University of Cyprus
Department of Economics

Working Paper 06-2024

Reciprocity in Peer Assessments

Lunzheng Li, Philippos Louis, Zacharias Maniadis and Dimitrios Xefteris

Department of Economics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus
www.ucy.ac.cy/econ/?lang=en

Reciprocity in Peer Assessments*

Lunzheng Li† Philippos Louis‡ Zacharias Maniadis§ Dimitrios Xefteris¶

December 4, 2024

Abstract

Peer assessment's reliability can be undermined when participants behave strategically. Using a formal model we show how reciprocity can lead to reviewers inflating their rating of each others' work, which is exacerbated when review takes place sequentially. We conduct a pre-registered online experiment and we find that reviewers engaged in mutual-review relationships inflate their reports more, compared to when reviews are one-sided. For sequential reviews, a non-trivial fraction of first movers maximally over-report. In accordance to our theoretical model, we also find that second movers are responsive to the review they received, but only when reviews are mutual. This reveals the potential for a quid-pro-quo element in mutual reviews. Our results highlight the importance of appropriately structuring peer assessment to take strategic reciprocity motives into account and ensure the system's reliability.

Keywords: Reciprocity, Lying, Peer Assessment, Experiment.
JEL Codes: D9, L2, M5

*This project has received funding from the European Union's Horizon 2020 Research and innovation Programme under Grant Agreement number: 857636 — SInnoPSis — H2020-WIDESPREAD-2018-2020/H2020-WIDESPREAD-2018-04. This project has also received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No [101079196] (Twinning for Excellence in Management and Economics of Research and Innovation [TWIN4MERIT]). This study has been approved by the Cyprus National Bioethics Committee (CNBC DECISION#2023.01.104), and registered with the AEA RCT registry (AEARCTR-0011281). We are grateful to Carlos Alós-Ferrer, Ido Erev, Shuguang Jiang, Lorenzo Zirulia, and seminar participants at Heriot-Watt University, as well as attendees of the 1st SInnoPSis Conference in Nicosia, Cyprus, the 2024 Asia Pacific ESA Meeting in Singapore, the 7th World Congress of the Game Theory Society in Beijing, China, the BREW-ESA 2024 Meeting in Delhi, India, and the Lancaster 2024 Psychoeconomics Workshop for their insightful comments.

†University of Cyprus, SInnoPSis. lunzheng.li@ucy.ac.cy.

‡University of Cyprus, Department of Economics. louis.philippos@ucy.ac.cy

§University of Cyprus, SInnoPSis and University of Southampton. Department of Economics, maniadis.zacharias@ucy.ac.cy.

¶University of Cyprus, Department of Economics. xefteris.dimitrios@ucy.ac.cy

1 Introduction

An accurate evaluation of an individual's or team's performance in their work is not only a prerequisite for any meritocratic system, but also a necessary condition for establishing proper incentive schemes in organizations. Performance evaluation is a straightforward process whenever objective and clearly quantifiable metrics exist. However, in many instances the output to be evaluated might elude objective measurements. Examples include a manager's performance as a leader, a teacher's ability to inspire students, and the contribution of a scientific paper. While most would agree that these dimensions should matter in determining decisions like promoting employees, awarding teaching prizes, or publishing articles, a similar agreement cannot be found when looking for ways to measure these.

One approach to overcome this issue is to identify domain-specific metrics that correlate with the object of interest. For example, employee turnover rates for managers, students' performance in standardized tests for teachers, or the number of citations for scientific papers.[1] Of course, even when such metrics exist, they might not be available when a related decision must be taken. A different approach embraces subjectivity, seeking feedback from peers in a structured manner, in what is commonly known as peer assessment, peer evaluation or peer review.

The reader is likely to be familiar with such processes in the academic sphere where they are applied to assist with decisions in the publication process, the awarding of research grants and elsewhere. The peer review system is trusted by researchers to enhance the quality and integrity of scholarly work, encourage accountability and transparency and ensure that standards of excellence are upheld across various fields and disciplines (Bornmann, 2011; Nicholas et al., 2015). Peer assessment is also found in organizations, where employees are periodically evaluated by coworkers, subordinates and managers, a process often referred to as "360-degree",

[1] See Nishii and Mayer (2009); Gilbert (2018); Zhu et al. (2015) and references therein, for discussions on the viability of the metrics in the examples.

multi-rater or multi-source feedback. This method of evaluation grew in popularity during the '80s and '90s and remains widely used today (Kane and Lawler, 1978; Tornow, 1993; Di Fiore and Souza, 2021). By providing individuals with diverse perspectives on their work, it offers constructive criticism and valuable insights that can lead to significant personal and professional development. More generally, it can foster a culture of continuous improvement and collaboration, enhancing workers' motivation and overall performance (London and Smither, 1995; Smither et al., 2005; Cappelli and Conyon, 2018; Wiles, 2018; Fleenor et al., 2020; Morgan et al., 2020).

Despite its success across domains, peer assessment is not immune to criticism. Since it relies on evaluators' subjective judgment, any biases affecting this judgment can influence the outcome of the process, e.g., favoritism due to personal relations, or prejudice towards specific individuals or groups (Love, 1981; Tsui and Barry, 1986; Prendergast and Topel, 1993; Lee et al., 2013; Sol, 2016; Tomkins et al., 2017; Frederiksen et al., 2020; Stelmakh, 2022). Heterogeneity in raters' skills, cognitive abilities and goals can add further noise or bias to peer assessment (DeNisi, 2003; Murphy et al., 2004; Moers, 2005; Wong and Kwong, 2007; Wang et al., 2010). These issues can be further exacerbated by poor choices in the design of such processes, ranging from the way ratings are operationalized (Saal et al., 1980) to broader issues of anonymity and accountability (London et al., 1997; Bamberger et al., 2005; Harari and Rudolph, 2017).

The reliability of peer assessment can also be put at risk by more deliberate *strategic behavior* on the part of evaluators. For instance, in environments where raters are also in competition with ratees, e.g., for resources or a possible promotion, there is evidence that some individuals might provide less positive feedback about their peers to make themselves or their close friends appear better (Carpenter et al., 2010; Balietti et al., 2016; Huang et al., 2019; Stelmakh et al., 2021; Hussam et al., 2022; Olckers and Walsh, 2022; Riedl et al., 2024). A different type of strategic behavior, which is the focus of this paper, is that of individuals strategically providing overly positive reviews due to anticipation of *reciprocal behavior* by their peers. In fact, rating inflation has been identified as one of the main contextual barriers to the success of systems based on peer assessment. It is therefore surprising that the role of reciprocity in peer assessment has received very little attention in the literature up to now (see Artz et al., 2023; Franke and Papadopoulos, 2024, for very recent exceptions). Our paper aims at better understanding when and how such strategic reciprocity can undermine a peer assessment exercise.

Ample evidence from the lab and the field shows that in both social and economic interactions people often treat kindly the ones that were kind to them, and expect others to do the same (see Malmendier et al., 2014, for a survey). In the context of peer evaluation, reciprocity of this kind would lead strategically-minded evaluators to be more positive towards their peers if they expect their roles to be reversed in the future. Such an expectation is very reasonable in organizations, where the 360-feedback process is repeated regularly and evaluations are often mutual (Artz et al., 2023). Similarly, in relatively narrow academic fields, researchers reviewing others' work for publication or the allocation of funding can be almost certain that their own future work will be evaluated by the other part at some point.[2]

While theoretically possible, observing such strategic rating behavior in the field is challenging. One important hurdle is that one lacks an objective measure to compare subjective ratings.[3] A second obstacle is the difficulty of disentangling pure other-regarding motives, such as treating one's close colleagues well, from self-regarding strategic behavior.[4] Finally, repeated interaction in the field can lead to

[2] In fact, one argument in favor of reviewers' anonymity and the double-blind system is that it can prevent such strategic behavior (Lee et al., 2013; Tomkins et al., 2017). Still, even with anonymity in place, it is often the case that scholars have strong indications about some of their reviewers' identities.

[3] One can argue that in situations where such measures exist, peer evaluation systems are redundant.

[4] In a recent (yet unpublished) paper, Artz et al. (2023) use proprietary data from an online retailer and report evidence of rating inflation among employees that chose to rate each other. Their results are consistent with the notion that strategic motives contribute to higher ratings, along with a selection effect. In another unpublished paper, Franke and Papadopoulos (2024) provide

the development of social norms that guide behavior in a way that does not allow a researcher to distinguish between norm compliance and strategic reciprocity. We overcome these difficulties through a controlled preregistered online experiment. The design and our hypotheses are formulated following the analysis of a formal model of reciprocity in the assessment of others.

The simple environment under consideration involves two individuals, each tasked with assessing someone else.[5] By varying the sequence between assessments (simultaneous vs. sequential) and the nature of the relationship between players (belonging to a group or not, and assessing the assessor or not), we can examine the performance of the model and explore the role of reciprocity in our setting. A key prediction of our theory is that, in an environment with sequential and mutual assessments among the two agents, the agent moving first will be inclined to provide a higher assessment of the second player, anticipating a higher assessment in return. We test four theoretical predictions and two preregistered hypotheses, the latter registered with the American Economic Association RCT registry, through online experiments conducted with participants representative of the UK and US populations. Importantly, the direction of the results is in line with all the predictions of our model, and statistical tests are supportive in about 50% of the cases. It is also noteworthy that the least inflated assessments occur in an experimental condition called NM (Non-mutual), where all possible factors conducive to reciprocity (namely, mutual assessments, belonging to a group, and sequential moves) are ruled out.

evidence from a German TV show and evince that, even in the absence of direct observability of past individual ratings, there is evidence for positive strategic reciprocity. It should be noted that learning and information acquisition is a potential driver of their results, a channel which we rule out by design.

[5] As will become evident in the presentation of our model and experiment, we are assuming that players are symmetric. This is indeed an important assumption, but our paradigm can be extended to examine settings with asymmetric players, e.g., with different levels in the hierarchy. We also assume for simplicity that the two players share similar information ex ante, which can be criticised as unrealistic, since individuals usually have a good ex ante idea of their performance. However, we argue that even in our simple environment, much can be learned about behavior in peer assessment tasks. For instance, workers who expect to be highly assessed are still likely to have a keen interest in the other person's assessment, as long as the element of subjectivity is significant. In general, the lessons learned from our setting are likely to carry over to more complex environments.

In particular, we inform our experimental design by analyzing first a formal model of reciprocity in assessments, within a framework of interacting altruism levels (Levine, 1998). In such an environment, lying-averse assessors, characterized by a privately observed altruism parameter, will judge the quality/performance of another agent differently depending: 1) on their own altruism, and 2) on the perceived altruism of the agent whose performance they assess. That is, agents want to be more generous to more altruistic individuals. Moreover, if an assessor expects to be assessed in the future by the agent she assesses now, she has incentives to overreport and thus exaggerate her altruism level. When reciprocity in assessment is not present, then this channel is closed and the report only depends on the performance and the level of altruism of the assessor.

In the experiments, the actual "quality" of individual performance is determined using a roll of a six-sided dice observed by the assessor, but individuals are free to report any value from zero to five. Disentangling the performance under assessment from decisions that may be affected by the applied peer-review procedure insulates our analysis from relevant endogeneity concerns. Indeed, if subjects were asked to perform a task that would be subsequently assessed, then the details of the assessment process could influence their performance, confounding the performance indicator in multiple non-trivial ways. Our design controls for the quality/performance under assessment, while also ensuring its non-verifiability to everyone except the assessor, which is key for our purposes.

