusive from the existing literature that female students are academically stronger in almost all the academic assessment criteria measured (Voyer and Voyer, 2014; Sheard, 2009). The explanations include that women have better self-control capacities (Duckworth and Seligman, 2006), self-regulation (Matthews, Ponitz and Morrison, 2009), and commitment (Sheard, 2009). These attributes are desirable for the completion of the academic assignment. In our

sample, females have a significant advantage in terms of exam performance, as shown in Table 1[12].

Second, as shown in the scatter plot, the refined Nash prediction that rating others 0 and self 10 is almost never played. We observe a high frequency of collusion but also rewarding contribution, as presented in Table 3, since the receiver's exam performance is a highly significant predictor of the ratings across all subsamples.

An alternative explanation is that female subjects are more strategic and self-serving, so they rate themselves higher and others lower. To validate this argument, we focus on the self-other bias of male subjects and female subjects. If such a statement is true, we would expect to observe that females show a greater bias towards rating themselves compared to what male subjects represent. Table 8 presents the regression outcome with the interaction term of the gender of the rater and the dummy variable *self*. For both D16-17 and D17-18, the coefficients of the interaction term are negative. Therefore, although statistically insignificant, females, on average, overrate themselves less. Therefore, we reject the hypothesis that women are more self-serving than their male counterparts.

Another alternative explanation is discrimination. Specifically, our results could be driven by the fact that male students, a minority group of approximately 1/3 of the course, could be negatively discriminated against. We cautiously reject this hypothesis for several reasons. First, the groups are formed endogenously, and thus the chance of observing taste-based discrimination is minimal, since those with a strong preference for working with only females/males could self-select themselves into single-gender groups.

Second, we analyse a joint evaluation approach.[13] Such an "evaluation nudge" may overcome gender bias in the evaluation process (Bohnet, Van Geen and Bazerman, 2016). Third, the group project requires intense interactions and regular group meetings. Therefore, from an information perspective, the subjects had enough opportunities to learn the contributions of their group members and update their beliefs. Recent research on discrimination in evaluation suggests that initial biases towards a specific gender can be mitigated by additional signals (Coffman and Klinowski, 2024).

Fourth, we investigate how ratings depend on the group's gender composition. Although empirical evidence to date has not been conclusive on whether the evaluators prefer candidates of the same gender or the opposite, most existing evidence finds that the gender composition of the evaluators influences the rating behavior (for a review, see (Bagues, Sylos-Labini and Zinovyeva, 2017)). Therefore, if gender-based discrimination plays an important role, we expect to observe that group composition would affect the rating behavior in our sample. To test this conjecture, we present the regression results (columns 3, 4, and 5 of Table 8),

[12]Mann-Whitney two-sample statistic $p < 0.001$ for all three datasets.

[13]Specifically in our design, subjects must submit their ratings for all group members (including themselves if requested) at the same time. Partial ratings, "Save for later" options, and multiple submissions were not allowed.

including the interaction term of gender and gender composition in the ratings. We are unable to identify the effect of gender composition on ratings. In short, we cautiously reject the hypothesis that discrimination is the explanation.

Table 8—Gender on self and gender composition

Dependent: Rate	(1) D16-17	(2) D17-18	(3) D16-17	(4) D17-18	(5) D18-19
Group Mark	0.0136*	0.0133	0.0138*	0.0130	-0.0164
	(0.00747)	(0.01000)	(0.00758)	(0.0100)	(0.0116)
Rater Female	-0.282***	-0.112	-0.295***	-0.142	-0.269***
	(0.100)	(0.0971)	(0.0991)	(0.0967)	(0.0991)
Receiver Female	0.354***	0.213***	0.251***	0.247***	-0.0207
	(0.0566)	(0.0546)	(0.0799)	(0.0844)	(0.103)
Rater Exam	-0.0105***	-0.0139***	-0.0105***	-0.0139***	-0.0107***
	(0.00332)	(0.00314)	(0.00332)	(0.00314)	(0.00314)
Receiver Exam	0.0196***	0.0220***	0.0197***	0.0220***	0.0278***
	(0.00169)	(0.00164)	(0.00169)	(0.00164)	(0.00194)
Self	0.683***	0.544***	0.624***	0.426***	
	(0.114)	(0.106)	(0.0567)	(0.0513)	
MFemale	-0.282***	-0.112	-0.295***	-0.142	-0.269***
	(0.100)	(0.0971)	(0.0991)	(0.0967)	(0.0991)
Self × MFemale	-0.0833	-0.161			
	(0.138)	(0.127)			
Major_Female			-0.109	0.102	0.307*
			(0.195)	(0.185)	(0.167)
RFemaler × Major_Female			0.156	-0.0963	0.0962
			(0.105)	(0.104)	(0.125)
Constant	7.325***	7.893***	7.379***	7.886***	9.372***
	(0.568)	(0.675)	(0.568)	(0.681)	(0.801)
Observations	3961	4016	3961	4016	3479

Note: Multilevel mixed-effects models using random intercepts for groups and individual subjects. RFemale is a dummy variable equal to one if the receiver of the rating is female. MFemale is a dummy variable equal to one if the marker is female. MExam stands for the marker's exam performance. RExam stands for the receiver's exam performance. Self is a dummy variable equal to one if the rate is self-rating. Major_Female is a dummy variable equal to one if the group has more females than males. Standard errors are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

B. Gender difference in self-deception?

Table 9—Determinants of egalitarian rating behaviour

				equalrate		
	(D16-17)	(D17-18)	(Combined)	(D16-17 w.o. absent)	(D17-18 w.o.absent)	(Combined w.o.absent)
Exam	-0.0328***	-0.0308***	-0.0310***	-0.0265***	-0.0256***	-0.0277***
	(0.00900)	(0.00899)	(0.00544)	(0.00994)	(0.00966)	(0.00612)
Female	-1.210***	-0.459*	-0.874***	-1.005***	-0.455	-0.731***
	(0.236)	(0.268)	(0.179)	(0.279)	(0.293)	(0.203)
Female_majority	0.229	0.567	0.372	0.576	0.657	0.630**
	(0.320)	(0.363)	(0.237)	(0.424)	(0.431)	(0.300)
Constant	4.325***	3.266***	3.807***	3.106***	2.690***	3.012***
	(0.753)	(0.591)	(0.433)	(0.859)	(0.658)	(0.498)
var(_cons[groupid])	2.128***	1.755***	1.945***	3.699***	2.567***	3.144***
	(0.502)	(0.512)	(0.358)	(0.880)	(0.759)	(0.579)
Observations	794	787	1581	672	689	1361

Note: The standard errors are presented in parentheses, and the significance levels are * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 10—Determinants of Rating Behaviour

				equalrate					
	(D16)	(D17)	(Comb)	(D16 w.o. a)	(D17 w.o.a)	(C w.o.a)	(D16 w.o. az)	(D17 w.o. a z)	(C w.o. a z)
mexam	-0.0328***	-0.0308***	-0.0310***	-0.0265***	-0.0256***	-0.0277***	-0.0248***	-0.0261***	-0.0256***
	(0.00900)	(0.00899)	(0.00544)	(0.00994)	(0.00966)	(0.00612)	(0.00943)	(0.00930)	(0.00548)
Gender	-1.210***	-0.459*	-0.874***	-1.005***	-0.455	-0.731***	-0.948***	-0.451	-0.704***
	(0.236)	(0.268)	(0.179)	(0.279)	(0.293)	(0.203)	(0.269)	(0.284)	(0.196)
major_female	0.229	0.567	0.372	0.576	0.657	0.630**	0.459	0.475	0.461**
	(0.320)	(0.363)	(0.237)	(0.424)	(0.431)	(0.300)	(0.322)	(0.333)	(0.230)
Constant	4.325***	3.266***	3.807***	3.106***	2.690***	3.012***	2.231***	1.956***	2.130***
	(0.753)	(0.591)	(0.433)	(0.859)	(0.658)	(0.498)	(0.758)	(0.608)	(0.421)
var(_cons[groupid])	2.128***	1.755***	1.945***	3.699***	2.567***	3.144***	1.335***	0.778**	1.048***
	(0.502)	(0.512)	(0.358)	(0.880)	(0.759)	(0.579)	(0.366)	(0.326)	(0.243)
Observations	794	787	1581	672	689	1361	502	471	973

Note:

VI. Conclusion

Our study provides robust evidence of self-serving bias in the context of peer evaluations, with significant implications for understanding gender dynamics in these contexts. Both male and female students exhibit this bias, although the extent varies slightly, with males showing a marginally higher tendency to overrate themselves. The lower self-social evaluation gap observed among female students, coupled with their higher social recognition, suggests that women may be less inclined to overstate their contributions in a group. This discrepancy in self-assessment behavior could reflect broader societal patterns of self-perception and confidence, which are often influenced by gender norms and expectations.

