Peer Evaluations: Exploiting Meritocratic Norms to Overcome Social Dilemmas*

Alexander L. Brown† Matt Van Essen‡ John Wooders§

December 23, 2025

Abstract

This paper demonstrates how social norms for meritocratic evaluations can be exploited to incentivize good service, high effort, and contributions to the public good. It develops a theoretical model showing how peer evaluation norms (e.g., meritocratic, truth-telling, or collusive) affect equilibrium effort and welfare. Our experimental results demonstrate subjects tend to follow meritocratic norms, and that appropriately structured peer evaluations are an effective means of increasing contributions to the public good and welfare. We find that groups that follow more meritocratic norms achieve higher welfare than groups that follow less meritocratic norms.

*We are grateful to Georgy Artemov, Jim Cox, Martin Dufwenberg, Nikos Nikiforakis, Tim Salmon, Tom Wilkening, and Christian Vossler for comments. Wooders gratefully acknowledges financial support from Tamkeen under the NYU Abu Dhabi Research Institute Award CG005.

†Department of Economics, Texas A&M University, alexbrown@tamu.edu.

‡Department of Economics, University of Tennessee, mvanesse@utk.edu.

§Center for Behavioral Institutional Design, New York University Abu Dhabi, United Arab Emirates. Division of Social Science, New York University Abu Dhabi. john.wooders@nyu.edu.

1 Introduction

Settings in which a supervisor or manager observes the overall result of the efforts of her subordinates but not their individual efforts are ubiquitous: A supervisor in a government office observes average client satisfaction, but not the quality of each of her subordinates' interactions with clients. A store manager observes overall sales, but not each salesperson's contribution to sales. In the classroom, an instructor observes the quality of a group project, but not the effort that each student in the group contributed. Students and employees, by contrast, are often better informed about the contributions of their peers, as they observe their peers' efforts directly. To address this information asymmetry, organizations solicit peer evaluations to incentivize good service, high effort, and contributions to the public good.

How people evaluate their peers is governed by a combination of explicit incentives and social or organizational norms. Here we focus on the role of norms. We will say a norm is meritocratic if higher effort elicits higher evaluations, and is more meritocratic as evaluations are more responsive to effort. A norm is truth-telling if evaluations match effort. A norm is collusive, by contrast, if evaluations are favorable regardless of effort. In our theoretical model, the norms that govern evaluations will shape incentives for the provision of effort.[1]

This paper develops a theoretical model of peer evaluations and tests it experimentally. The model captures settings in which a principal (e.g., a supervisor, manager, professor) wishes to incentivize agents (e.g., subordinates, employees, students) to contribute effort for the public good. In the model, each agent simultaneously chooses how much effort to contribute and the total effort contributed determines the amount, or quality, of the public

[1] According to Bicchieri (2006), a social norm exists when people conform to a behavioral rule when they expect a sufficiently large subset of the population follows the same behavioral rule and they believe that the population expects them to conform to the rule (and may sanction them if they don't).

good. Each agent observes the effort contributed by every other agent. The principal, by contrast, observes only the quality of the public good provided.[2]

To incentivize contributions, the principal collects, from each agent, the agent's evaluation of their own contribution and the agent's evaluations of the contributions of their peers. Each agent's payoff is determined by a combination of the quality of the public good and the evaluations the agent receives. We model an agent's payoff as a weighted average of a "group grade" and the agent's evaluation-based "individual grade." The group grade is the same for all agents and equals the quality of the public good. An agent's idiosyncratic individual grade is a function of the median evaluation they receive. In particular, their individual grade is equal to what the quality of the public good would have been had every agent contributed an effort equal to the median evaluation the agent received. Individual grades are, in this sense, fair grades. The focus of our theoretical and empirical results is on how the weight placed in payoffs on the individual grade affects effort and welfare, and how it interacts with the norms governing peer evaluations.

Our theoretical focus is on subgame perfect equilibria of the peer evaluation game in which evaluations are truthful, i.e., evaluations are equal to effort contributions.[3] There is a unique such equilibrium, and, in this equilibrium, effort and welfare are increasing in the payoff weight placed on the evaluation-based individual grade. We show that effort reaches its welfare-maximizing level when payoffs are determined entirely by evaluations and are independent of the actual quality of the public good provided. In this case, each agent fully internalizes the payoff externality that their effort exerts on the other agents, and when each agent maximizes their individual grade they maximize social welfare as well.

[2] Our model applies equally to settings where individual efforts are observable but not verifiable (i.e., not contractible).

[3] Abeler, Nosenzo, and Raymond (2019), in meta study of preferences for truth-telling, conclude that subjects have a preference for being seen as honest and a preference for being honest.

There are also subgame perfect equilibria in which evaluations are not truthful. If evaluations follow a collusive norm, and are independent of effort, then increasing the weight placed on the individual grade undermines incentives to provide the public good, and reduces effort and welfare. In the setting we study, truth-telling is a welfare-enhancing social norm.

A key feature of our setting, and a common feature in practice, is that an agent's own payoff does not depend on the evaluations they give to their peers.[4] Furthermore, since evaluations are given after effort choices are made, and since the game is one-shot, there is no scope for the evaluations an agent gives to beneficially influence either the future effort choices of his peers or the evaluations they receive from their peers. An agent therefore has neither a direct or indirect incentive to be strategic when making peer evaluations. It is in this setting that we expect social norms to be most powerful since they do not compete with other incentives.[5]

Nevertheless, as subjects gain experience, they learn about the evaluation norms of their peers via the evaluations they receive, and they may adjust the norm they follow. We will see evidence of this. It is worth emphasizing that our theoretical model does not predict evaluation norms. Subjects bring to the laboratory their own evaluation norms.

In our baseline experiment, subjects participate in a public good game without evaluations. We then introduce evaluations, studying three versions of the peer evaluation game which differ in the weight placed on evaluations (via the individual grade) in the determination of payoffs. In the first treatment, a subject's payoff equals their group grade, i.e., subjects give and receive evaluations but evaluations have no direct impact on payoffs. In the second treatment, a subject's payoff is an equally weighted average of the group grade and their individual grade. In the third treatment, a subject's payoff equals their individual grade, i.e., a subject's payoff is determined

[4] The efforts contributed by peers are observed costlessly and without noise.

[5] At the same time, an agent does not have any positive incentive to evaluate the contributions of their peers accurately or honestly.

entirely by the evaluations they receives.

Our baseline experiment reduces to a non-linear voluntary contribution mechanism (VCM). Consistent with prior research on VCMs, we find that subjects contribute more than the Nash equilibrium effort, but contributions substantially decay as subjects gain experience.

In the first treatment, subjects give and receive peer evaluations, but evaluations are cheap-talk with no impact on monetary payoffs. We find that peer evaluations are a powerful nudge to contributing effort to the public good. In particular, the introduction of peer evaluations increases effort and welfare relative to the baseline. Effort decays over time, although more slowly than in the baseline treatment, but effort remains well above the baseline and Nash levels, even at the end of the experiment. This result shows that subjects obtain non-monetary rewards purely from giving and receiving evaluations, and this incentivizes contributions.