The results provide support for our theoretical model across different dimensions. First of all, the existence of other-regarding preferences in the model implies that people should inflate their report across the board. This is exactly what we observe: for all conditions, reports are higher than 2.5, and this is significant for all conditions except NM. This consistent and strong evidence supports our first prediction. Secondly, the model's key result is that first movers in the sequential and mutual condition (SeM) report higher than all players in the simultaneous and mutual condition (SiM). In fact, the first movers' average report in SeM (3.05) is higher than average report for SiM (2.88), which is consistent with our model, although the difference is not significant at conventional levels (p=0.3342). Thirdly, the model predicts that (in expectation) reports of second movers in condition SeM should not differ from reports in condition SiM. In the experiment, average report of second movers in condition SeM (2.79) is relatively similar to average report in SiM (2.88), although equivalence cannot be supported statistically. Finally, the model predicts that the reports of second movers in SeM depend on the respective reports of first movers. The data indicate a statistically significant ($p < 0.1$) – albeit modest – positive correlation between the two types of reports. Moreover, regression analysis yields a significant and positive coefficient (approximately 0.16 with $p < 0.01$) on the first movers' reports in explaining the second movers' reports in SeM. In summary, the data point consistently in the direction predicted by our model, and in many cases the effects are significant.

The data also reveal interesting additional features. First movers in our two sequential conditions (SeM and SeNM) disproportionately report the maximum allowable value for others. This can be understood as an attempt to signal that they are willing to lie to benefit the other person. On the other hand, second movers seem to reciprocate this only in the mutual condition SeM. In addition, we do find meaningful and statistically significant difference in average reports between the polar opposite cases of NM and first movers in SeM. Recall that NM simulates interaction in a large anonymous field, while SeM approximates interaction in small domains, where the probability of repeated interaction is one. This provides some (exploratory) support for the idea that the shadow of repetition matters for peer assessment.

Distorting one's assessment of others' performance is a form of deception that involves some degree of lying. It is hence natural for our experimental design to be based on the paradigm introduced by Fischbacher and Föllmi-Heusi (2013), which forms the backbone of a significant part of the experimental literature on this topic. Most other experiments in the literature (Gneezy, 2005; Erat and Gneezy, 2012; Wiltermuth, 2011) do not consider reciprocal opportunities to lie for the other person. Instead, they examine various forms of splitting the surplus between an agent that can potentially lie, and others. Colzani et al. (2023) also employ the same dice paradigm and have participants matched in a simultaneous vs. a sequential treatment. However, participants' reports affect their own payoffs, rather than the payoffs of the other person. In Buckle et al. (2021) participants in one treatment can lie to benefit a passive participant. They find that individuals will still lie to benefit others without any personal gain. Nevertheless, they are significantly less willing to lie to benefit others compared to when they can benefit themselves. There is no possibility for reciprocal behavior in their setting. Ours is the first theoretically-driven empirical study on the potential of positive reciprocity in lying behavior. Alempaki et al. (2019) also consider the issue of whether lying can be used as a device for reciprocity, but the domain of reciprocity is different from ours: they have an initial task that involves splitting the pie in a dictator experiment, while we have a task that involves potential lying at both stages of the game. In addition, our predictions are supported by a formal model of reciprocity. Dato et al. (2019) also have a model of reciprocity that adapts Levine (1998) and they consider the effect of reciprocity on lying behavior. However, their setting is different. They consider interactions with negatively correlated payoffs, whereas our interest is in an environment where people could lie to benefit others, rather than themselves. Instead of a competitive setting with negative reciprocity, we consider an environment with potentially positive reciprocity. Importantly, our focus is on the optimal behavior of the first mover, and how it differs across environments where the reciprocal relationship can be present or absent (capturing different boundary conditions about the size of the underlying population, and therefore the probability of repeated play).

The structure of the paper is as follows. In Section 2 we present our theoretical model of reciprocity and its predictions. Section 3 presents our additional preregistered hypotheses, the different experimental conditions, and the details about the implementation of our experiments. Section 4 contains our empirical results, both in terms of descriptive and reduced-form statistical testing, as well as in terms of econometric analysis. Section 5 discusses the implications of our work and concludes.

2 The Model

Our model is based on the influential framework of Levine (1998), adapted to the environment of reciprocal assessments. There are two players P_1 and P_2, each observing a private draw from a known distribution. Unlike the standard die-rolling paradigm in which a player's payoff depends on their own report of the private draw, each player's payoffs depend on the other player's report in our model. As we shall see, for the players who move first, the prospect of inducing reciprocity by indicating that one is of an altruistic type will matter for the assessment of the other person.

Private Draws and Reports P_1 observes a private draw s_2 from distribution $\mathcal{N}(\mu_{s_2}, \sigma^2_{s_2})$ and she reports x (that is, they claim that $s_2 = x$). This report determines the payoff of P_2, which is equal to that report. Similarly, P_2 observes a private draw s_1 from $\mathcal{N}(\mu_{s_1}, \sigma^2_{s_1})$, and she reports y, which in turn becomes the payoff of P_1. Therefore, both players can potentially affect their opponents' payoffs by lying, i.e., by reporting an x (y) that is not equal to s_2 (s_1) but higher or lower.

Lying Costs The players are averse to lying. That is, whenever P_1 (P_2) observes s_2 (s_1) and reports x (y), she incurs a cost that is (quadratically) increasing in the distance between s_2 (s_1) and x (y).

Reciprocal Utility In addition to the monetary payoffs (x for P_2, and y for P_1) and those related to lying, players receive an additional component, namely "reciprocal utility". That is, they derive utility not only from their own payoff but also from the payoff of the other player. The key determinants of an agent's reciprocal utility are the two players' reciprocity coefficients. For player P_i, the reciprocity coefficient is $\alpha_i \sim \mathcal{N}(\mu_{\alpha_i}, \sigma^2_{\alpha_i})$. A positive value of this coefficient indicates altruism and a negative value indicates spite. The larger the absolute value of this parameter, the higher the level of altruism/spite.

Importantly, in the setting of Levine (1998) an agent's reciprocal utility is determined both by own reciprocity coefficient and by the (expected) reciprocity coefficient of the other player. That is, an agent derives more reciprocal utility when she is more altruistic herself, but also when she interacts with a more altruistic agent.[6]

Overall, the ex-post utilities of the two players are given by,

$$U_1(x, y) = -(s_2 - x)^2 + (\alpha_1 + \alpha_2)x + y,$$

$$U_2(x, y) = -(s_1 - y)^2 + (\alpha_1 + \alpha_2)y + x,$$

where the first part of the sum represents the lying costs, the second part represents reciprocal utility, and the final one is the utility from money.[7]

The key predictions of our model concern behavior in the "simultaneous moves" vs. "sequential moves" case.

Simultaneous Setting

In the case of P_1 and P_2 moving simultaneously, they do not observe their opponents' reports, and thus x and y are independent. P_1 solves the following maximization problem,

$$\max_x \quad -(s_2 - x)^2 + (\alpha_1 + E(\alpha_2))x + E(y)$$

where $-(s_2 - x)^2$ represents P_1's cost of lying and $(\alpha_1 + E(\alpha_2))x$ is the "reciprocal utility" that P_1 receives from the interaction. Intuitively, the higher the expected

[6] Levine (1998) considers binary types, i.e., altruists and non-altruists, while we allow for a richer type space that nests varying levels of altruism and spite.

[7] One could introduce different weights of own and others' altruism in the reciprocal utility part, without affecting the qualitative predictions of the model.

sum of reciprocity components, the happier P_1 becomes by choosing a high report x for the other person. Importantly, P_1 has no posterior information on the basis of which to estimate P_2's reciprocity, and thus he bases his decision on $E(\alpha_2) = \mu_{\alpha_2}$. Turning to P_2's maximization problem, we observe that it is similar:

$$\max_{y} \quad -(s_1 - y)^2 + (\alpha_2 + E(\alpha_1))y + E(x)$$

where $-(s_1 - y)^2$ is P_2's lying cost and $(\alpha_2 + E(\alpha_1))y$ is her "reciprocal utility". It is straightforward to calculate the first-order conditions. The equilibrium reports are as follows:

$$\begin{cases} x^* = \frac{1}{2}(\alpha_1 + \mu_{\alpha_2} + 2s_2) \\ y^* = \frac{1}{2}(\alpha_2 + \mu_{\alpha_1} + 2s_1) \end{cases} \tag{1}$$

These conditions indicate that the mere presence of reciprocal utility makes reports consistently greater that the draws (in expectation), if players are (on average) altruistic (i.e., if $\mu_{\alpha_1}, \mu_{\alpha_2} > 0$).[8]

Sequential Setting

Players move sequentially with P_1 being the first mover and P_2 being the second mover. In this case, x and y are dependent, as P_2 observes x before she reports for P_1, and P_1 knows about it. Therefore, P_1 solves the following maximization problem,

$$\max_{x} \quad -(s_2 - x)^2 + (\alpha_1 + E(\alpha_2))x + E(y|x)$$

where $-(s_2 - x)^2$ represents the cost of lying for P_1, $(\alpha_1 + E(\alpha_2))x$ is her reciprocal utility, and $E(y|x)$ is P_2's expected report y conditional on x, i.e., P_1's expected

[8] We believe it is reasonable to assume that players in our setting may act altruistically, a perspective supported by extensive literature providing behavioral and biological evidence for human altruistic behavior (see, for instance, Fehr and Fischbacher, 2003; Gintis et al., 2003; Fehr and Rockenbach, 2004). Furthermore, previous experiments have shown that agents engage in lying for the benefit of others when strategic incentives for lying are absent (e.g., Buckle et al., 2021), further supporting our premise.

monetary payoff conditional on P_2's report. That is, P_1 should now condition the expected report of her opponent on her own report. As for P_2, she observes x, and thus she knows her exact payoff, and her expectation of P_1's reciprocal coefficient is conditioned on P_1's report. P_2's maximization problem is as follows,

$$\max_{y} \quad -(s_1 - y)^2 + (E(\alpha_1|x) + \alpha_2)y + x.$$

We will prove the existence of a unique linear Perfect Bayesian Equilibrium of the game (henceforth, equilibrium), and we will fully characterize it. In such an equilibrium, P_1's expected monetary payoff is a linear function of x, given by $bx + z$.

Assume that an equilibrium exists. Then, P_1's maximization problem becomes,

$$\max_{x} \quad -(s_2 - x)^2 + (\alpha_1 + E(\alpha_2))x + bx + z$$

and its first-order condition yields,

$$\begin{aligned} & x = \frac{1}{2}(\alpha_1 + b + E(\alpha_2) + 2s_2) \\ \iff & 2x - b - \mu_{\alpha_1} - \mu_{\alpha_2} - 2\mu_{s_2} = \alpha_1 - \mu_{\alpha_1} + 2s_2 - 2\mu_{s_2} \end{aligned}$$

which can be used to rewrite P_2's expectation of P_1's reciprocity coefficient,

$$\begin{aligned} E(\alpha_1|x) &= E(\alpha_1 - \mu_{\alpha_1}|x) + \mu_{\alpha_1} \\ &= E(\alpha_1 - \mu_{\alpha_1}|2x - b - \mu_{\alpha_1} - \mu_{\alpha_2} - 2\mu_{s_2}) + \mu_{\alpha_1} \\ &= E(\alpha_1 - \mu_{\alpha_1}|\alpha_1 - \mu_{\alpha_1} + 2s_2 - 2\mu_{s_2}) + \mu_{\alpha_1} \end{aligned}$$

where $(\alpha_1 - \mu_{\alpha_1}) \sim \mathcal{N}(0,\, \sigma^2_{\alpha_1})$ and $(2s_2 - 2\mu_{s_2}) \sim \mathcal{N}(0,\, 4\sigma^2_{s_2})$.[9] Therefore, $E(\alpha_1 - \mu_{\alpha_1}|\alpha_1 - \mu_{\alpha_1} + 2s_2 - 2\mu_{s_2}) = \frac{\sigma^2_{\alpha_1}}{\sigma^2_{\alpha_1} + 4\sigma^2_{s_2}}(\alpha_1 - \mu_{\alpha_1} + 2s_2 - 2\mu_{s_2}) = \frac{\sigma^2_{\alpha_1}}{\sigma^2_{\alpha_1} + 4\sigma^2_{s_2}}(2x - b -$

[9] Notice that by writing $E(\cdot|x)$ we are slightly abusing notation, as $E(\cdot|x = c)$ is the proper way to express such conditional probabilities (i.e., by specifying that variable x takes the particular value c). With this in mind, it becomes clear why $E(\cdot|x) = E(\cdot|2x + \bar{c})$ (i.e., because $E(\cdot|x = c) = E(\cdot|2x + \bar{c} = 2c + \bar{c})$).