The analysis also highlights the role of academic performance in influencing both self- and peer-evaluations, as well as the propensity to abstain from rating. Males with weaker academic performance are particularly prone to abstain when self-assessment is required, suggesting that these students might be strategically avoiding situations where their self-image could be negatively impacted. Our findings have real-world implications, particularly in settings where self-assessments are increasingly used for performance evaluations and promotions. If self-assessment practices are widely adopted, the tendency of females to underrate themselves relative to their male counterparts could reinforce existing gender disparities in career advancement and recognition. This underscores the need for institutions to carefully consider the design of assessment systems to ensure that they do not inadvertently disadvantage women, potentially exacerbating gender inequality in the workplace.

REFERENCES

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for truth-telling." *Econometrica*, 87(4): 1115–1153. Publisher: Wiley Online Library.

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva. 2017. "Does the gender composition of scientific committees matter?" *American Economic Review*, 107(4): 1207–1238.

Bohl, Don L. 1996. "Minisurvey: 360-degree appraisals yield superior results, survey stows." *Compensation and Benefits Review*, 28(5): 16–19. Publisher: SAGE Publications Inc.

Bohnet, Iris, Alexandra Van Geen, and Max Bazerman. 2016. "When performance trumps gender bias: joint vs. separate evaluation." *Management Science*, 62(5): 1225–1234. Publisher: INFORMS.

Capraro, Valerio. 2018. "Gender differences in lying in sender-receiver games: a meta-analysis." *Judgment and Decision Making*, 13(4): 345–355. Publisher: Cambridge University Press.

Carpenter, Jeffrey, and Peter Hans Matthews. 2009. "What norms trigger punishment?" *Experimental Economics*, 12: 272–288. Publisher: Springer.

Coates, Dennis E. 1998. "Don't tie 360 feedback to pay." *Training*, 35: 68–78.

Coffman, Katherine, and David Klinowski. 2024. "Gender and Preferences for Performance Feedback." *Management Science*, mnsc.2023.02482.

Dana, Jason, Roberto A Weber, and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory*, 33: 67–80. Publisher: Springer.

Deb, Rahul, and Ludovic Renou. 2022. *Which wage distributions are consistent with statistical discrimination?* University of Toronto, Department of Economics.

DeNisi, Angelo S., and Avraham N. Kluger. 2000. "Feedback effectiveness: can 360-degree appraisals be improved?" *Academy of Management Perspectives*, 14(1): 129–139. Publisher: Academy of Management.

Dreber, Anna, and Magnus Johannesson. 2008. "Gender differences in deception." *Economics Letters*, 99(1): 197–199.

Duckworth, Angela Lee, and Martin EP Seligman. 2006. "Self-discipline gives girls the edge: gender in self-discipline, grades, and achievement test scores." *Journal of Educational Psychology*, 98(1): 198. Publisher: American Psychological Association.

Edwards, Mark R., and Ann J. Ewen. 1996. "How to manage performance and pay with 360-degree feedback: multisource assessment can work for both performance and pay management when participants know the system is fair. But doing it right requires a commitment." *Compensation and Benefits Review*, 28(3): 41–46. Publisher: SAGE Publications Inc.

Englmaier, Florian, Stefan Grimm, Dominik Grothe, David Schindler, and Simeon Schudy. 2024. "The effect of incentives in nonroutine analytical team tasks." *Journal of Political Economy*, 132(8): 2695–2747. Publisher: The University of Chicago Press.

Erat, Sanjiv, and U. Gneezy. 2010. "White lies." *Management Science*, 58: 723–733.

Espey, Molly. 2022. "Gender and peer evaluations." *Journal of Economic Education*, 53(1): 1–10. Publisher: Routledge _eprint: https://doi.org/10.1080/00220485.2021.2004277.

Exley, Christine L, and Judd B Kessler. 2022. "The gender gap in self-promotion." *Quarterly Journal of Economics*, 137: 1345–1381.

Fast, Nathanael J, Ethan R Burris, and Caroline A Bartel. 2014. "Managing to stay in the dark: managerial self-efficacy, ego defensiveness, and the aversion to employee voice." *Academy of Management Journal*, 57(4): 1013–1034. Publisher: Academy of Management Briarcliff Manor, NY.

Fischer, Christian, and Hans-Theo Normann. 2019. "Collusion and bargaining in asymmetric cournot duopoly—an experiment." *European Economic Review*, 111: 360–379.

Gneezy, Uri, Agne Kajackaite, and Joel Sobel. 2018. "Lying aversion and the size of the lie." *American Economic Review*, 108(2): 419–453. Publisher: American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Gneezy, Uri, Christina Gravert, Silvia Saccardo, and Franziska Tausch. 2017. "A must lie situation–avoiding giving negative feedback." *Games and Economic Behavior*, 102: 445–454. Publisher: Elsevier.

Golman, Russell, David Hagmann, and George Loewenstein. 2017. "Information avoidance." *Journal of Economic Literature*, 55(1): 96–135.

Greene, Kenneth V, and Phillip J Nelson. 2002. "If extremists vote how do they express themselves? An empirical test of an expressive theory of voting." *Public Choice*, 113(3): 425–436. Publisher: Springer.

Houser, Daniel, Stefan Vetter, and Joachim Winter. 2012. "Fairness and cheating." *European Economic Review*, 56(8): 1645–1655.

Jaroszewicz, Ania, George Loewenstein, and Roland Benabou. 2024. "Offering, asking, consenting, and rejecting: the psychology of helping interactions."

Krueger, Joachim I., Patrick R. Heck, and Jens B. Asendorpf. 2017. "Self-enhancement: Conceptualization and Assessment." *Collabra: Psychology*, 3(1): 28.

Kruger, Justin, and David Dunning. 1999. "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments." *Journal of Personality and Social Psychology*, 77(6): 1121–1134. Place: US Publisher: American Psychological Association.

Larrick, Richard P., Albert E. Mannes, and Jack B. Soll. 2024. "The Social Psychology of the Wisdom of Crowds (with a New Section on Recent Advances)." In *Behavioral Decision Analysis*. Vol. 350, , ed. Florian M. Federspiel, Gilberto Montibeller and Matthias Seifert, 121–143. Cham:Springer International Publishing.

Linton, Oliver, Esfandiar Maasoumi, and Yoon-Jae Whang. 2005. "Consistent testing for stochastic dominance under general sampling schemes." *Review of Economic Studies*, 72(3): 735–765.

Matthews, Jamaal S, Claire Cameron Ponitz, and Frederick J Morrison. 2009. "Early gender differences in self-regulation and academic achievement." *Journal of Educational Psychology*, 101(3): 689. Publisher: American Psychological Association.

Michailidou, Georgia, and Valentina Rotondi. 2019. "I'd lie for you." *European Economic Review*, 118: 181–192. Publisher: Elsevier.

Morgan, John, Susanne Neckermann, and Dana Sisak. 2021. "Peer evaluation and team performance: an experiment on complex problem solving." Downloaded.

Prendergast, Canice. 1999. "The provision of incentives in firms." *Journal of Economic Literature*, 37(1): 7–63. Publisher: American Economic Association.

Ramm, Joachim, Sigve Tjotta, and Gaute Torsvik. 2013. "Incentives and Creativity in Groups." *SSRN Electronic Journal*.

Sedikides, Constantine, and Aiden P. Gregg. 2008. "Self-Enhancement: Food for Thought." *Perspectives on Psychological Science*, 3(2): 102–116.

Sheard, Michael. 2009. "Hardiness commitment, gender, and age differentiate university academic performance." *British Journal of Educational Psychology*, 79(1): 189–204. Publisher: Wiley Online Library.

Spiekermann, Kai P., and Arne Weiss. 2016. "Objective and subjective compliance: a norm-based explanation of 'moral wiggle room'." *Games and Economic Behavior*, 96: 170–183.

Vazire, Simine, and Erika N. Carlson. 2011. "Others Sometimes Know Us Better Than We Know Ourselves." *Current Directions in Psychological Science*, 20(2): 104–108.