In the final two treatments, payoffs depend on evaluations. In the first of these, payoffs are an equally weighted average of group and individual grades. We find that effort and welfare both increase in this treatment: effort increases by 60% over the level when evaluations are a nudge, and effort approaches the level predicted by the subgame perfect equilibrium in which evaluations are truthful. Effort contributions increase over the entire course of the experiment. This result demonstrates that appropriately structured peer evaluations can be an effective means of increasing contributions to the public good and increasing welfare.

In the last treatment, each subject's payoff is determined entirely by the evaluations they receive. In this treatment, in the equilibrium with truth-telling, effort and welfare are predicted to reach their welfare-maximizing levels. This treatment is a stress test of the theory as subjects have a strong incentive to tacitly collude: If subjects follow the collusive evaluation norm in which subjects give the maximum possible evaluation regardless of their peers' efforts, then in equilibrium subjects obtain higher payoffs than in the

equilibrium with truth-telling. We find, indeed, that truth-telling breaks down in this treatment. Effort and welfare both decrease as a consequence. This result demonstrates that a well-designed peer-evaluation mechanism must weigh the trade off between welfare gains from a higher weight on peer-evaluations in payoffs, against the potential for participants to adopt a collusive evaluation norm.

Our theoretical results show that payoff-relevant peer evaluations increase equilibrium effort and welfare when evaluations are truthful. But do subjects tell the truth? In the treatment in which subjects are most truthful, we find that only 24% of evaluations match effort contributed. Nonetheless, while evaluations are less than fully truthful, evaluation norms are meritocratic and the evaluations by subjects of the efforts of their peers are accurate on average. Subjects' evaluations of their own effort, by contrast, are inflated relative to their actual effort, and this is especially true when payoffs are determined entirely by evaluations. These findings demonstrate that peer evaluations can be effective in raising effort and welfare, even if imperfectly truthful.

A subject's monetary incentive to contribute effort depends on the degree to which evaluations are meritocratic. We find that the subjects' evaluations of others are highly meritocratic when payoffs are an equally weighted average of the group grade and their own individual grade. The degree to which evaluations are meritocratic increases from the first half of the experiment to the second, and effort contributions increase as well. Furthermore, observed effort contributions are consistent with the degree of meritocracy observed in evaluations. By contrast, effort contributions decline over time when there are no evaluations or when evaluations don't enter in payoffs.

Evaluations are least meritocratic when payoffs are determined entirely by evaluations. The responsiveness of evaluations to effort declines from the first half of the experiment to the second half. Given that subjects receive favorable evaluations even when contributing zero effort, by the second half of

the experiment it is optimal to contribute no effort given the estimated evaluation norm. Subjects nevertheless contribute effort at an amount roughly equal to what they contribute when evaluations are purely a nudge.

Different equilibria may be played in different experimental sessions, as different norms for evaluating effort may develop in different sessions. When payoffs were determined entirely by evaluations, there was one session in which evaluations were largely truthful and, in this session, observed effort approached its welfare maximizing level. In another session of the same treatment, evaluations were highly inflated and observed effort took its lowest value across all sessions. For the two treatments with payoff-relevant evaluations, we find that effort is positively correlated across sessions with the truthfulness of evaluations. For the treatment with payoff-irrelevant evaluations, by contrast, effort was uncorrelated with the truthfulness of evaluations. These results demonstrate the social value of the truth-telling norm.

Related Literature

Game Theory and Social Norms

Kandori (1992) shows that there is a social norm that sustains cooperation in the Prisoner's Dilemma when the game is played among members of a community, with randomly selected opponents, and each player knows only their own history. According to the norm, a player forever defects after observing a defection; thus a single defection eventually spreads throughout the whole community. Kandori (1992) shows that when players are sufficiently patient, then the threat of contagious defection sustains cooperation. Kandori (1992), like the present paper, does not explain the origin of the norm. By contrast, Kandori, Mailath, and Rob (1993) studies an evolutionary model of behavior in 2×2 repeated coordination games. It shows that, as the mutation rate vanishes, the probability that the risk-dominant Nash equilibrium is played in the long run approaches one. In other words, a convention for how to play such games eventually emerges.

Krupka and Weber (2013) elicits social norms for behavior in several different framings of the Dictator game. It shows that different framings yield different outcomes, the result of different norms for giving versus taking. In the present paper, rather than eliciting norms, we show that the social norm for meritocratic evaluations can ameliorate social dilemmas.

Experimental Public Goods Games

A vast literature studies voluntary contributions mechanisms (VCM) for the provision of public goods. In a linear VCM with n players, each player i simultaneously chooses an amount to contribute g_i of their endowment ω_i to the public good. The payoff of player i is

$$\pi_i(g_1, \ldots, g_n) = \omega_i - g_i + \alpha \sum\nolimits_{i=1}^{N} g_i,$$

where α is the marginal per-capita return. When $\frac{1}{n} < \alpha < 1$, then it is a dominant strategy for each player i to choose $g_i = 0$, while total payoffs are maximized when each player chooses $g_i = \omega_i$. A robust experimental finding is that subjects make positive contributions, although contributions decline in repeated play. See Ledyard (1995) for an early survey. See Laury and Holt (2008) for a survey of results for non-linear VCMs with interior Nash equilibria. To establish a baseline for contributions without peer evaluations, our first treatment is a non-linear VCM.

A body of research has considered methods of fostering contributions in public good games when $g_i = 0$ is dominant strategy. Barron and Nurminen (2020) show that contributions increase if subjects are told in the instructions that a contribution above a certain threshold would be labelled as "good," and any other contribution would be labelled as "bad."

Fehr and Gächter (2000) modifies the VCM by adding a second stage in which players may, at a cost, punish co-players after observing their contributions. When groups are randomly formed at each round, it finds that the opportunity to punish increases contributions although, net of punishment

costs, it does not raise player payoffs. The punishment behavior observed is consistent with the social norm calling for punishing free riders, even when punishment is costly to the punisher. Cason and Gangadharan (2015) studies peer punishment in non-linear VCMs, arguing that "Nonlinear environments may be considered as more representative of many practical situations, since they typically lead to equilibrium and socially optimal choices that are not on the boundaries of the choice space." It finds that peer punishment is effective in reducing free riding, although the effect is weaker and takes more time to emerge than in linear environments.

Nikiforakis (2008) extends the VCM further by adding a third stage in which punished players can counter-punish their punishers. The addition of this third stage reduces the willingness of subjects to punish free riders, which in turn reduces contributions. Net of punishment costs, payoffs are lower when players can punish and counter-punish than in the baseline VCM. These results show that the opportunity to punish (and the associated possibility of counter-punishment) does not solve the underprovision of contributions.