$\mu_{\alpha_1} - \mu_{\alpha_2} - 2\mu_{s_2})$,[10] and

$$E(\alpha_1|x) = \frac{\sigma^2_{\alpha_1}}{\sigma^2_{\alpha_1} + 4\sigma^2_{s_2}}(2x - b - \mu_{\alpha_1} - \mu_{\alpha_2} - 2\mu_{s_2}) + \mu_{\alpha_1}.$$

Combining this expression with the first-order condition of the maximization problem for P_2 allows us to calculate y in expectation,

$$E(y|x) = \frac{\sigma^2_{\alpha_1}}{\sigma^2_{\alpha_1} + 4\sigma^2_{s_2}}x + \frac{1}{2}(\mu_{\alpha_2} + \mu_{\alpha_1} + 2\mu_{s_1} - \frac{\sigma^2_{\alpha_1}}{\sigma^2_{\alpha_1} + 4\sigma^2_{s_2}}(b + \mu_{\alpha_1} + \mu_{\alpha_2} + 2\mu_{s_2}))$$

and thus

$$\begin{aligned} b &= \frac{\sigma^2_{\alpha_1}}{\sigma^2_{\alpha_1} + 4\sigma^2_{s_2}}, \\ z &= \frac{1}{2}(\mu_{\alpha_2} + \mu_{\alpha_1} + 2\mu_{s_1} - \frac{\sigma^2_{\alpha_1}}{\sigma^2_{\alpha_1} + 4\sigma^2_{s_2}}(b + \mu_{\alpha_1} + \mu_{\alpha_2} + 2\mu_{s_2})). \end{aligned}$$

Given b, z, and the expression $E(\alpha_1|x)$, the equilibrium reports can be computed as follows,

$$\begin{cases} x^* = \frac{1}{2}(\alpha_1 + b + \mu_{\alpha_2} + 2s_2) \\ y^* = bx^* + z + \frac{1}{2}(\alpha_2 - \mu_{\alpha_2}) + (s_1 - \mu_{s_1}) \end{cases} \tag{2}$$

where $b = \frac{\sigma^2_{\alpha_1}}{\sigma^2_{\alpha_1} + 4\sigma^2_{s_2}}$, and $z = \frac{1}{2}(\mu_{\alpha_2} + \mu_{\alpha_1} + 2\mu_{s_1} - \frac{\sigma^2_{\alpha_1}}{\sigma^2_{\alpha_1} + 4\sigma^2_{s_2}}(b + \mu_{\alpha_1} + \mu_{\alpha_2} + 2\mu_{s_2}))$.[11]

Indeed, one can easily substitute x^* with its equilibrium value in y^* and also

[10] This calculation is based on the following fact. Given independent $X \sim N(\mu_X, \sigma^2_X)$, $Y \sim N(\mu_Y, \sigma^2_Y)$, and $Z = X + Y$, we wish to compute $E[X|Z = z]$. In a bivariate normal distribution,

$$E[X|Z = z] = E[X] + \rho_{XZ} \times \frac{\sigma_X}{\sigma_Z} \times (z - E[Z])$$

where $\sigma_Z = \sqrt{\sigma^2_X + \sigma^2_Y}$ and $\rho_{XZ} = \frac{Cov(X,Z)}{\sigma_X \sigma_Z} = \frac{Cov(X,X+Y)}{\sigma_X \sigma_Z} = \frac{Var(X)}{\sigma_X \sigma_Z} = \frac{\sigma_X}{\sqrt{\sigma^2_X + \sigma^2_Y}}$. Substituting the terms, we derive

$$E[X|Z = z] = \mu_X + \frac{\sigma^2_X}{\sigma^2_X + \sigma^2_Y} \times (z - \mu_X - \mu_Y).$$

[11] By substituting these values of b and z in the players' maximization problems, one can straightforwardly validate that the described strategies form an equilibrium. Finally, by the fact that the above exercise identified unique admissible values for b and z, it follows that this equilibrium is unique.

write the latter only in terms of parameters. However, this way of writing y^* makes it clear that the report of P_2 correlates with the report of P_1 in a positive manner, which is a key observation underlying this theoretical analysis.

Theoretical Predictions

When both players have identical reciprocity coefficient distributions, s_i's drawn from the same distributions, and all variances equal to 1, i.e., $\mu_{\alpha_1} = \mu_{\alpha_2} = \mu_\alpha$, $\mu_{s_1} = \mu_{s_2} = \mu_s$ and $\sigma_{\alpha_1} = \sigma_{\alpha_2} = \sigma_{s_1} = \sigma_{s_2} = 1$, b equals $\frac{1}{5}$ and z becomes $\frac{4}{5}(\mu_\alpha + \mu_s) - \frac{1}{50}$. Therefore, the equilibrium reports in (1) and (2) reduce to,

$$\begin{cases} x^*_{si} = \frac{1}{2}(\alpha_1 + \mu_\alpha + 2s_2) \\ y^*_{si} = \frac{1}{2}(\alpha_2 + \mu_\alpha + 2s_1) \end{cases} \tag{3}$$

$$\begin{cases} x^*_{se} = \frac{1}{2}(\alpha_1 + \mu_\alpha + 2s_2 + \frac{1}{5}) \\ y^*_{se} = \frac{1}{10}(\alpha_1 + 5\alpha_2 + 4\mu_\alpha) + \frac{1}{10}(10s_1 + 2s_2 - 2\mu_s) \end{cases} \tag{4}$$

where (x^*_{si}, y^*_{si}) is the equilibrium report profile for the simultaneous setting and (x^*_{se}, y^*_{se}) is the analogous profile for the sequential setting. In expectation, we have $E(x^*_{si}) = E(y^*_{si}) = E(y^*_{se}) = \mu_\alpha + \mu_s$, and $E(x^*_{si}) < E(x^*_{se}) = \mu_\alpha + \mu_s + \frac{1}{10}$. That is, P_2 believes that P_1 inflated her report in the sequential setting by $\frac{1}{5}$ compared to the simultaneous one, and P_1 optimally inflates her report by the same amount, leading in both cases to an accurate assessment of the P_1's reciprocity coefficient, and subsequently to the same expectation of y^* conditional on s_1. Note that all equilibrium reports are, in expectation, higher than μ_s when $\mu_\alpha > 0$.

Finally, the analysis of our model allows us to formulate concrete theoretical predictions:

Prediction 1 *All movers in both settings over-report.*

Prediction 2 *The first mover in the sequential setting reports more in expectation compared to the players in the simultaneous setting.*

Prediction 3 *The second mover in the sequential setting reports the same in expectation as the players in the simultaneous setting.*

Prediction 4 *The report of the second mover in the sequential setting is increasing in the report of the first mover in the same setting.*

3 Experimental Design

The general setting is based on the dice-rolling paradigm of Fischbacher and Föllmi-Heusi (2013), which utilizes events with known distributions (such as dice-rolling), and compares the self-reported distributions with distributions under truth-telling to detect lies at the aggregate level. We instruct participants to roll a dice (physically or via an online method) and to report the outcome, which shall determine the payments of another participant. If they report numbers 1, 2, 3, 4 or 5, the other participant receives an equivalent payment in British pounds (e.g., 1 representing £1), but if they report a 6, the payment is £0.[12] Our four experimental conditions are as follows, and all of them are one-shot.

One-sided/Non-mutual (NM) In a large group, everyone observes a personal dice roll and reports a number that determines the payoff of someone else. Participants do not know who the "someone else" is. However, they do know that this "someone else" is not the person whose reports will count for themselves.

Simultaneous and Mutual (SiM) Subjects are randomly paired into groups of two. In a given group with subject A and subject B, player A observes a dice roll

[12] Despite allowing subjects from both the UK and the US, our experiment only used British pounds. This was because the Prolific platform, being UK-based, only processed payments in British pounds at the time of the experiment (September and October 2023). It wasn't until August 2024, as announced here: https://participant-help.prolific.com/en/article/9b3cab, that they began offering USD as a payment currency. As a result, users who registered on Prolific and participated in our experiment were aware that all transactions would be in pounds.

and reports a number that determines B's payoff, and in the meantime, player B observes another dice roll and reports a number that determines A's payoff. Note that both of them report for the other person before observing what the other person reported for them, i.e., players move simultaneously.

Sequential and Mutual (SeM) Subjects are randomly paired into groups of two. In a group with subject A and subject B, A is the first mover, who first observes a dice roll and then reports a number that determines B's payoff. Then, the second mover B observes A's report, as well as a dice roll and then reports a number that determines A's payoff.

Sequential and Non-mutual (SeNM) Subjects are randomly paired into groups of three, namely subjects A, B and C. A is the first mover, who observes a dice roll and reports a number that determines B's payoff. Then, the second mover B observes A's report, as well as a dice roll and then reports a number for C. C serves as a terminal "receiver", receiving the report from B but reporting for no one.

Let us explain first the role of the first two conditions, SiM and NM. It is straightforward to see that SiM corresponds to the pure "simultaneous setting" of the model. In addition, with NM we can test an ancillary, non-model based (but preregistered) hypothesis. Using both SiM and NM, we shall isolate the effect of group membership, in particular the tendency to treat others well in small groups, comparing it with the purely anonymous environment in NM. For instance, people may instinctively extrapolate from a current interaction into the future, expecting relationships to last. Alternatively, the notion of belonging to a group may induce the feeling of social proximity, and hence positive other-regarding preferences.

A very important experimental condition is SeM, which corresponds to the sequential setting of our model. The sequential and mutual nature of the interaction allows for the prospect of tangible advantages of being lenient in the first period: a subject will be assessed in the second period by the same person whom she may treat leniently now. The second mover observes the first mover's report and decides whether she reciprocates or retaliates. In this sequential setting, reciprocity plays a potential role, which we are testing. Finally, condition SeNM is used to examine another preregistered hypothesis outside the main model (but informing its scope conditions). It examines the hypothesis that in small fields, a "culture" of being lenient to others may develop impersonally, rather than by directly reciprocating lenience.[13]

Our Preregistered Hypotheses

Our theoretical model induces discipline by generating very concrete hypotheses. However, we also opted to examine a few hypotheses that fall outside the scope of the model, and to make sure that these are preregistered. This "ties our hands" defining a clear set of confirmatory hypotheses: four theory-driven and two additional preregistered ones. The remaining statistical patterns and results will be of an exploratory nature.