Voyer, Daniel, and Susan D Voyer. 2014. "Gender differences in scholastic achievement: a meta-analysis." *Psychological Bulletin*, 140(4): 1174. Publisher: American Psychological Association.

Winterbotham, Mark, David Vivian, Genna Kik, Jessica Huntley, Mark Tweddle, Christabel Downing, Dominic Thomson, Naomi Morrice, and Sam Stroud. 2018. "Employer skills survey 2017." Department of Education, UK.

Appendix

A1. Proof of Lemma III.1

PROOF:

Consider player j. Let π_j denote player j's relative rating when she abstains from this peer review.

a) Consider the Condition Self case, i.e., when players must give ratings to themselves as well as to their groupmates in peer review. If $\pi_j = 1$ when player j abstains, then she will be indifferent between abstention and participating and giving herself 10 and all her groupmates 0, as the latter strategy will keep $\pi_i = 1$.
Instead, suppose that $\pi_j < 1$ if player j abstains. That $\pi_j < 1$ when player j abstains implies that at least one of player j's groupmates participates in the peer review and at least one of them gives player j a rating that is strictly lower than 10. That is, $r_{i,j} \leq 10$ for all $i \in I_j$ with at least one strict inequality. Moreover, $\pi_j < 1$ means that some other group member, say, player k, receives higher rating(s) than player j does when player j abstains. Let I_k denote the set of group members giving ratings to player k and I_j denote that for player j when player j abstains. Then,

$$\pi_j = \frac{\sum\limits_{i \in I_j} r_{i,j} / \#I_j}{\sum\limits_{i \in I_k} r_{i,j} / \#I_k} < 1.$$

If player j participates in the peer review and gives 10 to herself and 0 to all her groupmates, then π_j becomes

$$\pi'_j = \frac{\left(\sum\limits_{i \in I_j} r_{i,j} + 10 \right) / \left(\#I_j + 1 \right)}{\left(\sum\limits_{i \in I_k} r_{i,j} + 0 \right) / \left(\#I_k + 1 \right)} = \frac{\left(\sum\limits_{i \in I_j} r_{i,j} + 10 \right) / \left(\#I_j + 1 \right)}{\sum\limits_{i \in I_k} r_{i,j} / \left(\#I_k + 1 \right)}.$$

As $r_{i,j} \leq 10$ for all $i \in I_j$ with at least one strict inequality, $\left(\sum\limits_{i \in I_j} r_{i,j} + 10 \right) / \left(\#I_j + 1 \right)$

is strictly greater than $\sum_{i\in I_j} r_{i,j}/\#I_j$:

$$\begin{aligned}
\frac{\sum_{i\in I_j} r_{i,j}+10}{\#I_j+1}-\frac{\sum_{i\in I_j} r_{i,j}}{\#I_j} &= \frac{\left(\sum_{i\in I_j} r_{i,j}+10\right)\cdot \#I_j - \sum_{i\in I_j} r_{i,j}\cdot\left(\#I_j+1\right)}{\left(\#I_j+1\right)\cdot \#I_j} \\
&= \frac{\sum_{i\in I_j} r_{i,j}\cdot \#I_j + 10\cdot \#I_j - \sum_{i\in I_j} r_{i,j}\cdot \#I_j - \sum_{i\in I_j} r_{i,j}}{\left(\#I_j+1\right)\cdot \#I_j} \\
&= \frac{10\cdot \#I_j - \sum_{i\in I_j} r_{i,j}}{\left(\#I_j+1\right)\cdot \#I_j} \\
&> \frac{10\cdot \#I_j - \sum_{i\in I_j} 10}{\left(\#I_j+1\right)\cdot \#I_j} \\
&= \frac{10\cdot \#I_j - 10\cdot \#I_j}{\left(\#I_j+1\right)\cdot \#I_j} \\
&= 0.
\end{aligned}$$

Thus, $\pi'_j = \frac{\left(\sum_{i\in I_j} r_{i,j}+10\right)/(\#I_j+1)}{\sum_{i\in I_k} r_{i,j}/(\#I_k+1)} > \frac{\sum_{i\in I_j} r_{i,j}/\#I_j}{\sum_{i\in I_k} r_{i,j}/\#I_k} = \pi_j$, which means that player j strictly benefits from switching from abstention to participating and giving herself 10 and all her groupmates 0.
From above we can see that in the CONDITION SELF case, abstention is weakly dominated by participating and giving herself 10 and all her groupmates 0.

b) Consider the Condition No Self case, i.e., when players give ratings only to their groupmates in the peer review.
If $\pi_j = 1$ when player j abstains, then she will be indifferent between abstention and participating and giving all her groupmates 0, as the latter strategy will keep $\pi_i = 1$.
Instead, suppose that $\pi_j < 1$ if player j abstains. That $\pi_j < 1$ when player j abstains implies that at least one of player j's groupmates participates in the peer review and at least one of them gives player j a rating that is strictly lower than 10. That is, $r_{i,j} \leq 10$ for all $i \in I_j$ with at least one strict inequality. Moreover, $\pi_j < 1$ means that some other group member, say, player k, receives higher rating(s) than player j does when player j abstains. Let I_k denote the set of group members giving ratings to player k and I_j denote that for player j when player j abstains. Then, $\#I_k = \#I_j$ if player k also abstains and $\#I_k = \#I_j - 1$ if player k participates in the peer review.

Then,

$$\pi_j = \frac{\sum\limits_{i \in I_j} r_{i,j}/\#I_j}{\sum\limits_{i \in I_k} r_{i,j}/\#I_k} < 1.$$

If player j participates in the peer review and gives 10 to herself and 0 to all her groupmates, then π_j becomes

$$\pi'_j = \frac{\sum\limits_{i \in I_j} r_{i,j}/\left(\#I_j + 1\right)}{\left(\sum\limits_{i \in I_k} r_{i,j} + 0\right)/\left(\#I_k + 1\right)} = \frac{\sum\limits_{i \in I_j} r_{i,j}/\left(\#I_j + 1\right)}{\sum\limits_{i \in I_k} r_{i,j}/\left(\#I_k + 1\right)}.$$

As

$$\frac{\pi'_j}{\pi_j} = \frac{\sum\limits_{i \in I_j} r_{i,j}/\left(\#I_j + 1\right)}{\sum\limits_{i \in I_k} r_{i,j}/\left(\#I_k + 1\right)} \cdot \frac{\sum\limits_{i \in I_k} r_{i,k}/\#I_k}{\sum\limits_{i \in I_j} r_{i,j}/\#I_j} = \frac{\#I_j \cdot \#I_k + \#I_j}{\#I_j \cdot \#I_k + \#I_k},$$

π'_j equals π_j if $\#I_j = \#I_k$ and is greater than π_j if $\#I_j > \#I_k$, that is, $\pi'_j = \pi_j$ if player k abstains and $\pi'_j > \pi_j$ if player k participates in peer review.
From above we can see that in the CONDITION SELF case, abstention is weakly dominated by participating and giving all her groupmates 0.

A2. The theoretical models with details

In this part of the appendix, we show the individual decision-making process when the observation of $s_{i,j}$ is realised.

To determine when it is optimal for an agent to opt for excuse-based lying ($I_i = 1$), we analyse the utility function defined as:

$$U_i = \max_{r_{i,j}, I_i} \sum_j (r_{i,i} - r_{i,j}) - (1 - I_i) \sum_j [(|r_{i,j} - s_{i,j}| + 1)^{\theta_i} - 1] - I_i \gamma \tag{A1}$$

Given the constraints:

- If $I_i = 1$, then $s_{i,j} = \widetilde{s_{i,j}} = c \ \forall \ j$,
- If $I_i = 1$, then $r_{i,j} = \widetilde{s_{i,j}} \ \forall \ j$.

When $I_i = 1$:

$$U_i(I_i = 1) = -\gamma \tag{A2}$$

When $I_i = 0$:

$$U_i(I_i = 0) = \sum_j (r_{i,i} - r_{i,j}) - \sum_j [(|r_{i,j} - s_{i,j}| + 1)^{\theta_i} - 1] \tag{A3}$$

To find when $I_i = 1$ is optimal, we set:

$$U_i(I_i = 1) > U_i(I_i = 0) \tag{A4}$$

Simplifying the inequality:

$$-\gamma > \sum_j (r_{i,i} - r_{i,j}) - \sum_j [(|r_{i,j} - s_{i,j}| + 1)^{\theta_i} - 1] \tag{A5}$$

The condition under which it is optimal to opt for $I_i = 1$ is:

$$\gamma < \sum_j [(|r_{i,j} - s_{i,j}| + 1)^{\theta_i} - 1] - \sum_j (r_{i,i} - r_{i,j}) \tag{A6}$$

This inequality indicates that an individual self-deceives if the fixed cost γ is less than the variable cost of standard lying minus the net material payoff difference between the self-ratings and the ratings to others.