Peer Pressure and Peer Evaluations

Kandel and Lazear (1992) model peer pressure in public good and partnership environments by adding a "peer pressure" function to payoffs, with the property that peer pressure is decreasing in effort exerted. Peer pressure might, for example, result from guilt or shame following a deviation from a social norm for effort provision, or it might result from explicit punishment from peers who monitor effort provision.

Carpenter, Robbet,and Akbar (2017) experimentally examine costly peer reporting in a profit sharing (linear VCM) environment. It is costly for the players submit a peer report and it is also costly for the manager to punish. They find that profit sharing increases effort relative to effort under a fixed wage, and profit sharing combined with peer reporting – "shirk" or "not shirk" – further increases effort. In our setting, payoffs are determined by a combination of peer evaluations and effort choices; we study both the-

oretically and experimentally how social norms for meritocratic evaluations determine effort.

Carpenter, Matthews, and Schirm (2010) report the result of a real-effort experiment with peer evaluations. They show that supplementing piece rate compensation with a tournament bonus raises effort when output is evaluated objectively, but reduces effort when output is evaluated by peers. Subjects use evaluations to sabotage their peers and, anticipating this, subjects reduce their effort. Dufwenberg, Gölitz, and Gravert (2024) also study peer evaluations in tournaments. In their setting, players have exogenously known qualities. Each player reports their own quality and all the other players' qualities; the tournament winner is the player with the highest total reported quality. They find experimentally that players with higher quality are more likely to win the tournament, despite an incentive for players to overreport own quality and underreport rivals' qualities. They show that this result is consistent with an equilibrium of a psychological game in which players obtain disutility when others believe they have cheated. Our analysis, by contrast, examines social norms under classical preferences.[6]

2 The Peer Evaluation Game

We introduce a model of peer evaluations in which players first contribute effort to a group project and then evaluate the contributions of the others. The total effort contributed determines the quality of the group project.

The Game

There are $N \geq 3$ players and two stages. At the first stage, each player simultaneously chooses an effort to devote to a project. Let $e_i \in \mathbb{R}^+$ denote

[6] The papers differs in two other respects, as well. In the present paper (i) effort/quality is endogenous and the focus is on peer evaluations as a means to incentivize high effort/quality, and (ii) peer evaluations are non-rivalrous – a player's evaluations of others has no effect on own payoffs.

the effort of player i and let $e = (e_1, \ldots, e_N)$ denote the profile of all the player efforts. At the second stage, each player observes the effort profile e and then gives an evaluation of every player, including one of himself. Let $\tilde{e}_i^j \in \mathbb{R}^+$ denote player i's evaluation of player j, let $\tilde{e}_i = (\tilde{e}_i^1, \ldots, \tilde{e}_i^N)$ denote i's profile of evaluations of all the players, and let $\tilde{e} = (\tilde{e}_1, \ldots, \tilde{e}_N)$ denote the profile of evaluation profiles. An *evaluation strategy* for player i is a profile of functions $\tilde{\varepsilon}_i = (\tilde{\varepsilon}_i^1, \ldots, \tilde{\varepsilon}_i^N)$, where $\tilde{\varepsilon}_i^j : \mathbb{R}^N \to \mathbb{R}^+$ maps an effort profile $(e_1, \ldots, e_N)$ to i's evaluation of j.

In the game, player i's payoff will be independent of his evaluation of player j. Hence i's evaluation of j's effort will be governed by a social norm. If i follows a meritocratic social norm, i.e., player i believes that high effort *should* be rewarded with high evaluations, then i's evaluation of j is increasing in j's effort e_j. Of special theoretical interest is the norm for truthtelling. This is the meritocratic norm in which evaluations match effort. We say that i *evaluates* j *truthfully* if $\tilde{\varepsilon}_i^j(e_1, \ldots, e_N) = e_j$ for all $(e_1, \ldots, e_N)$. We say that *evaluations are truthful* if i evaluates j truthfully for every i and j. A non-meritocratic norm, by contrast, is one in which evaluations are independent of effort. A special case is the collusive norm, in which each player i gives j the highest possible evaluation, regardless of j's effort.

A *strategy* for player i is a pair $(e_i, \tilde{\varepsilon}_i)$, where e_i is i's effort and $\tilde{\varepsilon}_i$ is i's evaluation strategy. Let $\tilde{e}^i = (\tilde{e}_1^i, \ldots, \tilde{e}_N^i)$ denote the profile of evaluations received by player i. The median evaluation of player i is denoted by $m(\tilde{e}^i)$.

Given a strategy profile $(e, \tilde{\varepsilon}) = ((e_i, \tilde{\varepsilon}_i)_{i=1}^N)$, the payoff to player i is determined by (i) a group grade which captures the quality of the final product and which is the same for every player, (ii) an idiosyncratic individual grade which is determined by the evaluations player i receives, and (iii) the cost of their effort. The *group grade* is determined by the total effort of the players and is given by

$$f(e_1 + \ldots + e_N),$$

where $f(0) = 0$, $f' > 0$, and $f'' < 0$. We assume that $f'(0) > 1$ and there is

an effort $\bar{e}$ such that $Nf'(N\bar{e}) < 1$.[7]

Player i's *individual grade* is determined by the median evaluation $m(\tilde{e}^i)$ they receives from the other players (their peers) and is given by

$$f(Nm(\tilde{e}^i)).$$

The individual grade $f(Nm(\tilde{e}^i))$ of player i with median evaluation $m(\tilde{e}^i)$ is "fair" in the sense that i's individual grade is equal to the group grade that would result were all the players to contribute an effort equal to i's evaluation.

The payoff to player i is therefore

$$\pi_i(e_1, \ldots, e_N, \tilde{e}^i_1, \ldots, \tilde{e}^i_N) = (1-\alpha)f(e_1 + \ldots + e_N) + \alpha f(Nm(\tilde{e}^i)) - e_i,$$

where $1-\alpha$ is the weight placed on the group grade, $\alpha \in [0,1]$ is the weight placed on the individual grade, and e_i is player i's effort cost. Player i's payoff is independent of his evaluations of other players.

Of primary theoretical and empirical interest is how the weight α placed on peer evaluations (via the individual grade) affects incentives to provide effort. When $\alpha = 0$, then payoffs are independent of evaluations and the game is equivalent to a non-linear voluntary contribution mechanism.[8] At the other extreme, when $\alpha = 1$, then each player's payoff is determined solely by the evaluations they receive and their own effort, and is independent of the quality of the final project.

Last, we define (per-capita) *welfare for effort profile* e to be the average payoff of the players in the hypothetical scenario where payoffs are deter-

[7] These assumptions guarantee a unique interior Pareto optimal total effort.

[8] Linear VCMs, where

$$f(e_1 + \ldots + e_N) = \frac{b}{N}(e_1 + \ldots + e_N)$$

and b is the marginal per-capita return, are by far the most commonly studied.

mined solely by the group grade and the cost of effort, i.e.,

$$W(e_1, \ldots, e_N) = f(e_1 + \ldots + e_N) - \frac{1}{N} \sum_{i=1}^{N} e_i.$$

Welfare is defined from the perspective of the manager/instructor who cares about the quality of the final product and the costly effort of the employees/students, but not the individual grades.