Preregistered Hypothesis 1 *The distributions of participants' reports in condition NM differs from condition SiM in terms of means. In particular, the mean in condition SiM is higher than in condition NM.*

Preregistered Hypothesis 2 *Reporting behavior of 'B players' in condition SeNM depends on the reports of 'A players'.*

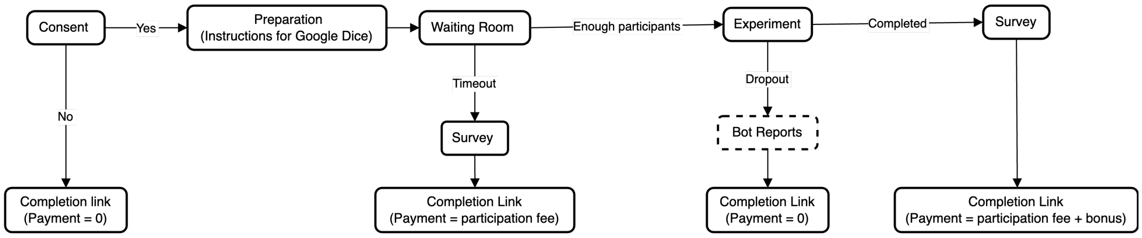
Implementation of the Experiments

We programmed the experiments using oTree (Chen et al., 2016) and conducted these sessions online via the platform 'Prolific'. Given the dynamic nature of online experiments, subjects could potentially withdraw at any stage. Moreover, certain

[13] If this was placed in a theoretical framework, it would correspond to an environment where altruism parameters are heterogeneous and the distribution of altruism parameters in the population is unknown. Observing kind behavior by the first mover may allow the second mover to infer a "better" distribution, and hence to treat others better.

experimental conditions necessitated the formation of groups with multiple participants online simultaneously. Hence, our experimental flow exhibited differences relative to standard lab sessions (see Figure 1).

Figure 1: Online Experimental Flow



To begin the experiment, participants log into their Prolific account and click on our oTree link. The first page they see is the "Consent" page, where they choose to attend the experiment ("Yes") or quit ("No"). Selecting "Yes" takes them to the "Preparation" page, where they receive instructions for setting up Google Dice. Selecting "No" redirects them to the "Completion link" with no payment. When the required number of participants is not achieved (for instance, conditions SiM and SeM require an even number of participants online), some subjects may be placed in a waiting room. If the required number is reached within a certain time frame, the experiment begins. Otherwise, waiting room participants are redirected to the survey and compensated with a fixed participation fee.

During the experiment, timeouts are set for all pages to handle potential dropouts. If a participant exceeds the timeout, they are considered a dropout and do not receive any payment. They are then replaced by a bot that reports a random number. The participant who was grouped with the dropout, without being aware of it, receives a bonus based on the bot's report. This data point is excluded from further analysis. Finally, the remaining participants are asked to complete a survey and are informed about their earnings, which are paid via Prolific the following day. The comprehensive set of instructions of the four experimental conditions can be found in Appendix A.

Experimental sessions for conditions NM, SiM, and SeM took place in September 2023, and sessions for the SeNM condition were held in October 2023. Participants were recruited randomly from the UK and the US populations. The total participants for NM, SiM, SeM, and SeNM conditions were 174, 187, 331, and 508, respectively. However, taking into account attrition (dropouts and participants who grouped with dropouts), and the fact that player C's in the SeNM condition do not report anything, the actual numbers of observations for each condition are 165 (NM), 163 (SiM), 302 (SeM), and 310 (SeNM). The median duration of the experiment was no more than 5 minutes. The average earnings for each subject were £4.32, including a fixed participation fee of £1.5.

4 Results

Some demographic characteristics of the participants are shown in Table 1. There are no particular differences in basic demographics across conditions, which are relatively similar in age, gender composition and religious affiliation. Figure 2 presents the overall average reporting across treatments (separating first from second movers in the sequential conditions). As can be seen, there is some evidence that first movers in the sequential treatments (even the non-mutual one) report high.[14] Note that in all treatments except NM participants were matched with the person they assessed in a small group (of two or three persons). Interestingly, the lowest reports can be found for the NM treatment, the most anonymous one.

In Figure 3a, we present in detail the reporting behavior per individual treatment. Across the board, there seems to exist some tendency for subjects to over-report in the direction of reporting 5. The model's predictions focus on the behavior of first movers in the (main) treatment SeM. As can be seen in Figure 3b, there is a high fraction of first-movers that report 5 in treatment SeM. A similarly high fraction

[14] Please note that in our experiment, reporting a dice roll of 6 results in a payoff of 0. The following analysis focuses on the "reported payoff " (ranging from 0 to 5) rather than the dice roll, as in Fischbacher and Föllmi-Heusi (2013).

Table 1: Basic Demographics

	N	Mean Age	Female (%)	No Religion (%)
NM	165	39.69	54.5	57.0
SiM	163	40.09	53.4	55.2
SeM	302	39.31	57.0	63.6
SeM: First	151	40.02	58.3	60.9
SeM: Second	151	38.60	55.6	66.2
SeNM	310	39.79	56.5	62.3
SeNM: First	155	41.99	57.4	57.4
SeNM: Second	155	37.59	55.5	67.1

Notes For the Sequential and Non-mutual (SeNM) treatment, we actually have a total of 465 participants, with 155 participants for each role. However, one-third of these participants are receivers who did not report anything. Therefore, the number of observations we have for SeNM is 310.

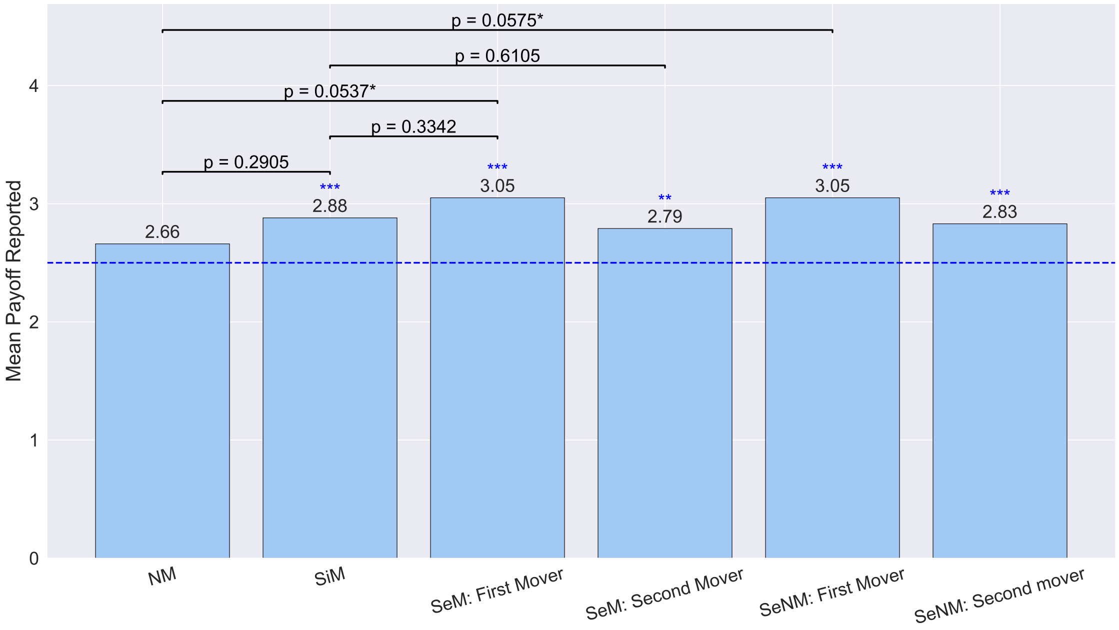
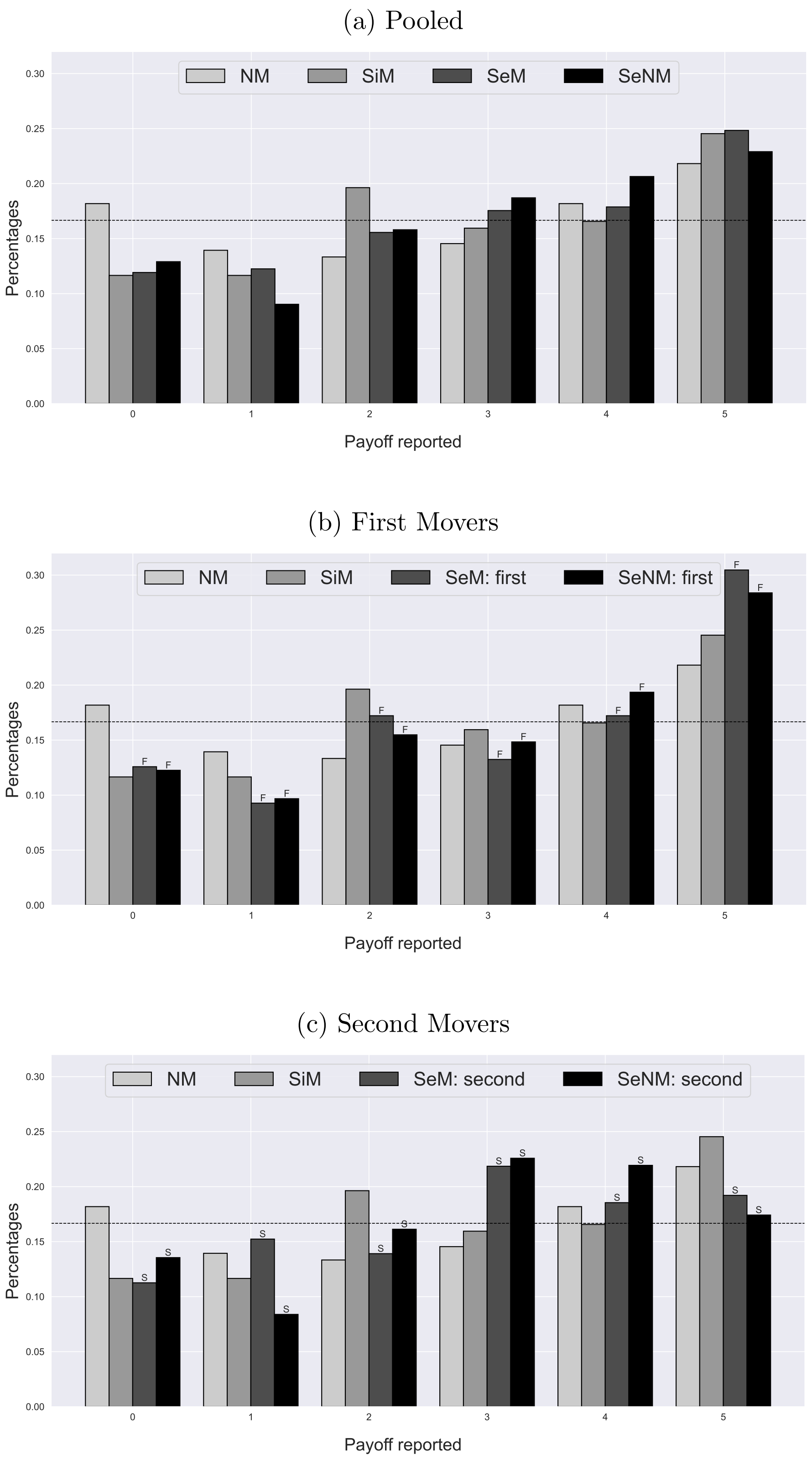


Figure 2: Overall mean reporting for all conditions.

Note: (1) The blue stars denote the significance levels of the Wilcoxon signed-rank test, which examines whether the median report exceeds 2.5 (the expected median report in the absence of lying). $*p < 0.1$, $**p < 0.05$, and $***p < 0.01$. (2) The p-values correspond to the two-sided Mann-Whitney U Test that assesses the mean differences between treatments.



(a) Pooled

(b) First Movers

(c) Second Movers

Figure 3: Reporting Distributions of All Conditions

Note: The dashed black lines represent 1/6. ‘F’ and ‘S’ above the bars represents ‘First mover’ and ‘Second mover’.

can be observed in the second sequential treatment we examine, namely SeNM. Juxtaposing the reporting behavior of first and second movers in the two sequential treatments (with the help of Figure 3c), we observe that while the reporting of first-movers indicates over-reporting of number 5, for second-movers the distribution is more even, with a much weaker tendency to report high. We can now focus in detail on the statistical evidence for our main experimental hypotheses.

4.1 Assessment of Hypotheses and Exploratory Analysis

Our model and our preregistration provide discipline and rigor in our data analysis, specifying clear patterns that the data should follow in order to support our hypotheses.