A3. Additional simulation results

In this part of the appendix, we provide some additional simulation results focusing on the impact of the cost of self-deception and the distribution of actual contribution. The simulation results for Condition No Self are similar and available upon request.

Cost of self-deception γ

What would happen if people had different self-deception costs? Intuitively, if this cost γ is higher (lower), we would expect more (less) frequent observations. The following simulation results confirm our expectations.

Dispersion of the actual distribution $s_{i,j}$

What would happen if the distribution of the actual contribution changes? Intuitively, the optimal rating strategy of people varies with their actual observations. For example, when they observe more egalitarian contributions, they have fewer incentives for self-deception.

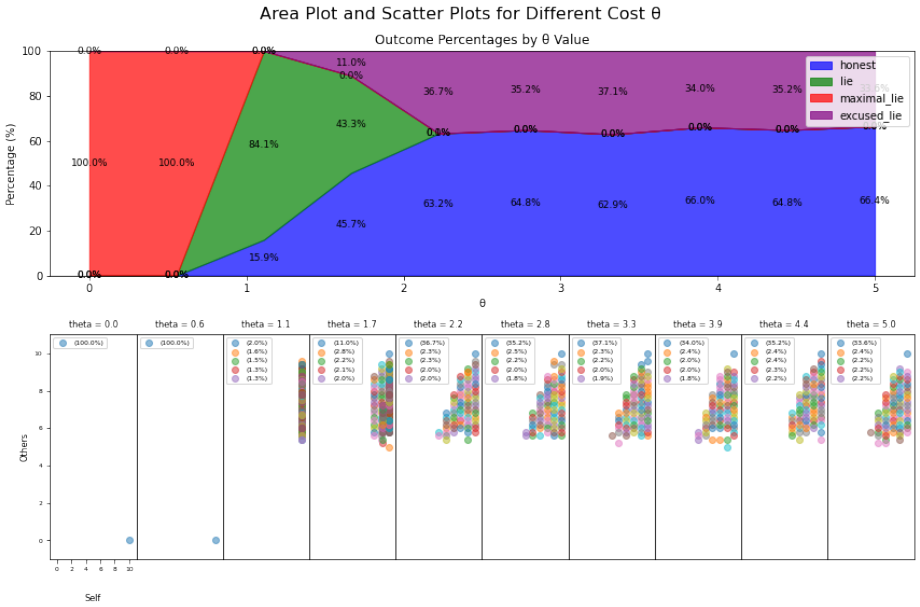


FIGURE A1. LOW γ AREA AND SCATTER PLOTS.

Note: θ ranges from $[0, 5]$, $\gamma = 5$ and $s_{i,j} \sim U(5, 10)$. For the scatter plots, the legend details the five most frequent occurrences.

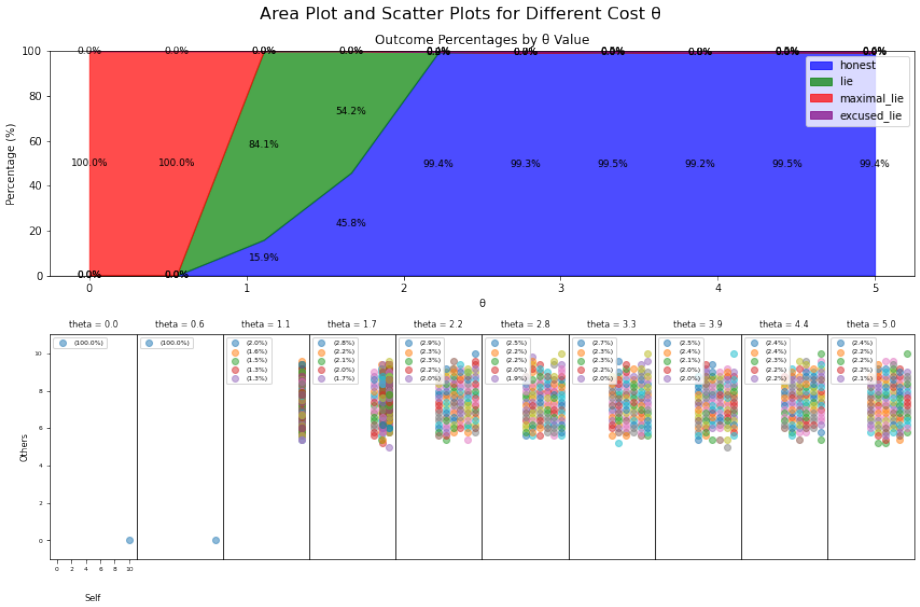


FIGURE A2. HIGH γ AREA AND SCATTER PLOTS.

Note: θ ranges from $[0, 5]$, $\gamma = 20$ and $s_{i,j} \sim U(5, 10)$. For the scatter plots, the legend details the five most frequent occurrences.

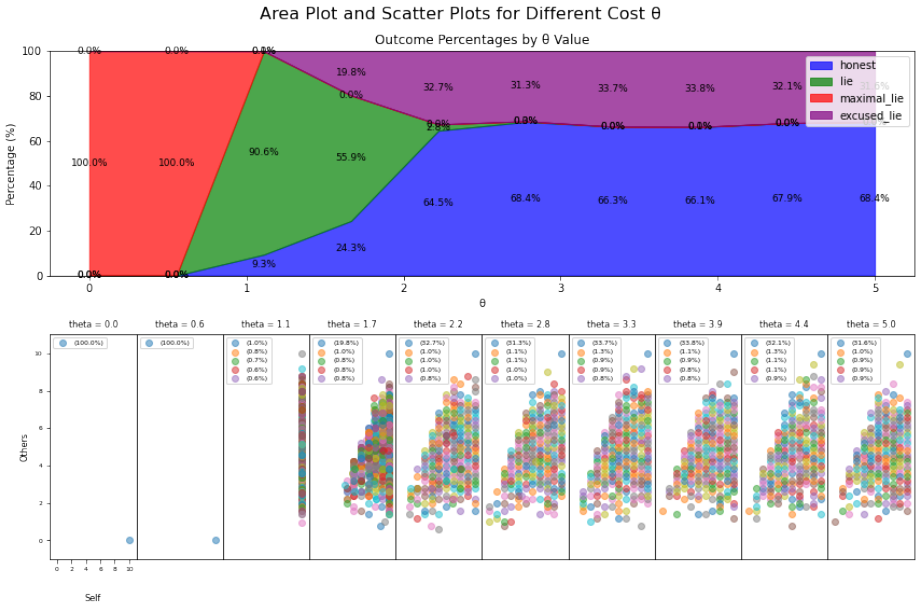


FIGURE A3. DISPERSED $s_{i,j}$ AREA AND SCATTER PLOTS.

Note: θ ranges from $[0, 5]$, $\gamma = 10$ and $s_{i,j} \sim U(0, 10)$. For the scatter plots, the legend details the five most frequent occurrences.

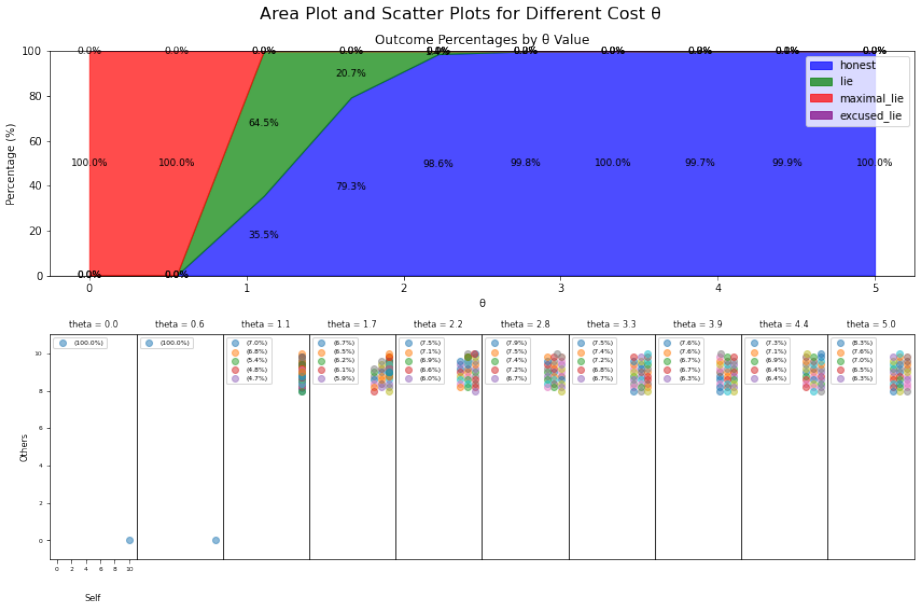


FIGURE A4. CONDENSED $s_{i,j}$ AREA AND SCATTER PLOTS.