Welfare generally differs from the average of player payoffs since welfare is independent of evaluations. Nevertheless, welfare trivially coincides with the average payoffs if either there are no evaluations or there are evaluations but they don't "count," i.e., $\alpha = 0$. More important, welfare also coincides with the average of the players' payoffs in the symmetric equilibrium with truth-telling when $\alpha > 0$, as we will see shortly.[9]

EQUILIBRIUM

The peer evaluation game has many subgame perfect equilibria. To give one a simple example, for $\alpha < 1$ let $e^{FR}(\alpha)$ the symmetric Nash equilibrium effort of the game with payoffs for player i of

$$(1 - \alpha) f(e_1 + \ldots + e_N) - e_i,$$

and let $c \geq 0$ be an arbitrary constant. It is easy to verify it is a subgame perfect equilibrium of the peer evaluation game for each player to choose effort $e^{FR}(\alpha)$ and give every player an evaluation of c for every effort profile e.[10] In this equilibrium, effort is decreasing in the weight placed on peer evaluations.

We focus on equilibria in which evaluations are truthful. Proposition 1 characterizes such equilibria. When $\alpha > 0$, there is a unique equilibrium

[9] See Claim 1 in the Appendix for the proof.

[10] More formally, $e_1 = \ldots = e_N = e^{FR}(\alpha)$ and $\tilde{\varepsilon}_i^j(e_1, \ldots, e_N) = c$ for all i, j, and $(e_1, \ldots, e_N)$ is a subgame perfect equilibrium.

in which evaluations are truthful and equilibrium is symmetric. When $\alpha = 0$ then total equilibrium effort is determined, but equilibrium need not be symmetric.

Proposition 1: *(i) If $\alpha > 0$ then there is a unique subgame perfect equilibrium $(e^*, \tilde{\varepsilon}^*)$ in which evaluations are truthful, i.e., in which $\tilde{\varepsilon}_i^{j*}(e_1, \ldots, e_N) = e_j$ for each i, j, and $(e_1, \ldots, e_N)$. Equilibrium is symmetric. Each player's effort is e^*, which is positive and the unique solution to*

$$[1 - \alpha + \alpha N] f'(Ne^*) = 1.$$

(ii) If $\alpha = 0$, then any strategy profile $(e^*, \tilde{\varepsilon}^*)$ where $e^* = (e_1^*, \ldots, e_N^*)$ satisfies

$$f'(e_1^* + \ldots + e_N^*) = 1$$

is a subgame perfect equilibrium, regardless of whether evaluations are truthful. Total effort $e_1^* + \ldots + e_N^*$ is uniquely determined, but efforts need not be symmetric.

While we assume a player's individual grade is determined by their median evaluation, the results of Proposition 1 are robust to different rules for determining individual grades. We provide two such alternative rules: each player's individual grade is determined by (i) the average of the evaluations they receive from the $N-1$ other players, or by (ii) a single evaluation drawn at random from the $N-1$ evaluations received from the other players. The key feature of both examples is that a player's self evaluation has no effect on their own individual grade. The median rule has the same feature so long as the evaluations by the other players agree, as they do when evaluations by the other players are truthful.

Let e^{WM} denote the symmetric effort level that maximizes welfare, i.e.,

$$e^{WM} \in \arg\max_e W(e, \ldots, e) = \arg\max_e f(Ne) - e.$$

Clearly $e^{WM} > e^{FR}(\alpha)$.[11] Let $W^* = f(Ne^{WM}) - e^{WM}$.

Proposition 2 below shows that equilibrium effort is increasing in α when evaluations are truthful. Furthermore, when $\alpha = 1$ then equilibrium effort maximizes welfare, with each player choosing effort e^{WM}. This follows since the marginal social benefit of an increase in e_i is equal to player i's marginal private benefit.[12]

Proposition 2: *In the subgame perfect equilibrium in which evaluations are truthful, equilibrium effort is given by $e^*(\alpha)$ and is increasing in the weight α given to peer evaluations. Moreover, equilibrium effort is welfare maximizing when the final grade is determined entirely by peer evaluations, i.e., $e^*(1) = e^{WM}$.*

While the welfare maximizing effort results in the symmetric equilibrium with truth-telling when $\alpha = 1$, there is a strong incentive for players to tacitly collude by contributing low effort and providing high evaluations. In particular, in the subgame perfect equilibrium in which each player i chooses effort $e_i = 0$ and gives every other player the highest possible evaluation, the players obtain higher payoffs than in the truth-telling equilibrium.[13] In light of this observation, it is natural to question whether peer evaluations raise welfare in practice.

[11] This follows since $f'(Ne^{FR}) = 1/(1-\alpha) \geq 1 > 1/N = f'(Ne^{WM})$ and f' is decreasing.

[12] When $\alpha = 1$, the marginal private benefit to an increase in e_i is $Nf'(Ne_i) - 1$, while the marginal social benefit to an increase in e_i is $Nf'(e_1 + \ldots + e_N) - 1$. These are the same when $e_1 = \ldots = e_N$.

[13] Claim 2 establishes that this is a subgame perfect equilibrium. In this equilibrium players follow the collusive evaluation norm $\tilde{\varepsilon}_i^j(e_1, \ldots, e_N) = c$ for all $(e_1, \ldots, e_N)$, where c is a constant.

3 Experimental Design

The experimental game had 3 players per group. The group grade was given by

$$f(x) = Ax - Bx^2,$$

where $x = e_1 + e_2 + e_3$, and $A = 1.15$, and $B = .01$. At the first stage, each player i chose an effort $e_i \in [0, 15]$. At the second stage, each player i observed the first-stage effort of every player j in their group and then gave player j an evaluation $\tilde{e}_i^j \in [0, 15]$. We focus on how the weight α on the individual grade affects effort and welfare.

$$[1 - \alpha + \alpha N] f'(Ne^*) = 1.$$

By Proposition 1, in the unique subgame perfect equilibrium in which evaluations are truthful, equilibrium effort e^* is given by the unique solution to

$$[1 - \alpha + \alpha N](A - 2BNe^*) = 1.$$

Substituting for A, B, and N, yields

$$e^*(\alpha) = \frac{230\alpha + 15}{12\alpha + 6}.$$

By Proposition 2, the welfare maximizing effort level is $e^{WM} = e^*(1) = 13.61$.

By contrast, in any subgame perfect equilibrium in which evaluations do not depend on effort, equilibrium effort e^{FR} is the solution to

$$(1 - \alpha)(A - 2BNe^{FR}) = 1,$$

i.e.,

$$e^{FR}(\alpha) = \begin{cases} \frac{15-115\alpha}{6-6\alpha} & \text{if } \alpha < \frac{3}{23} \\ \\ 0 & \text{if } \alpha \geq \frac{3}{23}. \end{cases}$$

Figure 0 illustrates these theoretical benchmarks. As established in the prior section, equilibrium effort $e^*(\alpha)$ is increasing in the weight placed on peer evaluations in the equilibrium with truth-telling, while equilibrium effort is decreasing in the weight placed on evaluations in the free riding equilibrium.