Hypothesis 1 *In condition SeM and in condition SiM all movers over-report.*

This hypothesis comes from Prediction 1 of our model. First of all, as can be seen from Table 2 when we pool the first and second movers in SeM, players over-report number 5 in their reports (p-values of binomial tests are less than 0.05). As Figure 2 shows, in all conditions, and for all types of movers, average reports exceed the unbiased expected report of 2.5, and an one-sample Wilcoxon signed-rank test indicates that the median reports of all conditions, except for NM, are significantly higher than 2.5. In addition, the Kolmogorov-Smirnov test and the chi-square test both reject the hypothesis that the distribution of reports is uniform under any experimental condition and for all types of movers.[15] This suggests that there is a systematic pattern of divergence from pure honesty, driven by the experimental environment, where lying in reports can benefit others, providing evidence for Hypothesis 1.

[15] The results of the Kolmogorov-Smirnov test indicate that the p-values are below 0.01 across all groups. The chi-square test yields p-values below 0.05 for all groups, with the exception of NM and the second movers in SeM, where the p-values are not significant.

Table 2: Share of payoff reported (in percent) and binomial tests

	0	1	2	3	4	5
NM	18.2	13.9	13.3	14.5	18.2	21.8*+
SiM	11.7*--	11.7*--	19.6	16	16.6	24.5**+++
SeM	11.9**--	12.3**--	15.6	17.5	17.9	24.8***+++
SeM: First	12.6	9.3**---	17.2	13.2	17.2	30.5***+++
SeM: Second	11.3*--	15.2	13.9	21.9+	18.5	19.2
SeNM	12.9*--	9.0***---	15.8	18.7	20.6*++	22.9***+++
SeNM: First	12.3-	9.7**---	15.5	14.8	19.4	28.4***+++
SeNM: Second	13.5	8.4***---	16.1	22.6*++	21.9*+	17.4

(a) Two-sided binomial test, $*p < .1, **p < .05, ***p < .01$
(b) One-sided binomial test (greater then 100%/6), $+p < .1, ++p < .05, +++p < .01$
(c) One-sided binomial test (less then 100%/6), $-p < .1, --p < .05, ---p < .01$

Hypothesis 2 *Participants in condition SiM and second movers in condition SeM report the same.*

Hypothesis 2 directly follows from Prediction 3 of our model. To assess equivalence, we use a Two One-Sided Test (TOST). The null hypothesis is that $|mean(SiM) - mean(SeM : second)| > \delta$, where δ is the equivalence margin representing the largest acceptable difference between the group means. Instead of pre-specifying a value for δ, we adopt a significance level of 0.05 and solve for the requisite δ that achieves this threshold. A permutation test revealed that our data only permit rejection of the null hypothesis of meaningful differences (and support equivalence) when the margin δ is at least 0.4, which is a relatively modest margin. Also, second movers' average report (2.79) is very similar to average report for SiM (2.88). We thus conclude that the evidence provides some support for this hypothesis.[16]

Hypothesis 3 *In condition SeM, first-movers report on average higher that in condition SiM.*

This hypothesis is derived from Prediction 2 of our model, which predicts that first-movers in the sequential environment will report higher than those in the simultaneous environment. As shown in Figure 2, the average reports are 2.88 in SiM

[16] A t-test is also performed, yielding a smallest δ of 0.331 when the significance level of 0.1, and 0.4 when the significance level reduce to 0.05. However, we should exercise caution in interpreting the t-test results due to the non-normality of the data.

and 3.05 for first movers in SeM. Again, the direction of the evidence is consistent with our model. However, the difference is not significant ($p = 0.3342$) and thus we lack sufficient evidence to support Hypothesis 3 statistically.

Hypothesis 4 *There is a positive correlation between the reports of First Movers and Second Movers in treatment SeM.*

This hypothesis is derived from Prediction 4 of our model, which predicts a positive correlation between the behavior of the two movers in the sequential and mutual condition. As Table 3 shows, the data indicate that the correlation is positive and statistically significant at the 10% level. Consequently, the data provide some evidence to support Hypothesis 4.

Hypothesis 5 *Participants in SiM report more than participants in NM.*

This hypothesis corresponds to our first preregistered hypothesis. The key difference between SiM and NM lies in the group cohesion participants might feel in SiM, in contrast to the one-sided/non-mutual nature of NM. We hypothesized that SiM participants, due to the sense of belonging to a group, might show more kindness towards one another, potentially yielding higher reports. While we have observed a lower mean report in NM compared to SiM, this difference lacks statistical significance (Kolmogorov-Smirnov: $p = 0.509$, Permutation: $p = 0.2658$, MWU: $p = 0.2905$, see Table B.2).

Table 3: Correlations between the reports of first and second movers

	beta coef (OLS)	**Pearson correlation**	**Spearman correlation**
SeM	1.726*	0.14*	0.147*
SeNM	-0.397	-0.032	-0.038

Notes $*p < .1, **p < .05, ***p < .01$

Hypothesis 6 *There is a positive correlation between the behavior of 'Player Bs' in and 'Player As' in treatment SeNM.*

This corresponds to preregistered hypothesis 2. As Table 3 shows, the data reveal no evidence in favour of this hypothesis, as the actual correlation is negative. This indicates that generalized reciprocity does not seem to play a strong role in the behavior of our participants. It is worth noting that this lack of correlation is consistent with our main model, which assumes only direct reciprocity, but not generalized one.

Sequential versus Non-mutual

Additional exploratory analysis shows that first movers in sequential conditions (both SeM and SeNM) report more than participants in condition NM. Getting back to our motivating example of the academic peer review process, NM simulates a well-implemented peer review process preserving full anonymity, while SeM mimics a process where anonymity cannot be fully preserved in a small field. The mean report for condition NM is 2.66, while for the first movers of SeM it is equal to 3.05. The results of the MWU test, as illustrated in Figure 2, reveal a statistically significant difference in average payoffs at the 10% significance level. Furthermore, a Fisher-Pitman Permutation test was conducted, yielding a significance level of 10% ($p = 0.0608$). Notably, the effect remains significant when the assessment is sequential but non-mutual. When comparing NM to the first movers in SeNM, the results of both the MWU test and Permutation test signal at the 10% level. Therefore, we receive evidence about the importance of anonymity (or the "size of the academic field") at the exploratory level.

To summarize our findings: in all conditions the reporting distribution deviates from the uniform one, and average reports exceed the unbiased expectation of 2.5. These effects are more prolonged in the conditions where subjects were matched in a group together with the participants for whom they reported. Differences in average reports were typically not statistically significant in comparisons between treatments, except that first movers in sequential conditions report significantly higher compared to NM. Overall, the data followed the patterns predicted by all four predictions of our theoretical model, although some of them were not significant at conventional levels.

4.2 Regression Analysis

To further investigate the between-treatment effects, we employ the linear model shown in Equation (5) where the dependent variable ($Report_i$) is the payoff associated with the reported dice roll of participant i. The model is specified as follows:

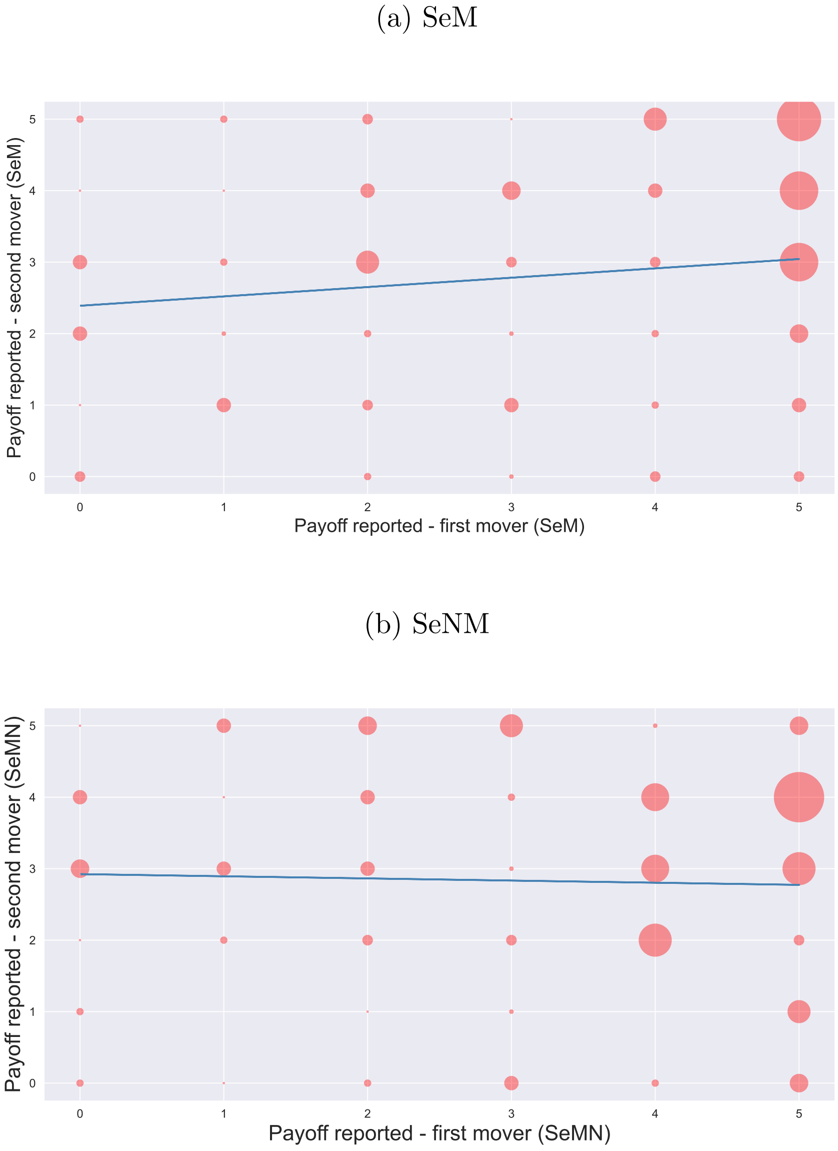
$$Report_i = \beta_0 + \beta_1 SiM_i + \beta_2 SeM_i + \beta_3 SeNM_i + \boldsymbol{X_i}\boldsymbol{\gamma} + \epsilon_i, \tag{5}$$

where SiM_i, SeM_i and $SeNM_i$ indicate whether it is an observation in experimental condition SiM, SeM and $SeNM$, respectively. The vector of participant characteristics $\boldsymbol{X_i}$, encompasses a range of demographic and socio-economic indicators. These include: age (continuous variable), gender (categorical variable), employment status (binary indicator with 1 representing employed and 0 otherwise), educational attainment (binary indicator with 1 representing possession of a bachelor's degree or higher, and 0 otherwise), residency (binary indicator with 1 indicating residency in the USA and 0 otherwise), and religious status (binary indicator with 1 indicating religious and 0 otherwise). Standard errors are clustered at the session level to account for potential correlation within experimental sessions.[17]

In experimental conditions with sequential settings (SeM and SeNM), participants have distinct roles: the first mover and the second mover. The reports of paired first and second movers may be correlated, and introducing a role indicator would inevitably cause multicollinearity issues,[18] necessitating separate regression

[17] Due to the capacity of the oTree program, multiple sessions were conducted for each experimental condition to ensure sufficient sample size. For both conditions NM and SiM, we conducted three sessions, with 40 to 70 participants per session. For condition SeM, we conducted three sessions, each with 100 to 110 participants. For condition SeNM, we conducted four sessions, and one session had 36 participants, while the remaining sessions each had more than 150 participants.