Note: θ ranges from $[0, 5]$, $\gamma = 10$ and $s_{i,j} \sim U(8, 10)$. For the scatter plots, the legend details the five most frequent occurrences.

A4. Did the rating mechanism affect group formation?

Conditional on the endogeneity of our group formation, one natural question is: Did students form groups differently facing these two mechanisms? Based on several observable characteristics, our answer is no. Exam performance, gender composition, and group size are the variables of interest. For the former two variables, we focus on the variance of each group to examine whether assortative matching is more likely in one treatment.

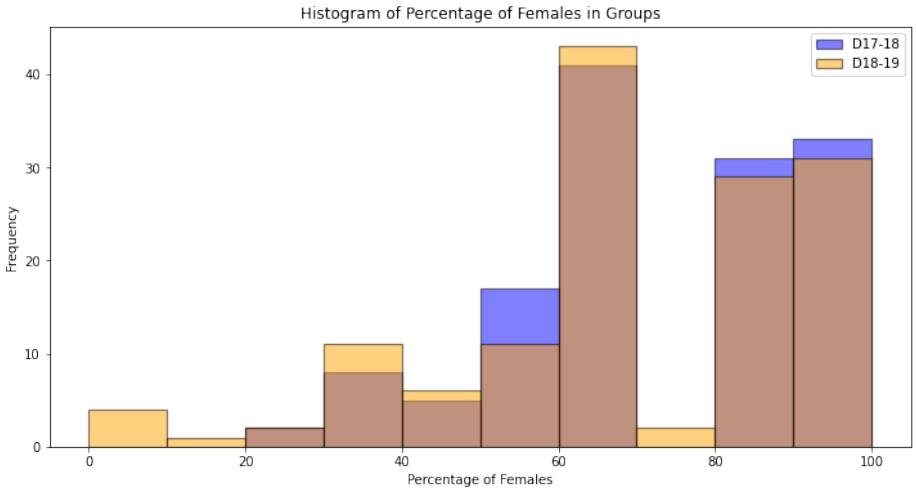


FIGURE A5. GROUP GENDER COMPOSITION

The results of the Levene test indicate that there are no statistically significant differences in the variances of the gender composition ($p = 0.911$), the midterm exam scores ($p = 0.633$) and the final exam scores ($p = 0.286$) between the groups in data set 2 and 3. The Mann-Whitney U test on the group size is also not significant ($p = 0.192$).

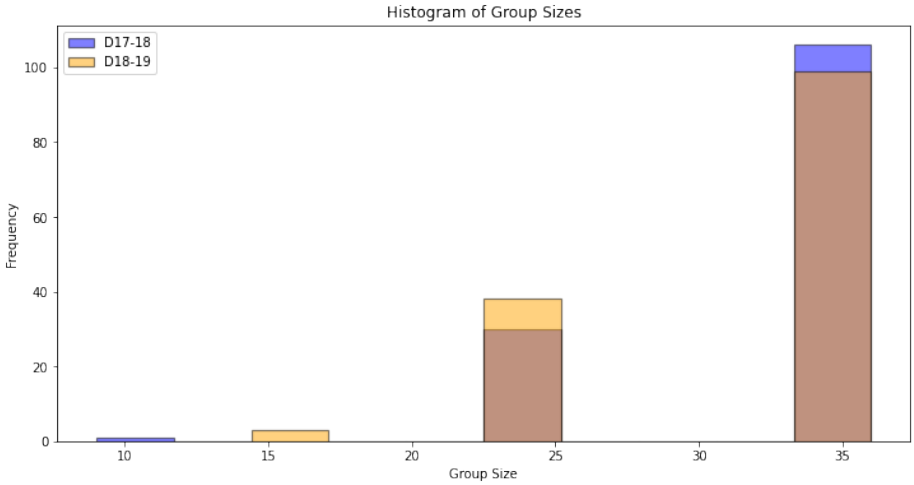


FIGURE A6. GROUP SIZES

A5. Detailed information with a focus on gender

Figure A7 reproduces the findings presented in Figure 2, with the data segmented by gender. Both male and female subjects exhibit comparable rating patterns, despite a higher likelihood of abstention among males.

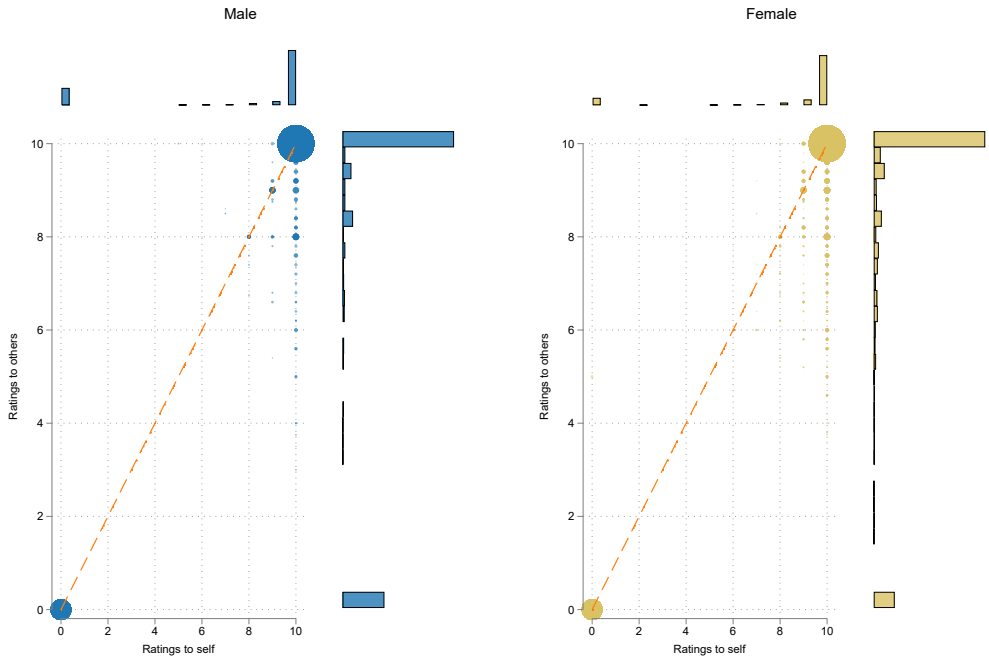


FIGURE A7. RATINGS TO SELF AND TO OTHERS

Note: The diameter of the bubbles is directly proportional to the frequency of the observation. The dashed line represents the line of equality at a 45-degree angle. The histograms displayed along the x- and y-axes illustrate the distribution of observations along each axis. The point located at coordinates (0, 0) indicates individuals who did not provide a response.

WHAT DETERMINES THE GROUP ASSIGNMENT SCORE?

In Table A1, we report a strong positive correlation between average exam performance and coursework scores.

Table A1—Group assignment and exam performance

Dependent: *Assignment* score	(1) D16-17	(2) D17-18	(3) D18-19	(4) Combined
Average exam	0.257*	0.399***	0.185***	0.283***
	(0.137)	(0.0762)	(0.0645)	(0.0557)
Group size	1.972	-0.383	-0.0301	0.894
	(1.634)	(1.344)	(0.884)	(0.779)
Female majority	3.737*	-1.477	-0.460	0.945
	(2.053)	(1.439)	(1.088)	(0.935)
D17-18				4.368***
				(1.427)
D18-19				5.908***
				(1.449)
Constant	33.94***	46.95***	57.84***	40.04***
	(12.82)	(9.006)	(5.943)	(5.934)
Observations	138	137	140	415

Note: D17-18 and D18-19 are dummy variables with the omitted category of D16-17 as 0. Standard errors are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

A6. Results excluding zero-variance groups

In this part of the appendix, we replicate important regressions excluding zero-variance groups. Compared to the full data results, excluding zero-variance groups can eliminate the impact of collusive groups (Observations from the collusive groups are noisy since ratings are uniform and independent of any potential explanatory variables).