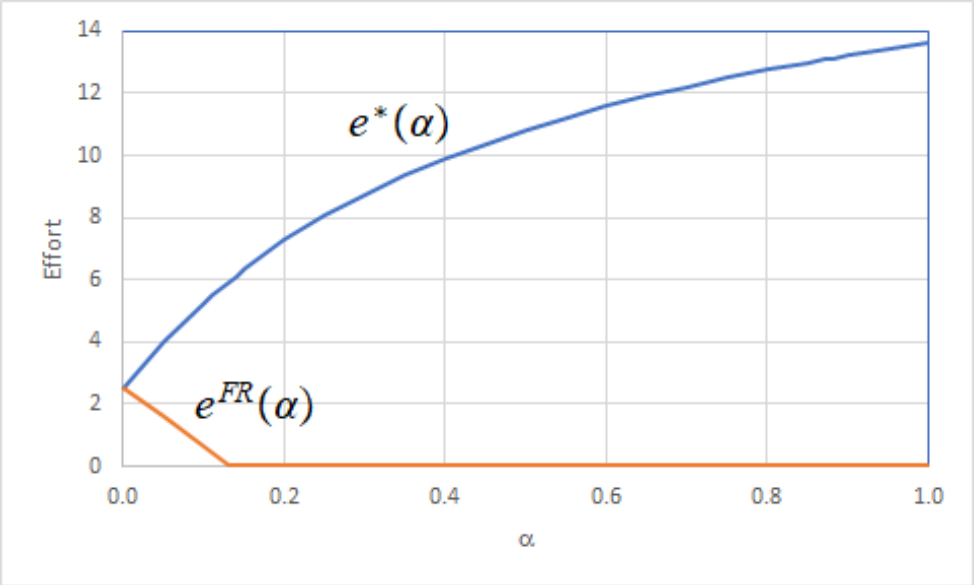


Figure 0: Effort as a Function of α.

The experiment had four treatments. The baseline No Evaluations ("No Eval.") treatment is the non-linear public goods game without peer evaluations and payoff function $\pi_i(e_1, e_2, e_3) = f(e_1 + e_2 + e_3) - e_i$. The "Peer Eval. $\alpha = 0$" treatment introduces peer evaluations, but has the same payoff function. This treatment accommodates the possibility that effort choices

respond to evaluations even when monetary payoffs are independent of evaluations. In the "Peer Eval. $\alpha = .5$" treatment, a subject's payoff is an equal-weighted average of their group grade and their individual grade minus their effort cost. Finally, in the "Peer Eval. $\alpha = 1$" treatment, a subject's payoff is their individual grade minus their effort cost.

Table 1 lists the four treatments and theoretical benchmarks for effort, payoffs, and welfare.

		Truth-telling $\underline{\tilde{\varepsilon}_i^j(e) = e_j}$		Free Riding $\underline{\tilde{\varepsilon}_i^j(e) = 15}$		
Treatments		$e^*(\alpha)$	$\pi^*(\alpha)$	$e^{FR}(\alpha)$	$\pi^{FR}(\alpha)$	W^{FR}
1.	No Eval.	2.50	5.56	2.50	5.56	5.56
2.	Peer Eval. $\alpha = 0$	2.50	5.56	2.50	5.56	5.56
3.	Peer Eval. $\alpha = .5$	10.83	15.98	0	15.75	0
4.	Peer Eval. $\alpha = 1$	$e^{WM} = 13.61$	$W^* = 16.67$	0	31.50	0

Table 1: Theoretical Benchmarks

The columns labeled $\pi^*(\alpha)$ and $\pi^{FR}(\alpha)$ show, respectively, payoffs in the equilibrium with truth-telling and payoffs in the free-riding equilibrium that is most favorable to the subjects (the one in which each player exerts effort $e^{FR}(\alpha)$ and receives the maximal evaluation of 15). They show that when $\alpha = .5$, then payoffs in the equilibrium with truth-telling are roughly the same as in the free-riding equilibrium most favorable to the subjects (payoffs of 15.98 and 15.75, respectively). By contrast, when $\alpha = 1$, the equilibrium with truth-telling is Pareto dominated by the free-riding equilibrium in which each player exerts no effort at all and receives the maximal evaluation. The $\alpha = 1$ treatment is therefore a "stress" test of the theory. These comparisons illustrate a practical trade-off when choosing the weight to place on peer

evaluations: a high weight on evaluations increases welfare and payoffs in the truth-telling equilibrium, but at the same time increases the subjects' incentives to tacitly collude.[14]

In the experiment there were 6 sessions, with 12 subjects each, for each treatment. A total of 288 subjects participated. Each session lasted 20 rounds. Each round, the subjects were randomly matched by zTree into groups of three to play the peer evaluation game. Subjects were informed of the number of rounds and the matching protocol. The subjects' decisions were framed in terms of a number of tokens to contribute, rather than in terms of effort. At the end of each round, a subject observed the total number of tokens contributed by their peers, their own contribution, the median of their evaluations, the group grade, their individual grade, their final grade, and their earnings. One round was chosen at random for payment.[15]

The sessions were conducted at the LINEEX Laboratory for Research in Behavioural Experimental Economics of the University of Valencia. The participants were undergraduate students and no subject participated in more than one session. Instructions were read aloud and, to avoid framing effects, a neutral framing was used for the instructions and questions. All questions were answered in private. No communication between subjects was allowed. The show up fee was 5 Euros. Subjects earned on average 13.90 Euros, in addition to the show up fee.

[14]Note that payoffs and welfare are the same in the equilibrium with truthtelling, for every value of α.

[15]Subjects were provided with a paper handout showing how the group grade was determined based the total number of token contributed. Subjects also had an on-screen "What If" earnings calculator. A subject could enter (i) a guess of the total number of tokens contributed by the other players in their group, (ii) their own contribution of tokens, and (iii) a guess of the median evaluation they would receive, and the calculator would provide a projection of their group grade, individual grade, and final grade. Subjects were free to make as many guesses as they wished.

4 Results: Effort

Evaluations as a Nudge

Figure 1 shows average effort for each treatment in all 20 rounds in the No Evaluations and the $\alpha = 0$ treatments.[16] It is evident that average effort in the baseline No Evaluations treatment exceeds the free-riding effort of 2.5. This result is consistent with a large prior literature on voluntary contribution mechanisms (without peer evaluations) which finds that contributions exceed the Nash equilibrium amounts. Introducing evaluations that don't count (i.e., $\alpha = 0$) raises average effort above the baseline effort.