[18] Consider a representative sample of four observations: a first mover in SeM, a second mover in


(a) SeM

(b) SeNM

Figure 4: First movers' reports against second movers'

Note: The bubble size represents frequency, and large bubbles in the plot are intentionally exaggerated for better visibility. The highest frequency in SeM is for the combination (5, 5), with a frequency of 11. The highest frequency in SeNM is for the combination (5, 4), with a frequency of 12. The complete table of all combinations is available in Appendix B.

analyzes for first movers and second movers. The OLS results are presented in Table 4 and Table 5, respectively. The regression analysis of first movers (Table 4) reveals that participants in the SeM condition report more than 0.37 units higher compared to the NM condition. Interestingly, first movers in the SeNM condition also report higher values with a comparable magnitude, providing some support for "generalized reciprocity". On the other hand, coefficients of SiM_i in both Table 4 and Table 5 are statistically insignificant.

Table 4: Effects of Experimental Conditions on Reported Payoffs (Second Movers Excluded)

	Dependent Variable: Payoff Reported						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SiM	0.210	0.210	0.214	0.217	0.226	0.220	0.217
	(0.214)	(0.215)	(0.215)	(0.216)	(0.207)	(0.203)	(0.203)
SeM	0.373**	0.373**	0.377**	0.382**	0.392**	0.388**	0.386**
	(0.165)	(0.165)	(0.162)	(0.159)	(0.154)	(0.155)	(0.148)
SeNM	0.398**	0.398**	0.398**	0.398**	0.411**	0.401**	0.385**
	(0.164)	(0.164)	(0.167)	(0.168)	(0.159)	(0.163)	(0.163)
Age	-0.008	-0.008	-0.008	-0.008	-0.007	-0.007	
	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	
Employed	0.098	0.098	0.108	0.107	0.085		
	(0.149)	(0.147)	(0.139)	(0.139)	(0.136)		
College-educated	-0.125	-0.125	-0.121	-0.122			
	(0.194)	(0.193)	(0.198)	(0.197)			
USA	-0.044	-0.044	-0.058				
	(0.178)	(0.178)	(0.186)				
Male	0.015	0.016					
	(0.113)	(0.114)					
Nonbinary gender	-0.494	-0.494					
	(0.662)	(0.661)					
Religious	-0.003						
	(0.098)						
Constant	3.010***	3.009***	2.984***	2.976***	2.879***	2.952***	2.661***
	(0.436)	(0.436)	(0.415)	(0.412)	(0.373)	(0.340)	(0.115)
Observations	634	634	634	634	634	634	634

Notes: The dependent variable is the payoff associated with the reported dice roll, which can be integers from 0 to 5. The indictor variables for "NM condition" and "female" are omitted. Robust standard errors in parentheses are clustered at the session level. $*p < .1, **p < .05, ***p < .01$

To examine how second movers respond to first movers' reports, we focus on

SeM, a first mover in SeNM, and a second mover in SeNM. The corresponding indicator vectors are given by $\mathbf{v}_{\text{SeM}} = (1, 1, 0, 0)$ and $\mathbf{v}_{\text{SeNM}} = (0, 0, 1, 1)$. Introducing a role indicator $\mathbf{v}_1 = (1, 0, 1, 0)$ would result in perfect collinearity with the interaction terms. Specifically, the interactions between SeM/SeNM and the role are captured by $\mathbf{v}_{\text{SeM1}} = (1, 0, 0, 0)$ and $\mathbf{v}_{\text{SeNM1}} = (0, 0, 1, 0)$, and $\mathbf{v}_1 = \mathbf{v}_{\text{SeM1}} + \mathbf{v}_{\text{SeNM1}}$.

Table 5: Effects of Experimental Conditions on Reported Payoffs (First Movers Excluded)

	Dependent variable: Payoff Reported						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SiM	0.217	0.217	0.221	0.216	0.221	0.220	0.217
	(0.221)	(0.220)	(0.219)	(0.215)	(0.208)	(0.203)	(0.203)
SeM	0.120	0.116	0.116	0.114	0.119	0.119	0.127
	(0.168)	(0.168)	(0.163)	(0.160)	(0.156)	(0.157)	(0.146)
SeNM	0.154	0.151	0.148	0.148	0.156	0.155	0.172
	(0.187)	(0.187)	(0.191)	(0.188)	(0.188)	(0.187)	(0.169)
Age	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	
Employed	0.030	0.027	0.034	0.032	0.018		
	(0.147)	(0.141)	(0.135)	(0.135)	(0.129)		
College-educated	-0.070	-0.068	-0.073	-0.073			
	(0.188)	(0.188)	(0.184)	(0.184)			
USA	0.082	0.083	0.069				
	(0.238)	(0.239)	(0.241)				
Male	0.226	0.223					
	(0.153)	(0.152)					
Nonbinary Gender	-0.282	-0.282					
	(0.556)	(0.556)					
Religious	0.035						
	(0.110)						
Constant	2.885***	2.895***	2.993***	3.009***	2.955***	2.971***	2.661***
	(0.444)	(0.434)	(0.403)	(0.391)	(0.366)	(0.342)	(0.115)
Observations	634	634	634	634	634	634	634

Notes: The dependent variable is the payoff associated with the reported dice roll, which can be an integer from 0 to 5. The indictor variables for "NM condition" and "female" are omitted. Robust standard errors in parentheses are clustered at the session level. $*p < .1, * * p < .05, * * *p < .01$

observations in experimental conditions SeM and SeNM, and estimate the following linear model:

$$Second_i = \beta_0 + \beta_1 First_i + \beta_2 SeM_i + \beta_3 (First_i \times SeM_i) + \boldsymbol{X_i}\boldsymbol{\gamma} + First_i \boldsymbol{X_i}\boldsymbol{\phi} + \epsilon_i, \quad (6)$$

where $Second_i$ denotes the payoff associated with the second mover i's reports, $First_i$ represents the payoff associated with the report that i observes (the first mover's report). The indicator SiM_i and the vector $\boldsymbol{X_i}$ have been defined in Equation 5, and $First_i \boldsymbol{X_i}\boldsymbol{\phi}$ denotes all interactions between the first mover's report and the demographic variables.

The OLS estimates of Equation 6 are presented in Table 6. Results show that second movers in the SeM condition report approximately 0.5 less units compared to the SeNM condition. Moreover, religious participants and those identifying as non-binary gender report lower amounts on average. However, the interaction terms of these variables with the first movers' reports are positive and statistically significant. This suggests that second movers in the SeM condition, religious individuals, and people who identify themselves as the third gender are more responsive to the first movers' generosity. The lack of significance in the "$First$" coefficient for the SeNM treatment suggests that second movers' actions are not influenced by the first movers' report. This observation further demonstrates that there is insufficient evidence to support Hypothesis 6.

In both Equation 5 and Equation 6, the dependent variables are ordinal, ranging from 0 to 5, with 5 denoting the maximum extent of potential lying. Given this ordinal nature, we supplement our analysis with Ordered Logistic Regression (OLR) for robustness. The results obtained from this method are consistent with the OLS estimates in terms of coefficient signs and statistical significance.[19] Notably, our pri-

[19] See Appendix B for comprehensive results. The analysis shows that the majority of coefficients exhibit lower p-values in the OLR model compared to OLS. Interestingly, certain specifications demonstrate an inverse relationship between the first mover's reports and the probability of the second mover reporting a higher level in treatment SeNM (Table B.6).

mary finding – that first movers in sequential settings (both SeM and SeNM) report systematically higher values – remains robust across these alternative specifications.

5 Discussion and Conclusions

Our results can be summarized as follows. In general, participants' behavior in our experiments is consistent with a willingness to over-report to benefit others, while also displaying some degree of lying aversion. These observations concur with the assumptions of our theoretical model, which provides us with predictions regarding the differences in reporting behavior between the simultaneous and sequential conditions of mutual rating. In accordance with these, we find that in the sequential condition, first movers report more on average than raters in the simultaneous one, while second movers report at a similar level. While the direction of these results agrees with the model's predictions, some of the effects are not large enough to be statistically significant given our design's power.

Can we attribute what we observe to strategic reciprocity? There is no conclusive response to this question from our data. On the one hand, we observe that the first movers' average reports in the the two sequential conditions are practically the same, even if only in one of them can raters expect some direct reciprocity from second movers. This points to the direction of over-reporting being driven mainly by a nonstrategic other-regarding motive, and not by the strategic reciprocity incentives we hypothesize. On the other hand, second movers' behavior across the two conditions is different. When ratings are mutual, second movers seem to (weakly) respond to the rating received by first movers, as our model predicts. Such responsiveness is absent in the non-mutual sequential treatment. Thus, "pay-it-forward" motives could explain first mover behavior in these two treatments, but not that of second movers. At the same time, strategic reciprocity could explain the behavior of both raters in the mutual rating condition as well as that of second movers in

Table 6: First Movers' Reports Against Second Movers'

	Dependent variable: Payoff Reported (Second Mover)				
	(1)	(2)	(3)	(4)	(5)
First	-0.227	-0.216	-0.193	-0.080	-0.030
	(0.234)	(0.159)	(0.173)	(0.044)	(0.026)
SeM	-0.475*	-0.459*	-0.509**	-0.543**	-0.534**
	(0.235)	(0.206)	(0.190)	(0.174)	(0.191)
Age	-0.015	-0.016	-0.017		
	(0.013)	(0.012)	(0.011)		
Religious	-0.359	-0.400**	-0.449	-0.480*	
	(0.251)	(0.163)	(0.239)	(0.237)	
Nonbinary Gender	-0.731*	-0.777***	-0.854***	-0.716***	
	(0.316)	(0.138)	(0.217)	(0.182)	
Male	0.198	0.164			
	(0.414)	(0.452)			
Employed	-0.056				
	(0.192)				
College-educated	-0.085				
	(0.559)				
USA	-0.195				
	(0.828)				
SeM×First	0.140*	0.134***	0.159***	0.165***	0.161***
	(0.066)	(0.035)	(0.035)	(0.032)	(0.029)
Age×First	0.003	0.003	0.003		
	(0.004)	(0.004)	(0.003)		
Religious×First	0.141*	0.152*	0.139	0.143	
	(0.069)	(0.066)	(0.089)	(0.087)	
Nonbinary Gender×First	0.470***	0.489***	0.450***	0.432***	
	(0.105)	(0.090)	(0.101)	(0.080)	
Male×First	0.104	0.116			
	(0.082)	(0.093)			
Employed×First	0.020				
	(0.087)				
College-educated×First	-0.009				
	(0.158)				
USA×First	0.065				
	(0.231)				
Constant	3.623***	3.531***	3.694***	3.075***	2.924***
	(0.505)	(0.336)	(0.436)	(0.115)	(0.125)
Observations	306	306	306	306	306

Notes: The dependent variable is the payoff associated with reported dice rolls of the second movers, which range from 0 to 5. "First" is the payoff associated with the report of the corresponding first mover. The indictor variable for "female" is omitted. Robust standard errors in parentheses are clustered at the session level. $*p < .1, **p < .05, ***p < .01$

the non-mutual condition, but not that of first movers in this condition. Further research would be necessary to provide a more definite answer to this puzzle.

One of the motivations behind our experimental study was to investigate whether individuals possess the strategic sophistication necessary to realize the potential payoff of reciprocal rating behavior in a sequential mutual rating scenario. The one-shot nature of the interaction ensures that any observation of strategic behavior cannot be attributed to learning or the formation of specific norms. Arguably, the required cognitive abilities are higher for a first mover. This might explain why it is only for second movers that we observe behavior, namely their responsiveness to first mover ratings, that we can confidently claim to be evidence of strategic reciprocity. Nevertheless, one could imagine that in an environment where such interactions occur repeatedly, first movers might learn that exaggerated positive ratings can benefit them, or a social norm of providing gracious ratings to others might evolve.