Table A2 replicates the results presented in Table 3.

Table A2—Determinants of Rating Behaviour Excluding Zero-Variance Groups

Dependent: Rate	Condition Self						Condition No Self
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	D16-17	D16-17	D17-18	D17-18	Combined	Combined	D18-19
Self	0.835***	1.032***	0.620***	0.784***	0.731***	0.913***	
	(0.0856)	(0.175)	(0.0768)	(0.179)	(0.0578)	(0.125)	
MarkerFemale	-0.381***	-0.607***	-0.234*	-0.424**	-0.309***	-0.519***	-0.314***
	(0.131)	(0.168)	(0.139)	(0.188)	(0.0959)	(0.126)	(0.122)
ReceiverFemale	0.431***	0.473***	0.328**	0.365**	0.382***	0.422***	0.155
	(0.125)	(0.144)	(0.139)	(0.160)	(0.0921)	(0.106)	(0.127)
MarkerExam	-0.0120**	-0.0120***	-0.0156***	-0.0156***	-0.0140***	-0.0140***	-0.0124***
	(0.00467)	(0.00466)	(0.00538)	(0.00538)	(0.00357)	(0.00356)	(0.00470)
ReceiverExam	0.0221***	0.0221***	0.0260***	0.0260***	0.0241***	0.0241***	0.0321***
	(0.00427)	(0.00426)	(0.00515)	(0.00515)	(0.00334)	(0.00333)	(0.00637)
Self × MarkerFemale		0.270		0.226		0.250*	
		(0.201)		(0.214)		(0.146)	
D17-18					0.406***	0.404***	
					(0.140)	(0.140)	
Constant	7.896***	7.838***	8.296***	8.243***	7.902***	7.847***	7.981***
	(0.438)	(0.458)	(0.430)	(0.455)	(0.339)	(0.354)	(0.514)
Observations	2937	2937	2742	2742	5679	5679	2624

Note: Standard errors are presented in parentheses at the following significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Multilevel mixed-effects models with robust standard errors, incorporating random intercepts for both groups and individual subjects. The categorical variables D17-18 and D18-19 represent different datasets, with D16-17 as the reference category. Self is a dummy variable that indicates whether the rating is a self-assessment. MarkerFemale is a dummy variable equal to one if the marker is female. ReceiverFemale is a dummy variable equal to one if the receiver of the rating is female. MarkerExam stands for the marker's exam performance. ReceiverExam stands for the receiver's exam performance. Self × MarkerFemale denotes the interaction between Self and MarkerFemale.

Table A3 presents a replication of the findings from Table 4. The findings are consistent with the full sample. As expected, the results also indicate that gender differences become more pronounced in social recognition when excluding groups that may have engaged in collusion.

Figure A8 replicates the scatter plot of SSEG over exam scores in Figure 5.

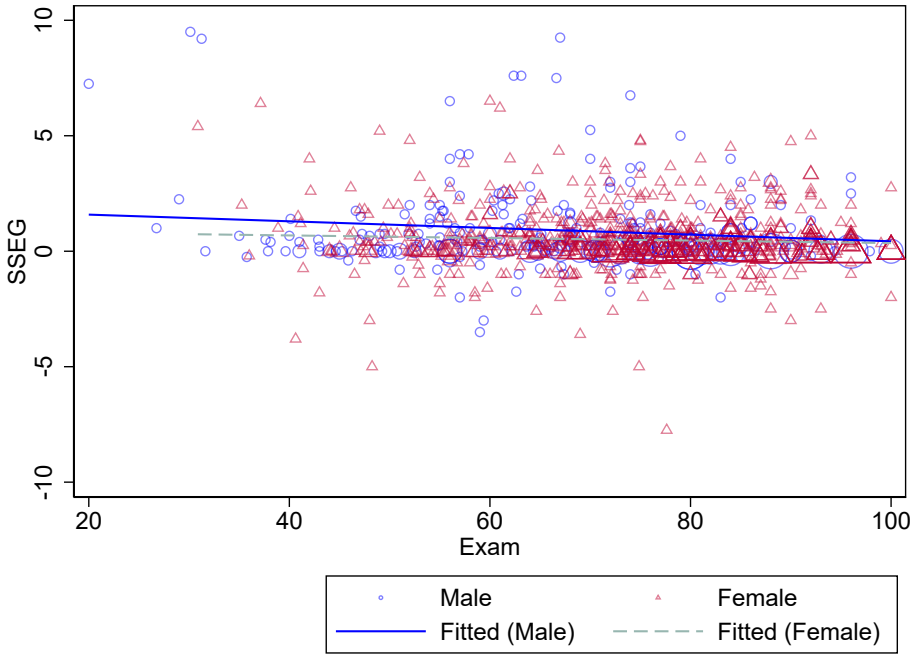
Table A4 replicates the regression on the determinants of SSEG in table 5.

TABLE A3—RATINGS RECEIVED BY SELF AND OTHERS EXCLUDING ZERO-VARIANCE GROUPS

	D16-17		D17-18		D18-19	
	M	F	M	F	M	F
Self	9.77	9.59**	9.73	9.70	NA	NA
Social	8.14	8.86***	8.67	9.11**	8.89	9.19***

Note: Abstention and groups with zero variance have been excluded from the analysis. *Self* denotes the rating an individual assigned to herself, while *Social* represents the average ratings received from peers, excluding the self-assigned rating. Mann-Whitney two-sample statistic significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

FIGURE A8. SSEG, ACADEMIC PERFORMANCE AND GENDER EXCLUDING ZERO-VARIANCE GROUPS



Note: The size of the circles and triangles is proportional to the observation frequency.

TABLE A4—DETERMINANTS OF SSEG, EXCLUDING ZERO-VARIANCE GROUPS

	(1) D16-17	(2) D16-17	(3) D17-18	(4) D17-18	(5) Combined	(6) Combined
Exam	-0.00726	-0.00809	-0.0200***	-0.0281*	-0.0102**	-0.0151
	(0.00485)	(0.0117)	(0.00694)	(0.0149)	(0.00410)	(0.00935)
Female	-0.525***	-0.620	-0.292	-1.064	-0.418***	-0.916
	(0.167)	(1.053)	(0.191)	(1.203)	(0.129)	(0.793)
Female_Majority	0.311*	0.311*	-0.0304	-0.0362	0.141	0.142
	(0.174)	(0.176)	(0.186)	(0.186)	(0.125)	(0.125)
Female × Exam		0.00128		0.0121		0.00723
		(0.0136)		(0.0173)		(0.0107)
Constant	1.429***	1.488*	2.085***	2.594**	1.551***	1.877***
	(0.410)	(0.869)	(0.505)	(1.007)	(0.321)	(0.667)
Observations	498	498	467	467	965	965

Note: The standard errors are presented in parentheses, and the significance levels are * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Multilevel mixed-effects models with robust standard errors, incorporating random intercepts for individual subjects. Female_Majority is a dummy variable which equals one if the group contains strictly more female students.

A7. Assignment instructions

2016/17 Instructions

This assignment consists of a group project (5 to 6 students) and a presentation, where the objective is for students to learn

- how to collect, process, and analyze data effectively using the descriptive statistics functions and plotting tools;
- how to present and communicate the results they find using clear and interpretable graphs and tables;
- how to collaborate with others.

Marking scheme:
A student's mark for this assignment will be calculated using the following formula:
(mark for the group project×0.7+mark for presentation×0.3)×contribution parameter

1 **The group project**
The total mark for the group project is 100, consisting of data collection (40 marks) and a report that analyzes the data (60 marks).

1) Each group will be randomly assigned a location (e.g., a state in US, a country in Europe) and a time period (e.g., 1997-2017). A group must collect data on at least for different economic variables (e.g., GDP, saving rates, household income, residences' education levels, etc.) for the assigned location during the assigned period.
2) Once a group has finished collecting the required data, the group needs to analyze the collected data (e.g., estimating the correlation between different variables, discussing the trends and changes of variables in the assigned period, etc.) using both statistical tools (e.g., regressions) and plotting tools (e.g., graphical techniques).
3) Each group must write a report on their analysis, including both the process and the results of the data analysis. The report must present the usage of multiple statistical tools and multiple graphical techniques.
   - Each group must submit two files, a data sheet in Excel format and a report in pdf format. The two files should be combined into one compressed zip file and submitted through ICE before the due date.
   - The data sheet and the report must be clear and easy to read.
   - The report should provide an economic explanation for the analytical results. For example, using the data collected, a group estimates a regression line between GDP and government expenditure which has

a positive slope. Then the group should discuss the possible reasons for why this slope is positive and the economic implication for this estimated linear relationship. The discussions must be logical.