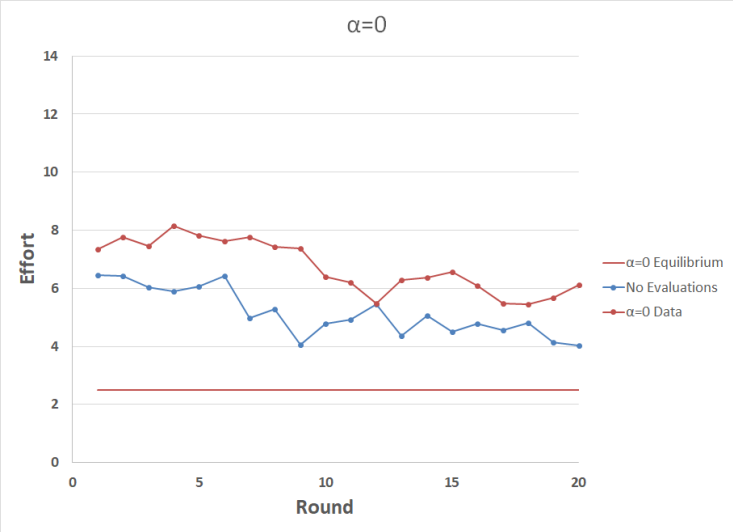


Figure 1: Average Effort, No Evaluations and $\alpha = 0$

Table 2 gives average player effort and welfare, over the last 10 rounds, for each session of each treatment. Average effort is statistically significantly higher in the $\alpha = 0$ treatment than in the No Evaluations treatment. In particular, the null hypothesis that average efforts in the two treatments are drawn from the same distribution is rejected at the 1% level ($n_1 = n_2 = 6$,

[16] Averaged across the 72 subjects in a treatment.

exact p-value 0.0022).[17] We can, likewise, reject the null hypothesis that welfare in the two treatments is drawn from the same distribution – welfare is higher when players give and receive evaluations, even though a player's payoff does not depend on the evaluations they receive (exact p-value 0.0022). Giving and receiving evaluations increased average welfare from 8.96 to 10.80, a 20.54% increase relative to the voluntary contributions game without evaluations.[18]

	Effort e_i				Welfare $W(e_1, e_2, e_3)$			
Session	No Eval.	$\alpha = 0$	$\alpha = .5$	$\alpha = 1$	No Eval.	$\alpha = 0$	$\alpha = .5$	$\alpha = 1$
1	4.84	6.71	9.53	3.62	9.03	11.96	14.79	7.14
2	4.77	5.00	10.44	5.79	9.19	9.51	15.00	10.57
3	4.72	5.07	9.91	9.96	9.23	9.48	15.35	15.28
4	4.74	6.05	9.66	10.16	9.16	10.88	15.02	15.46
5	4.14	7.26	7.97	12.49	8.12	12.49	13.52	16.45
6	4.78	5.72	10.09	4.85	9.06	10.48	15.27	9.21
Average	4.66	5.97	9.60	7.81	8.96	10.80	14.82	12.35

Table 2: Average Effort and Welfare, rounds 11-20

Result 1: *Peer evaluations are an effective nudge. Effort and welfare both increase when players give and receive evaluations, even when the evaluations received have no direct effect on payoffs.*

Result 1 shows that evaluations have a substantial impact on effort and welfare, even when a model of rational decision making predicts none. We now turn to assessing the effect of incentivized evaluations, taking effort and

[17] The two samples are maximally separated: the lowest average in the $\alpha = 0$ treatment exceeds the highest average in the No Evaluations treatment. The exact p-value is computed by https://ccb-compute2.cs.uni-saarland.de/wtest/

[18] Since welfare is a concave (non-linear) function of effort, average welfare is not determined solely by average effort, although they are highly correlated in the data.

welfare when $\alpha = 0$ as the baseline.

THE POWER OF TRUTHINESS

In the $\alpha = .5$ treatment, each player's payoff is an equally weighted average of the group grade and their individual grade, where the later is determined by the median evaluation the player receives. Figure 2 shows average effort in all 20 rounds in the $\alpha = 0$ and $\alpha = .5$ treatments. Recall from Table 1 that equilibrium effort is 2.50 when $\alpha = 0$ and 10.83 when $\alpha = .5$.

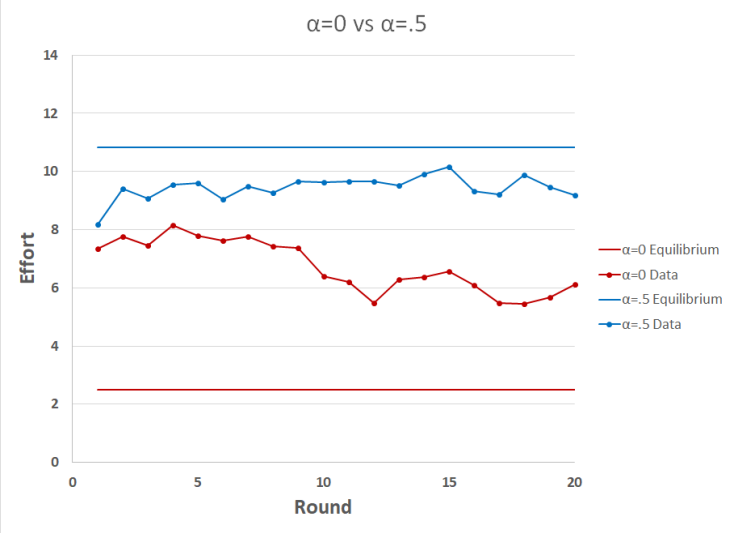


Figure 2: Average Effort, $\alpha = 0$ and $\alpha = .5$

Figure 2 shows that effort increases when payoffs are an equally weighted average of group and individual grades. The null hypothesis that the average efforts reported in Table 2 for $\alpha = 0$ and $\alpha = .5$ are drawn from the same distribution is rejected at the 1% level (exact p-value 0.0022). Averaging across the 6 sessions, increasing α from 0 to .5 increases average effort from 5.97 to 9.60, a 60.8% increase. Average effort of 9.60 approaches its theoretically predicted value of 10.83.

The null hypothesis that welfare in the $\alpha = 0$ and the $\alpha = .5$ treat-

ments are drawn from the same distribution is likewise rejected (exact p-value 0.0022). Increasing α from 0 to .5 increases average welfare from 10.80 to 14.82, a 37.2% increase, and welfare is 65.4% higher with $\alpha = .5$ than when there are no evaluations at all. Average welfare of 14.82 approaches its theoretically predicted value, in the equilibrium with truth-telling, of 15.98.

Result 2: *Effort and welfare both increase when grades are an equally weighted average of the group grade and evaluation-based individual grades, rather than being determined solely by the group grade. Effort and welfare approach the levels predicted in the equilibrium with truth-telling.*

BENDING THE TRUTH UNTIL IT BREAKS

When grades are determined solely by evaluations, i.e., $\alpha = 1$, then, in the equilibrium with truth-telling, effort and welfare reach their Pareto efficient levels. In particular, increasing α from .5 to 1 increases equilibrium effort from 10.83 to 13.61, an increase of 25.7%.