Peer assessment is a crucial building block of modern performance management systems in organizations and still forms the backbone of most academic evaluation exercises, in the form of peer review. Despite the vast literature studying such systems and the acknowledgment that rating inflation is an important concern, only recently have scholars started to take a closer look to the possibility of strategic reciprocity and its effect on the reliability and effectiveness of peer assessment (Artz et al., 2023; Franke and Papadopoulos, 2024). Of course, by considering peer-evaluation in the publication process, the reader might be reminded of their own experiences with unfriendly reviewers, leading them to question the idea of positive reciprocity playing a role in the process. However, even in many areas within academia, direct peer competition may not be a concern. For example, in grant funding allocation, evaluators are often selected from a set that is disjoint to that of the reviewed scholars. Secondly, the traditional peer review system is often criticized and alternative approaches are proposed. Most of these retain some form of peer assessment, but differ in other dimensions, such as the degree of anonymity they require, the type and direction of accountability they impose on participants, and the level of peer competition. It is important to understand the potential of such innovations for introducing incentives for strategic reciprocity into the system, and this work makes a first step in this direction.

References

Alempaki, D., G. Doğan, and S. Saccardo (2019): "Deception and Reciprocity," *Experimental Economics*, 22, 980–1001.

Artz, M., C. Deller, and S. Leonelli (2023): "You Rate Me and I'll Rate You: Mutual Rating Relationships in Multi-Rater Performance Evaluation Systems," *Available at SSRN 4310277*.

Balietti, S., R. L. Goldstone, and D. Helbing (2016): "Peer review and competition in the Art Exhibition Game," *Proceedings of the National Academy of Sciences*, 113, 8414–8419.

Bamberger, P. A., I. Erev, M. Kimmel, and T. Oref-Chen (2005): "Peer assessment, individual performance, and contribution to group processes: The impact of rater anonymity," *Group & Organization Management*, 30, 344–377.

Bornmann, L. (2011): "Scientific peer review," *Annual review of information science and technology*, 45, 197–245.

Buckle, G. E., S. Füllbrunn, and W. J. Luhan (2021): "Lying for others: The impact of agency on misreporting," *Economics Letters*, 198, 109677.

Cappelli, P. and M. J. Conyon (2018): "What do performance appraisals do?" *ILR Review*, 71, 88–116.

Carpenter, J., P. H. Matthews, and J. Schirm (2010): "Tournaments and office politics: Evidence from a real effort experiment," *American Economic Review*, 100, 504–517.

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Colzani, P., G. Michailidou, and L. Santos-Pinto (2023): "Experimental evidence on the transmission of honesty and dishonesty: A stairway to heaven and a highway to hell," *Economics Letters*, 231, 111257.

Dato, S., E. Feess, and P. Nieken (2019): "Lying and reciprocity," *Games and*

Economic Behavior, 118, 193–218.

DeNisi, A. (2003): *A cognitive approach to performance appraisal*, Routledge.

Di Fiore, A. and M. Souza (2021): "Are peer reviews the future of performance evaluations," *Harvard Business Review*.

Erat, S. and U. Gneezy (2012): "White lies," *Management Science*, 58, 723–733.

Fehr, E. and U. Fischbacher (2003): "The nature of human altruism," *Nature*, 425, 785–791.

Fehr, E. and B. Rockenbach (2004): "Human altruism: economic, neural, and evolutionary perspectives," *Current opinion in neurobiology*, 14, 784–790.

Fischbacher, U. and F. Föllmi-Heusi (2013): "Lies in Disguise-an Experimental Study on Cheating: Lies in Disguise," *Journal of the European Economic Association*, 11, 525–547.

Fleenor, J. W., S. Taylor, and C. Chappelow (2020): *Leveraging the impact of 360-degree feedback*, Berrett-Koehler Publishers.

Franke, J. and A. Papadopoulos (2024): "Strategic Reciprocity in a Contest with Large Stakes," .

Frederiksen, A., L. B. Kahn, and F. Lange (2020): "Supervisors and performance management systems," *Journal of Political Economy*, 128, 2123–2187.

Gilbert, M. (2018): "Student performance is linked to connecting effectively with teachers," *Journal of Research in Innovative Teaching & Learning*, 12, 311–324.

Gintis, H., S. Bowles, R. Boyd, and E. Fehr (2003): "Explaining altruistic behavior in humans," *Evolution and human Behavior*, 24, 153–172.

Gneezy, U. (2005): "Deception: The role of consequences," *American Economic Review*, 95, 384–394.

Harari, M. B. and C. W. Rudolph (2017): "The effect of rater accountability on performance ratings: A meta-analytic review," *Human Resource Management Review*, 27, 121–133.

Huang, Y., M. Shum, X. Wu, and J. Z. Xiao (2019): "Discovery of bias and strategic behavior in crowdsourced performance assessment," *arXiv preprint arXiv:1908.01718*.

Hussam, R., N. Rigol, and B. N. Roth (2022): "Targeting high ability entrepreneurs using community information: Mechanism design in the field," *American Economic Review*, 112, 861–898.

Kane, J. S. and E. E. Lawler (1978): "Methods of peer assessment." *Psychological bulletin*, 85, 555.

Lee, C. J., C. R. Sugimoto, G. Zhang, and B. Cronin (2013): "Bias in peer review," *Journal of the American Society for information Science and Technology*, 64, 2–17.

Levine, D. K. (1998): "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, 1, 593–622.

London, M. and J. W. Smither (1995): "Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance-related outcomes? Theory-based applications and directions for research," *Personnel psychology*, 48, 803–839.

London, M., J. W. Smither, and D. J. Adsit (1997): "Accountability: The Achilles' heel of multisource feedback," *Group & Organization Management*, 22, 162–184.

Love, K. G. (1981): "Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction." *Journal of Applied Psychology*, 66, 451.

Malmendier, U., V. L. te Velde, and R. A. Weber (2014): "Rethinking reciprocity," *Annu. Rev. Econ.*, 6, 849–874.

Moers, F. (2005): "Discretion and bias in performance evaluation: the impact of diversity and subjectivity," *Accounting, Organizations and Society*, 30, 67–80.

Morgan, J., S. Neckermann, and D. Sisak (2020): "Peer evaluation and team performance: An experiment on complex problem solving," Tech. rep., mimeo.

Murphy, K. R., J. N. Cleveland, A. L. Skattebo, and T. B. Kinney

(2004): "Raters who pursue different goals give different ratings." *Journal of applied Psychology*, 89, 158.

Nicholas, D., A. Watkinson, H. R. Jamali, E. Herman, C. Tenopir, R. Volentine, S. Allard, and K. Levine (2015): "Peer review: Still king in the digital age," *Learned Publishing*, 28, 15–21.

Nishii, L. H. and D. M. Mayer (2009): "Do inclusive leaders help to reduce turnover in diverse groups? The moderating role of leader–member exchange in the diversity to turnover relationship." *Journal of applied psychology*, 94, 1412.

Olckers, M. and T. Walsh (2022): "Manipulation and peer mechanisms: a survey," *arXiv e-prints*, arXiv–2210.

Prendergast, C. and R. Topel (1993): "Discretion and bias in performance evaluation," *European Economic Review*, 37, 355–365.

Riedl, C., T. Grad, and C. Lettl (2024): "Competition and Collaboration in Crowdsourcing Communities: What happens when peers evaluate each other?" *Organization Science*.

Saal, F. E., R. G. Downey, and M. A. Lahey (1980): "Rating the ratings: Assessing the psychometric quality of rating data." *Psychological bulletin*, 88, 413.

Smither, J. W., M. London, and R. R. Reilly (2005): "Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings," *Personnel psychology*, 58, 33–66.

Sol, J. (2016): "Peer evaluation: Incentives and coworker relations," *Journal of Economics & Management Strategy*, 25, 56–76.

Stelmakh, I. (2022): "Making Scientific Peer Review Scientific," Ph.D. thesis, Carnegie Mellon University.

Stelmakh, I., N. B. Shah, and A. Singh (2021): "Catch me if i can: Detecting strategic behaviour in peer assessment," *Proceedings of the AAAI Conference on Artificial Intelligence*, 35, 4794–4802.

Tomkins, A., M. Zhang, and W. D. Heavlin (2017): "Reviewer bias in singleversus double-blind peer review," *Proceedings of the National Academy of Sciences*, 114, 12708–12713.

Tornow, W. W. (1993): "Editor's note: Introduction to special issue on 360-degree feedback," *Human Resource Management*, 32, 211.

Tsui, A. S. and B. Barry (1986): "Interpersonal affect and rating errors," *Academy of Management Journal*, 29, 586–599.

Wang, X. M., K. F. E. Wong, and J. Y. Kwong (2010): "The roles of rater goals and ratee performance levels in the distortion of performance ratings." *Journal of Applied Psychology*, 95, 546.

Wiles, J. (2018): "Peer feedback boosts employee performance," .

Wiltermuth, S. S. (2011): "Cheating more when the spoils are split," *Organizational Behavior and Human Decision Processes*, 115, 157–168.

Wong, K. F. E. and J. Y. Kwong (2007): "Effects of rater goals on rating patterns: Evidence from an experimental field study." *Journal of Applied Psychology*, 92, 577.

Zhu, X., P. Turney, D. Lemire, and A. Vellino (2015): "Measuring academic influence: Not all citations are equal," *Journal of the Association for Information Science and Technology*, 66, 408–427.

A Experimental Instructions

Instructions

♢ One-sided/non-mutual

At the beginning of the experiment, you will be paired with a randomly selected other participant from the *Prolific* participant pool. You will determine this participant's payoff by reporting your dice roll.

Meanwhile, there is **another** randomly selected participant, who will determine your payoff by reporting their own dice roll. **Note that *Prolific* has a huge participant pool, and this participant will not be the one you are paired with at the beginning.**

As explained, your report of the dice roll determines some other participant's payment. You can see the exact payoff from the following chart.

Number rolled	1	2	3	4	5	6
Resulting payoff	£1	£2	£3	£4	£5	£0

You will also receive a £1.50 participation fee.

Note: **Due to the nature of online experiments, participants may be in different time zones. If they cannot submit their dice roll at the same time as you, you won't know your dice-roll related payment at the end of experiment. We will inform you about your total payment in a couple of days.**

♢ Simultaneous and Mutual

At the beginning of the experiment, you will be paired with a randomly selected other participant from the *Prolific* participant pool. You will determine the other participant's payoff by reporting your dice roll. Meanwhile, the other participant will determine your payoff by reporting their own dice roll.

As explained, **your report of your dice roll determines how much the other participant earns. The other participant's report of their dice roll determines how much you earn.**

You can see the exact payoff from the following chart.

Number rolled	1	2	3	4	5	6
Resulting payoff	£1	£2	£3	£4	£5	£0

You will also receive a £1.50 participation fee.

◇ Sequential and Mutual

At the beginning of the experiment, you will be paired with a randomly selected other participant from the *Prolific* participant pool. The experiment has two stages, and participants make decisions sequentially:

- In Stage 1, the first mover reports their dice roll. The second mover's payoff is determined by the first mover's report.
- In Stage 2, the second mover sees the first mover's report, reports their own dice roll. This report determines the first mover's payoff.

Whether you or the other participant moves first is determined randomly. As explained, **your report of your dice roll determines how much the other participant earns. The other participant's report of their dice roll determines how much you earn.**
You can see the exact payoff from the following chart.

Number rolled	1	2	3	4	5	6
Resulting payoff	£1	£2	£3	£4	£5	£0

You will also receive a £1.50 participation fee.

◇ Sequential and Non-mutual

At the beginning of the experiment, you will be part of a randomly selected group of **3 participants** from the *Prolific* participant pool. The 3 participants are randomly assigned to different roles: ***First Mover***, ***Second Mover*** and ***Receiver***.
The experiment has two stages. The ***First and Second Movers*** make decisions sequentially, and the ***Receiver*** does not need to make any decisions:

- In Stage 1, the ***First Mover*** reports their dice roll. The ***Second Mover***'s payoff is determined by the ***First Mover***'s report.
- In Stage 2, the ***Second Mover*** sees the ***First Mover***'s report, reports their own dice roll. This determines the ***Receiver***'s payoff.