- The report should also include detailed information on the data resource, such as the name of the websites where the data were found and the links to the websites. If data is collected from printed resources like yearbooks, the name of these resources and detailed information (e.g., publishers, ISBN, etc.) should be provided.

2 **The presentation**

Each group should select one member as the representative to present the group work. The presentation should include:

- what data the group collected and where the group found the data (10 marks);
- what analysis the group did on the data (presentations of used statistical and graphic tools must be presented) (40 marks);
- what the economic explanations and implications of the data analysis are (30 marks).

Students will need to answer the teacher's questions during their presentations. Up to 20 marks will be awarded to a group if their presenter answers the teacher's questions correctly.

3 **The contribution parameter**

- This contribution parameter is designed to punish free riding and to enhance team collaboration.

At the end of the last teaching week,each student will have a chance to do a peer review to evaluate his/her teammates' contribution to the group work. During this peer review, a student will give grades to all the team members in his/her group (including him/herself), which reflects his/her evaluation of his/her teammates' contribution to the group work. Each grade is between 0 and 100. A student who thinks that his/her teammate A contributes more to the group work than another teammate B does should give a higher grade to A and a lower grade to B. (Each student's evaluations on his/her teammates and him/herself will be private, that is, the grades he/she gives and receives can only be seen by him/herself and the module leader.)
A student's grade will be the arithmetic mean of the sum of the grades he/she receives from all the members (including him/herself) in his/her group. A student's contribution parameter is the ratio between his/her grade and the highest grade in his/her group.

Example (I will use a group with three members as an example. However, each actual group for the assignment will consist of 5 to 6 students.)

A group has three students, Aaron, Betty, and Charlie. This group received 89 for their group project and 93 for the presentation.

Both Aaron and Betty worked hard on the group project, spending a lot of time on collecting data and analyzing the data. However, Charlie did very little for the group work.

Aaron knows Betty has contributed a lot to the group project, so he gives 95 to Betty in the peer review. He thinks himself contributes a little more to the group project than Betty does, so he gives 99 to himself. He gives 10 to Charlie as he knows that Charlie did little for the group project.

Betty gives 96 to both Aaron and herself. She gives 20 to Charlie. Charlie gives 60 to Aaron, 70 to Betty, and 90 to himself.

Aaron's grade: (99+96+60)/3=85

Betty's grade: (95+96+70)/3=87

Charlie's grade: (10+20+90)/3=40

Since Betty receives the highest grade in peer review, her contribution parameter is 1.

Aaron's contribution parameter is 85/87=0.977.

Charlie's contribution parameter is 40/87=0.460.

The, Betty's mark for the assignment is (89×0.7+93×0.3)×1=90.2;

Aaron's mark for the assignment is (89×0.7+93×0.3)×0.977=88.1;

Charlie's mark for the assignment is (89×0.7+93×0.3)×0.=41.5.

2017/18 Instructions

This assignment consists of a group project (60%) and a presentation for the group project (40%). This assignment aims to help students learn

- how to collect, process, and analyze data effectively using the descriptive statistics functions and plotting tools;
- how to present and communicate the results they find using clear and interpretable graphs and tables;
- how to collaborate with others.

1 **Formation of Groups**

Each student must join and can only join, one group. Each group will consist of five to six students. Any student who fails to join any group by {a date specified by the module convener}[14] will be randomly assigned to a group by the module leader.

2 **The assignment**

This assignment consists of two parts, a group project (60

[14]Should I include the explicit date here?

3 The group project

The total mark for the group project is 100, consisting of data collection (40 marks) and a report that analyzes the data (60 marks).

1) Each group will be randomly assigned a location (e.g., a state in the US, a country in Europe) and some time period(s) (e.g., 1997-2017, or 1975-1985 and 2005-2015). A group must collect data on at least four different economic variables (e.g., GDP, saving rates, household income, residents' education levels, etc.) for the assigned location during the assigned period(s).
2) Once a group have finished collecting the required data, the group needs to analyze the collected data (e.g., estimating the correlation between different variables, discussing the trends and changes of variables in the assigned period, etc.) using at least (but not limited to) all the statistical techniques listed below and at least two different graphical techniques.
   - Mean, median, variance, standard deviation, and growth rates
   - Correlation and regressions, the measure of the goodness of fit, prediction
   - Techniques of finding the underlying trend in time series data
3) Each group must write a report on their data analysis, including both the process and the results of the data analysis. The report must present the usage of multiple statistical tools and multiple graphical techniques.
   - Each group must submit two files, a data sheet in Excel format and a report in pdf format. The two files should be combined into one compressed zip file and submitted through ICE before the due date. The data sheet and the report MUST be named using the group name as "group name + content." For example, for the group 01 Group D1/01 A,
     - their data sheet MUST be named as 01-Group-D101-A-data;
     - their report MUST be named as 01-Group-D101-A-report.
   - The data sheet and the report must be clear and easy to read.
   - The report should provide economic explanations for the analytical results, and their analysis should provide economic insights. For example, using the data collected, a group estimates a regression line between GDP and government expenditure which has a positive slope. Then the group should discuss the possible reasons for why this slope is positive and the economic implication for this estimated linear relationship. The discussions must be logical.
   - The report should also include detailed information on the data resource, such as the name of the websites where the data were found and the links to the websites. If data is collected from printed resources like yearbooks, the name of these resources and detailed information (e.g., publishers, ISBN, etc.) should be provided.

4 **The presentation**

Each group should select one (and only one) member as the representative to present the group work. The presentation should include:

- what data the group collected and where the group found the data (10 marks);
- what analysis the group did on the data. The presentations must show what graphic and statistical techniques are used in analyzing the data; (40 marks);
- what the economic explanations and implications of the data analysis are (30 marks).

Students will need to answer teacher questions during their presentations. Up to 20 points will be awarded to a group if their presenter correctly answers the teacher's questions.

5 **Submission of the assignment**

Each group only needs to submit one copy of the zip file of their project. This zip file MUST be named as "group name + project." For example, group 01 Group D1/01 A MUST name the zip file as 01-Group-D101-A-project.
Each group must submit their report together with a Coursework Submission Cover Sheet that has the signatures of all group members. On this cover sheet, there is a form called Summary of Contributions. Each student needs to state what he/she contributed to the group project.

TABLE A5—**Example: Summary of Contributions**

Student name	Student ID	Contributions to the group project (Maximum words: 30)
Amit	********	I collected the data for our group project.
Bella	********	I did statistical analysis on the data.

6 **Marking scheme**

Each group will receive a mark on their report and a mark on their presentation. The mark for a group project is calculated as following:

Mark for the project=mark for the report$\times$0.6+mark for the presentation$\times$0.4

Based on what a student did for the group project, which is stated in the Summary of contributions form on the cover sheet, each student will receive an individual mark. A student's mark for this assignment will be calculated using the following formula:

Mark for the project$\times$contribution parameter$\times$0.4 + individual mark$\times$0.6

Example:
The group project of Amit's group receives 70 for the report and 60 for the presentation. Then the mark for the group project of Amit's group is 70×0.6+60×0.4=66. Amit receives 65 as his individual mark and his contribution parameter is 0.75. Then Amit's mark for the assignment is 66×0.75×0.4+65×0.6=62.76.

7 The contribution parameter

- This contribution parameter is designed to punish free riding and to enhance team collaboration.

When submitting their assignments, each student will have the opportunity to do a peer review to evaluate his/her teammates' contribution to the group work. During this peer review, a student will give grades to all the team members in his/her group (including him/herself), which reflects his/her evaluation of his/her teammates' contribution to the group work. Each grade is between 0 and 10. A student who thinks that his/her teammate A contributes more to the group work than another teammate B does should give a higher grade to A and a lower grade to B. (Each student's evaluations on his/her teammates and him/herself will be private, that is, the grades he/she gives to his/her group members can only be seen by him/herself and the module leader.)
A student's grade will be the arithmetic mean of the sum of the grades he/she receives from all the members (including him/herself) in his/her group. A student's contribution parameter is the ratio between his/her grade and the highest grade in his/her group.

Example: (I will use a group with three members as an example. However, each actual group for the assignment will consist of 5 to 6 students.)

A group has three students, Aaron, Betty, and Charlie. The mark for their group project is 90.