Figure 3 shows average effort in all 20 rounds in the $\alpha = .5$ and $\alpha = 1$ treatments. It shows that effort, in fact, *falls* as α increases to 1. Table 2 shows that average effort (in rounds 11-20) falls from 9.60 when $\alpha = .5$ to 7.81 when $\alpha = 1$, an 18.6% decrease. In the next section we shall see that peer evaluations lose effectiveness when $\alpha = 1$ as a result of a breakdown in

truth-telling.

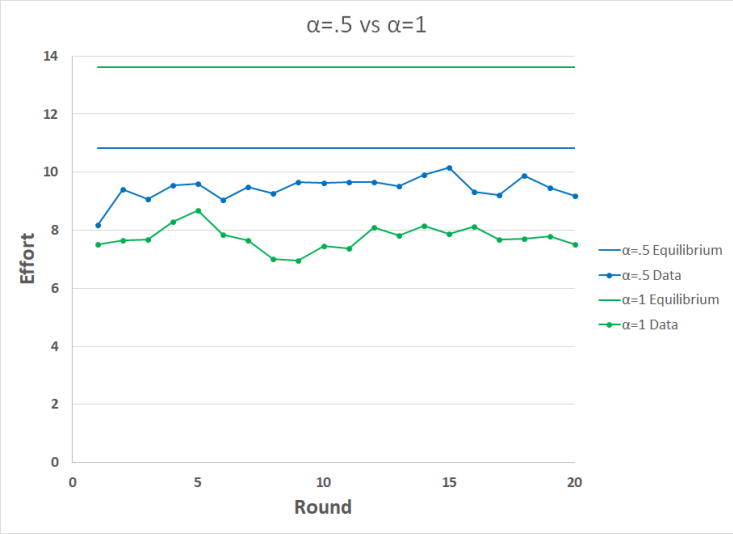


Figure 3: Average Effort, $\alpha = .5$ and $\alpha = 1$

Table 3 presents regression results for effort, welfare, and individual and group grades by treatment. (All regression models use cluster-robust standard errors at the session level.) When there are no evaluations (the baseline), mean effort is 5.152, which is above the theoretical level of 2.50, a difference that is statistically significant. Introducing evaluations that don't count raises effort by 1.586, an increase that is statistically significant at the 1% significance level, confirming that evaluations are an effective nudge. Subjects provide more effort in anticipation of receiving evaluations. In the $\alpha = .5$ treatment, effort rises by a further $2.709 = 4.295 - 1.586$ units, a difference which is statistically significant at the 1% level.[19] Average effort of $9.447 = 5.152 + 4.295$ approaches its theoretical level of 10.83. By contrast, further increasing the weight placed on evaluations, from $\alpha = .5$ to $\alpha = 1$, leads to a reduction in effort.[20] In all cases, effort is higher with evaluations

[19] The corresponding t-statistic is -4.81 ($p < 0.001$).

[20] The reduction in effort of $2.593 - 4.295 = -1.702$ is not statistically significant, with a t-statistic is 1.48 ($p = .153$).

than without.

			Grade	
Treatment	Effort	Welfare	Group	Indiv.
	e_i	$W(e_1, e_2, e_3)$	$f(e_1 + e_2 + e_3)$	$f(3m(\tilde{e}_i))$
No Eval.	5.152***	9.659***	14.811***	17.920***
(constant)	(0.184)	(0.249)	(0.432)	(0.755)
$\alpha = 0$	1.586***	2.105***	3.691***	
	(0.433)	(0.540)	(0.971)	
$\alpha = 0.5$	4.295***	4.977***	9.273***	7.514***
	(0.445)	(0.436)	(0.872)	(1.460)
$\alpha = 1$	2.593**	2.837**	5.430**	6.436***
	(1.095)	(1.182)	(2.267)	(1.613)
N	5,760	1,920	1,920	4,320
R-squared	0.118	0.217	0.231	0.121

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

Table 3: Effort, welfare, and grades (all rounds)[21]

Figures 2 suggests the presence of a time trend in effort when $\alpha = 0$. Denote by e_{it} the effort of subject i at round t. To test for time trends, we estimate the regression model

$$e_{it} = \sum_{j \in \{No\ Eval, \alpha=0, \alpha=.5, \alpha=1\}} \left(\frac{1}{t} \mathbb{I}_i^j \beta_1^j + \frac{t-1}{t} \mathbb{I}_i^j \beta_\infty^j \right) + \epsilon_{it},$$

where $\mathbb{I}_i^j = 1$ if subject i is assigned to treatment $j \in \{No\ Eval., \alpha = 0, \alpha = .5, \alpha = 1\}$ and $\mathbb{I}_i^j = 0$ otherwise, β_1^j is initial effort at round 1 in treatment j, and β_∞^j is limit effort in treatment j in the hypothetical scenario in which there are infinite number of rounds of play. In other words, in treatment j,

[21] The baseline treatment is the omitted term.

effort at round t is a weighted average of β_1^j and β_∞^j with weights $1/t$ and $(t-1)/t$, respectively.

Regression results are reported in Table 4 below.

Treatment	Initial and Limit Efforts		
	β_1^j	β_∞^j	$\beta_\infty^j - \beta_1^j$
No Eval. (Constant)	7.195	4.704	-2.492***
	(0.550)	(0.122)	(0.471)
$\alpha = 0$	8.322	6.390	-1.932***
	(0.461)	(0.414)	(0.415)
$\alpha = 0.5$	8.341	9.690	1.349***
	(0.478)	(0.409)	(0.302)
$\alpha = 1$	7.624	7.771	0.147
	(0.621)	(1.281)	(1.301)
N		5,760	
R-squared		0.754	

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

Table 4: Estimates of β_1^j and β_∞^j

In the No Eval and $\alpha = 0$ treatments, the differences $\beta_\infty^{NoEval} - \beta_1^{NoEval} = -2.492$ and $\beta_\infty^{\alpha=0} - \beta_1^{\alpha=0} = -1.923$ are both statistically significant, demonstrating that effort decays in both treatments. The first of these differences is consistent with prior experimental evidence showing that contributions decay in voluntary contribution public good games. In contrast, when $\alpha = .5$ there is a statistically significant increasing trend in effort ($\beta_\infty^{\alpha=.5} - \beta_1^{\alpha=.5} = 1.349$), which shows that the effort-enhancing effect of peer evaluations are robust to experience.

Figure 4 shows average payoffs for each treatment in all 20 rounds. Consistent with the decline over time in effort in the No. Eval and $\alpha = 0$ treatments, payoffs also decline over time. In $\alpha = .5$ and $\alpha = 1$ treatments,

by contrast, payoffs increase over time. In the $\alpha = 1$ treatment, payoffs ultimately exceed the welfare maximizing payoffs of 16.67. This is possible only if evaluations are inflated relative to average effort, thereby raising payoffs.