Please note that roles are assigned randomly, and the payments for each role are as follows:

- The ***First mover*** receives a random integer payment between £0 and £5, and they does not know about the exact number until the end of the experiment.

- The report of the ***First mover***'s die roll determines how much the ***Second Mover*** earns.

- The report of the ***Second mover***'s die roll determines how much the ***Receiver*** earns.

You can see the exact payoff from the following chart.

Number rolled	1	2	3	4	5	6
Resulting payoff	£1	£2	£3	£4	£5	£0

You will also receive a £1.50 participation fee.

B Additional Tables

Table B.1: Tests for the uniformity of reports and for the median report

	Kolmogorov-Smirnov	χ^2	**Wilcoxon signed-rank**
NM	0.0000***	0.3731	0.1319
SiM	0.0000***	0.0364**	0.0029***
SeM	0.0000***	0.0011***	0.0000***
SeM: First	0.0000***	0.0002***	0.0001***
SeM: Second	0.0000***	0.2307	0.0162**
SeNM	0.0000***	0.0002***	0.0000***
SeNM: First	0.0000***	0.0011***	0.0001***
SeNM: Second	0.0000***	0.0216**	0.0048***

(a) Both Kolmogorov-Smirnov test and χ^2 test are two-sided. $*p < .1, **p < .05, ***p < .01$
(b) Wilcoxon signed-rank test checks if the median report is greater than 2.5 (the expected median report when nobody lies). $*p < .1, **p < .05, ***p < .01$

Table B.2: Mean Payoff Reported Differences Between Treatments

	Mean Difference	**Permutation**	**MWU**
(NM vs. SiM)	-0.2167	0.2658	0.2905
(NM vs. SeM: First)	-0.3858	0.0608*	0.0537*
(NM vs. SeM: Second)	-0.1275	0.5094	0.5979
(SiM vs. SeM: First)	-0.1691	0.3891	0.3342
(SiM vs. SeM: Second)	0.0892	0.6681	0.6105
(NM vs. SeNM: First)	-0.3846	0.0594*	0.0575*
(NM vs. SeNM: Second)	-0.1717	0.379	0.4944
(SiM vs. SeNM: First)	-0.1679	0.3924	0.3468
(SiM vs. SeNM: Second)	0.045	0.8153	0.7821
(SeM: First vs. SeNM: First)	0.0012	1	0.949
(SeM: Second vs. SeNM: Second)	-0.0442	0.8318	0.8146

(a) Mean Difference: The mean payoff reported of the first element in brackets minus the mean of the second element.
(b) Permutation: Fisher-Pitman Permutation Test; MUW: Mann-Whitney U Test
(c) All tests are two-sided. $*p < .1, **p < .05, ***p < .01$

Table B.3: The frequency of reports of first and second movers in SeM and SeNM

first mover	***second mover***	***frequency (SeM)***	***frequency (SeNM)***
0	0	4	3
	1	1	3
	2	5	1
	3	5	6
	4	1	5
	5	3	1
1	0	0	1
	1	5	0
	2	2	3
	3	3	5
	4	1	1
	5	3	5
2	0	3	3
	1	4	1
	2	3	4
	3	7	5
	4	5	5
	5	4	6
3	0	2	5
	1	5	2
	2	2	4
	3	4	2
	4	6	3
	5	1	7
4	0	4	3
	1	3	0
	2	3	9
	3	4	8
	4	5	8
	5	7	2
5	0	4	6
	1	5	7
	2	6	4
	3	10	9
	4	10	12
	5	11	6

Table B.4: Ordered Logistic Regression: Effects of Experimental Conditions on Reported Payoffs (Second Movers Excluded)

	Dependent variable: Payoff Reported						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SiM	0.197	0.196	0.201	0.205	0.216	0.212	0.205
	(0.215)	(0.216)	(0.214)	(0.216)	(0.203)	(0.199)	(0.202)
SeM	0.388**	0.389**	0.393**	0.398**	0.405***	0.402**	0.397***
	(0.168)	(0.170)	(0.163)	(0.159)	(0.155)	(0.156)	(0.149)
SeNM	0.392**	0.392**	0.392**	0.392**	0.409***	0.401***	0.384**
	(0.166)	(0.166)	(0.167)	(0.168)	(0.152)	(0.156)	(0.159)
Age	-0.008	-0.008	-0.008	-0.008	-0.007	-0.007	
	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	
Employed	0.090	0.091	0.096	0.095	0.080		
	(0.154)	(0.151)	(0.146)	(0.146)	(0.142)		
College-educated	-0.101	-0.102	-0.098	-0.097			
	(0.200)	(0.200)	(0.201)	(0.201)			
USA	-0.056	-0.055	-0.062				
	(0.187)	(0.186)	(0.196)				
Male	0.024	0.024					
	(0.126)	(0.126)					
Nonbinary Gender	-0.525	-0.525					
	(0.653)	(0.654)					
Religious	-0.023						
	(0.098)						
cut1	-1.950***	-1.946***	-1.923***	-1.912***	-1.833***	-1.902***	-1.606***
	(0.515)	(0.517)	(0.484)	(0.477)	(0.450)	(0.425)	(0.201)
cut2	-1.207**	-1.203**	-1.181***	-1.170***	-1.092***	-1.161***	-0.867***
	(0.472)	(0.472)	(0.437)	(0.431)	(0.398)	(0.353)	(0.124)
cut3	-0.447	-0.442	-0.421	-0.411	-0.333	-0.402	-0.110
	(0.443)	(0.444)	(0.408)	(0.404)	(0.372)	(0.331)	(0.111)
cut4	0.149	0.153	0.174	0.185	0.262	0.193	0.484***
	(0.423)	(0.424)	(0.388)	(0.386)	(0.344)	(0.302)	(0.106)
cut5	0.950**	0.954**	0.975**	0.986**	1.063***	0.993***	1.284***
	(0.434)	(0.436)	(0.398)	(0.394)	(0.354)	(0.317)	(0.109)
Observations	634	634	634	634	634	634	634

Notes: The dependent variable is the payoff associated with the reported dice roll, which can be integers from 0 to 5. The indictor variables for “NM condition” and “female” are omitted. Robust standard errors in parentheses are clustered at the session level. $*p < .1, * * p < .05, * * *p < .01$

Table B.5: Ordered Logistic Regression: Effects of Experimental Conditions on Reported Payoffs (First Movers Excluded)

	Dependent variable: Payoff Reported						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SiM	0.216	0.216	0.214	0.210	0.219	0.219	0.211
	(0.237)	(0.237)	(0.234)	(0.230)	(0.219)	(0.215)	(0.219)
SeM	0.104	0.103	0.099	0.097	0.103	0.104	0.108
	(0.163)	(0.164)	(0.157)	(0.155)	(0.151)	(0.152)	(0.143)
SeNM	0.120	0.118	0.122	0.123	0.133	0.133	0.149
	(0.178)	(0.179)	(0.185)	(0.182)	(0.182)	(0.182)	(0.165)
Age	-0.008	-0.008	-0.009	-0.009	-0.008	-0.008	
	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	
Employed	0.013	0.011	0.010	0.007	-0.007		
	(0.160)	(0.151)	(0.149)	(0.148)	(0.144)		
College-educated	-0.067	-0.066	-0.078	-0.078			
	(0.198)	(0.199)	(0.191)	(0.191)			
USA	0.051	0.051	0.056				
	(0.250)	(0.250)	(0.250)				
Male	0.232	0.231					
	(0.167)	(0.165)					
Nonbinary Gender	-0.260	-0.261					
	(0.765)	(0.762)					
Religious	0.013						
	(0.117)						
cut1	-1.998***	-2.003***	-2.110***	-2.124***	-2.064***	-2.058***	-1.723***
	(0.542)	(0.528)	(0.506)	(0.496)	(0.466)	(0.433)	(0.207)
cut2	-1.198**	-1.202**	-1.312***	-1.327***	-1.267***	-1.260***	-0.927***
	(0.503)	(0.487)	(0.456)	(0.441)	(0.412)	(0.367)	(0.131)
cut3	-0.478	-0.482	-0.594	-0.609	-0.549	-0.543	-0.212*
	(0.471)	(0.457)	(0.430)	(0.418)	(0.392)	(0.349)	(0.120)
cut4	0.282	0.278	0.163	0.148	0.208	0.214	0.541***
	(0.459)	(0.445)	(0.407)	(0.393)	(0.366)	(0.330)	(0.103)
cut5	1.200***	1.196***	1.079***	1.064***	1.123***	1.129***	1.455***
	(0.451)	(0.438)	(0.392)	(0.379)	(0.353)	(0.330)	(0.130)
Observations	634	634	634	634	634	634	634

Notes: The dependent variable is the payoff associated with the reported dice roll, which can be integers from 0 to 5. The indictor variables for "NM condition" and "female" are omitted. Robust standard errors in parentheses are clustered at the session level. $*p < .1, * * p < .05, * * *p < .01$

Table B.6: Ordered Logistic Regression: First Movers' Reports Against Second Movers'

	Dependent variable: Payoff Reported (Second Mover)				
	(1)	(2)	(3)	(4)	(5)
First	-0.189	-0.209	-0.120***	-0.077*	-0.031
	(0.261)	(0.160)	(0.044)	(0.045)	(0.030)
SeM	-0.546**	-0.516**	-0.529***	-0.588***	-0.598***
	(0.265)	(0.221)	(0.195)	(0.165)	(0.181)
Age	-0.016	-0.016			
	(0.012)	(0.012)			
Religious	-0.356	-0.411**	-0.437***	-0.493**	
	(0.232)	(0.162)	(0.169)	(0.230)	
Nonbinary Gender	-0.441***	-0.537**	-0.437	-0.517***	
	(0.166)	(0.244)	(0.275)	(0.173)	
Male	0.217	0.154	0.177		
	(0.417)	(0.481)	(0.440)		
Employed	-0.052				
	(0.217)				
College-educated	-0.142				
	(0.634)				
USA	-0.255				
	(0.842)				
SeM × First	0.171**	0.161***	0.161***	0.180***	0.182***
	(0.080)	(0.042)	(0.038)	(0.033)	(0.030)
Age × First	0.002	0.002			
	(0.005)	(0.004)			
Religious × First	0.128**	0.144**	0.144**	0.135	
	(0.058)	(0.060)	(0.063)	(0.084)	
Nonbinary× First	0.721***	0.770***	0.767***	0.710***	
	(0.165)	(0.164)	(0.168)	(0.152)	
Male × First	0.107	0.130	0.122		
	(0.084)	(0.098)	(0.090)		
Employed × First	-0.001				
	(0.099)				
College-educated × First	-0.011				
	(0.177)				
USA * First	0.081				
	(0.249)				
cut1	-2.872***	-2.754***	-2.141***	-2.248***	-2.086***
	(0.555)	(0.361)	(0.343)	(0.206)	(0.196)
cut2	-2.042***	-1.923***	-1.315***	-1.434***	-1.272***
	(0.477)	(0.256)	(0.294)	(0.066)	(0.129)
cut3	-1.315***	-1.196***	-0.589**	-0.720***	-0.562***
	(0.488)	(0.282)	(0.279)	(0.109)	(0.108)
cut4	-0.369	-0.254	0.350	0.202**	0.353***
	(0.501)	(0.298)	(0.303)	(0.102)	(0.129)
cut5	0.715	0.824**	1.425***	1.261***	1.392***
	(0.575)	(0.377)	(0.330)	(0.143)	(0.164)
Observations	306	306	306	306	306

Notes: The dependent variable is the payoff associated with reported dice rolls of the second movers, which range from 0 to 5. "First" is the payoff associated with the report of the corresponding first mover. The indictor variable for "female" is omitted. Robust standard errors in parentheses are clustered at the session level. $*p < .1, **p < .05, ***p < .01$