Both Aaron and Betty worked hard on the group project, spending a lot of time collecting data and analyzing the data. However, Charlie did not do much for the group work.

Aaron's individual mark is 92, Betty's individual mark is 95, and Charlie's individual mark is 60. Aaron knows Betty has contributed a lot to the group project, so he gives 9 to Betty in the peer review. He thinks himself contributes a little more to the group project than Betty does, so he gives 10 to himself. He gives a 0 to Charlie as he thinks that Charlie did little for the group project.

Betty gives 10 to both Aaron and herself. She gives 1 to Charlie. Charlie gives 4 to Aaron, 8 to Betty and 9 to himself.

Aaron's grade: (10+10+4)/3=8 Betty's grade: (9+10+8)/3=9 Charlie's grade: (0+1+9)/3=3.33

Since Betty receives the highest grade in peer review, her contribution parameter is 1.

Aaron's contribution parameter is 8/9=0.89.

Charlie's contribution parameter is 3.33/9=0.37.

Then Betty's mark for the assignment is 90*1*0.4+95*0.6=93.6; Aaron's mark for the assignment is 90*0.89*0.4+92*0.6=87.24; Charlie's mark for the assignment is 90*0.37*0.4+60*0.6=49.32.

Note that the contribution parameters will be calculated by the evaluations given by students in the peer review, that is, if a group has six members, but only four of these six members do the peer review, then the contribution parameters for all six members will be calculated by the evaluations made by the four students who do the peer review. For a group such that no member of this group does the peer review, all members of this group will receive the same contribution parameter, which is 1.

2018/19 Instructions

This assignment consists of a group project (60%) and a presentation for the group project (40%). This assignment aims to help students learn

- how to collect, process, and analyse data effectively using the descriptive statistics functions and plotting tools;
- how to present and communicate the results they find using clear and interpretable graphs and tables;
- how to collaborate with others.

1 **Formation of Groups**
Each student must join and can only join one group. Each group will consist of five to six students. Any student who fails to join any group by {a date specified by the module convener}[15] will be randomly assigned to a group by the module leader.

2 **The assignment**
This assignment consists of two parts, a group project (60%) and a presentation for this group project (40%).

3 **The group project**
The total mark for the group project is 100, consisting of data collection (40 points) and a report that analyzes the data (60 points).

1) Each group will be randomly assigned a location (e.g., a state in the US, a country in Europe) and some time period (e.g., 1997-2017 or 1975-1985 and 2005-2015). A group must collect data on at least four economic variables (e.g., GDP, savings rates, household income, resident education levels, etc.) for the assigned location during the assigned period(s).

[15]Should I include the explicit date here?

2) Once a group has finished collecting the required data, the group needs to analyze the collected data (e.g., estimating the correlation between different variables, discussing the trends and changes of variables in the assigned period, etc.) using at least (but not limited to) all the statistical techniques listed below and at least two different graphical techniques.
   - Mean, median, variance, standard deviation, and growth rates
   - Correlation and regressions, measures of the goodness of fit, prediction
   - Techniques of finding the underlying trend in time series data
3) Each group must write a report on their data analysis, including both the process and results of data analysis. The report must present the usage of multiple statistical tools and multiple graphical techniques.
   - Each group must submit a SINGLE .pdf file consisting of three parts, the cover sheet, the report, and the data set. The cover sheet must be signed by all members of the group by hand and then be scanned and put before the report. The data set must be turned into a .pdf file and be attached to the end of the report.
   - The data sheet and the report must be clear and easy to read.
   - The report should provide economics explanations for the analytical results, and their analysis should provide economics insights. For example, using the data collected, a group estimates a regression line between GDP and government expenditure which has a positive slope. Then the group should discuss the possible reasons for why this slope is positive and the economic implication for this estimated linear relationship. The discussions must be logical.
   - The report should also include detailed information about the data resource, such as the name of the websites where the data were found and the links to the websites. If data are collected from printed resources like yearbooks, the name of these resources and detailed information (e.g., publishers, ISBN, etc.) should be provided.

4 **The presentation** Each group should select one (and only one) member as the representative to present the group work. The presentation should include:

- what data the group collected and where the group found the data (10 marks);
- what analysis the group did on the data. The presentations must show what graphic and statistical techniques are used in analyzing the data; (40 marks);
- what the economic explanations and implications of the data analysis are (30 marks). Students will need to answer the teacher's questions during their presentations. Up to 20 marks will be awarded to a group if their presenter answers the teacher's questions correctly.

5 **Submission of the assignment**

Each group only needs to submit one copy of the pdf file of their project. This pdf file MUST be named as "group name + project." For example, the group 01 Group D1/01 A MUST name the pdf file as 01-Group-D101-A-project. Each group must submit their report together with a Coursework Submission Cover Sheet that has the signatures of all the group members. On this cover sheet, there is a form called Summary of Contributions. Each student needs to state what he/she contributed to the group project.

TABLE A6—**Example: Summary of Contributions**

Student name	Student ID	Contributions to the group project (Maximum words: 30)
Amit	********	I collected the data for our group project.
Bella	********	I did statistical analysis on the data.

6 **Marking scheme**

Each group will receive a mark on their report and a mark on their presentation. The mark for a group project is calculated as follows: Mark for the project = mark for the report × 0.6 + mark for the presentation × 0.4 Based on what a student did for the group project, which is stated in the Summary of contributions form on the cover sheet, each student will receive an individual mark. A student's mark for this assignment will be calculated using the following formula: Mark for the project × contribution parameter × 0.4 + individual mark × 0.6 Example: The Amit group project received 70 for the report and 60 for the presentation. Then the mark for the group project of Amit's group is 70 ×0.6+60 ×0.4=66. Amit receives 65 as his individual mark and his contribution parameter is 0.75. Then Amit's mark for the assignment is 66 ×0.75 ×0.4+65 ×0.6=62.76.

7 **The contribution parameter**

- This contribution parameter is designed to punish free riding and to improve team collaboration.

When submitting their assignments, each student will have the opportunity to do a peer review to evaluate the contribution of her teammates to the group work. During this peer review, a student will give grades to all team members in her group EXCEPT herself, which reflects her evaluation of the contribution of her teammates to the group work. Each grade is between 0 and 10. A student who thinks that his/her teammate A contributes more to the group work than another teammate B does should give a higher grade to A and a lower grade to

B. (Each student's evaluations on his/her teammates and him/herself will be private, that is, the grades he/she gives to his/her group members can only be seen by him/herself and the module leader.) A student's grade is the arithmetic mean of the sum of the grades he/she receives from all his/her teammates in his/her group. A student's contribution parameter is the ratio between his/her grade and the highest grade in his/her group.

Example: (I will use a group with three members as an example. However, each actual group for the assignment will consist of 5 to 6 students.)

A group has three students, Aaron, Betty, and Charlie. The mark for their group project is 90.

Both Aaron and Betty worked hard on the group project, spending a lot of time on collecting data and analyzing the data. However, Charlie did not do much for the group work.

Aaron's individual mark is 92, Betty's individual mark is 95, and Charlie's individual mark is 60.

Aaron knows Betty has contributed a lot to the group project, so he gives 9 to Betty in the peer review. He thinks himself contributes a little more to the group project than Betty does, so he gives 10 to himself. He gives 0 to Charlie as he thinks that Charlie did little for the group project.

Betty gives 10 to both Aaron and herself. She gives 1 to Charlie.

Charlie gives 4 to Aaron, 8 to Betty, and 9 to himself.

Aaron's grade: (10+4)/(3-1)=7 Betty's grade: (9+8)/(3-1)=8.5 Charlie's grade: (0+1)/(3-1)=0.5

Since Betty receives the highest grade in peer review, her contribution parameter is 1. Aaron's contribution parameter is 7/8.5=0.82.

Charlie's contribution parameter is 0.5/8.5=0.06.

Then Betty's mark for the assignment is 90×1*0.4+95×0.6=93;

Aaron's mark for the assignment is 90×0.82×0.4+92×0.6=84.72;

Charlie's mark for the assignment is 90×0.06×0.4+60×0.6=38.16.

Note that the contribution parameters will be calculated by the evaluations given by the students in the peer review, that is, if a group has six members, but only four of these six members do the peer review, then the contribution parameters for all six members will be calculated based on the evaluations made by the four students who do the peer review. For a group such that no member of this group does the peer review, all members of this group will receive the same contribution parameter, which is 1. If only one member in a group does the peer review, though this student cannot give him/herself a mark in this peer review, this student will receive the contribution parameter 1.