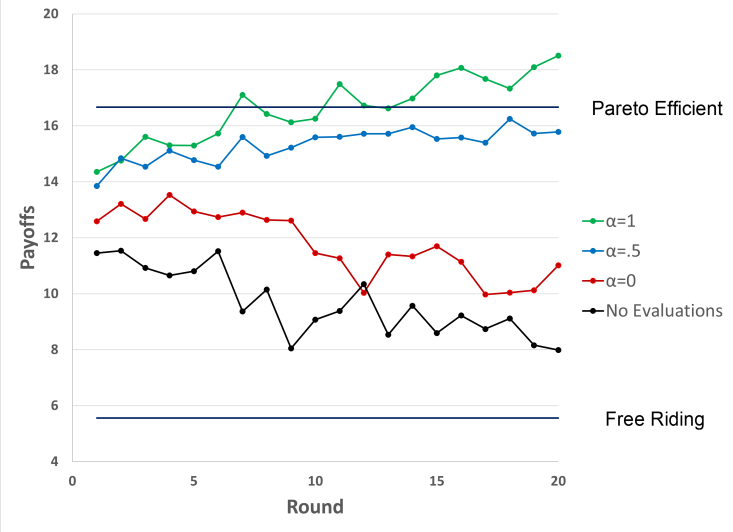


Figure 4: Average Payoffs

The effect on payoffs of inflated evaluations is clear in Figure 5, which compares average payoffs and welfare for $\alpha = .5$ and $\alpha = 1$. When $\alpha = .5$, payoffs slightly exceed welfare indicating that evaluations are only slightly elevated relative to effort. When $\alpha = 1$, payoffs substantially exceed welfare, demonstrating the breakdown of truth-telling and the inflation of evaluations

relative to actual effort.

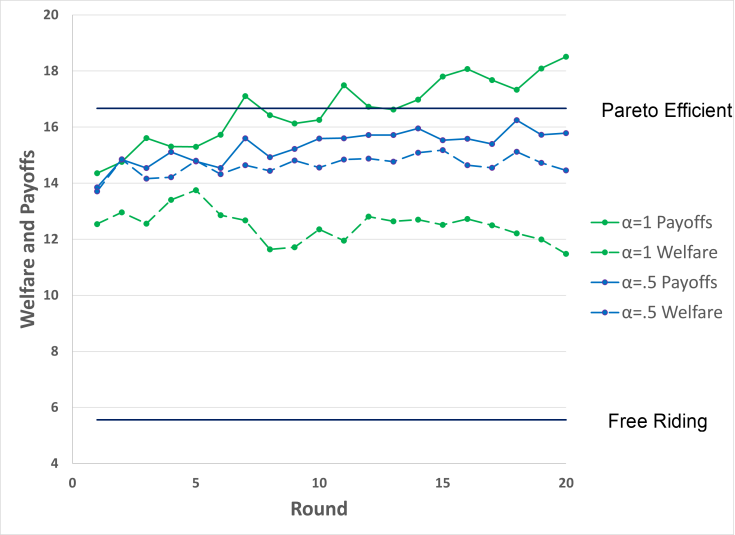


Figure 5: Payoffs

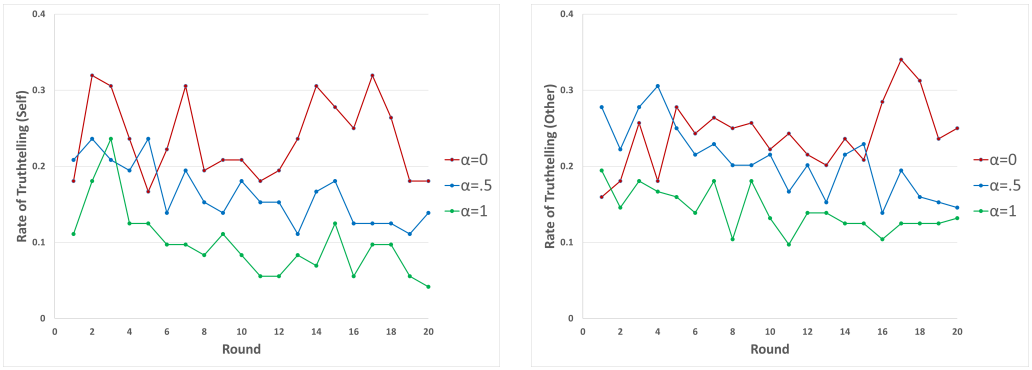
5 Results: Evaluation Norms

Next we study the evaluation norms followed by subjects, and how the responsiveness of evaluations to effort shapes incentives to provide effort. In the peer evaluation game a player has no monetary incentive to provide truthful evaluations: the payoff of player i does not depend on the i's evaluation of player j. Hence we would expect truthful evaluations only so long as truth-telling is a social norm.[22]

Recall that i's evaluation of j is *truthful* if $\tilde{e}_i^j = e_j$. Figure 2AB shows the fraction of truthful evaluations at each round for each treatment. It is evident that evaluations are largely not truthful. Qualitatively, evaluations are most truthful when $\alpha = 0$ and least truthful when $\alpha = 1$. Self evaluations

[22]If a player has other regarding preferences, he might inflate or deflate his evalations for reasons of beneficence or spite.

tend to be less truthful than evaluations of others.



(a) Self (b) Others

Figure 2AB: Fraction of Truthful Evaluations

Table 5 presents regression results. The first column shows regression results when the dependent variable is an indicator function that takes the value 1 if i's evaluation of j is truthful, and takes the value zero otherwise. It shows that 24.1% of evaluations are truthful when evaluations don't count. Evaluations are less truthful by approximately 3.3 percentage points when $\alpha = .5$, and by 10.0 percentage points when $\alpha = 1$, with the later difference being statistically significant. Self evaluations are even less truthful, although

the coefficient estimates are not statistically significant.

	Truthful $I_{\tilde{e}_i^j = e_j}$	Difference $\tilde{e}_i^j - e_j$
$\alpha = 0$ (Constant)	0.241***	-0.099
	(0.028)	(0.274)
$\alpha = 0.5$	-0.033	0.401
	(0.053)	(0.519)
$\alpha = 1$	-0.100***	1.067
	(0.032)	(0.758)
Self $(i = j)$	-0.004	2.216***
	(0.024)	(0.296)
Self $(i = j) \times \alpha = 0.5$	-0.040	0.441
	(0.029)	(0.554)
Self $(i = j) \times \alpha = 1$	-0.037	1.304**
	(0.033)	(0.500)
N	12,960	12,960
R-squared	0.016	0.085

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

Table 5: Truthfulness of Evaluations

The last column of Table 5 shows that the difference between i's evaluation of j and j's effort is not statistically significantly different from zero in any treatment. Hence, while evaluations of others are mostly not truthful, they are unbiased. By contrast, self-evaluations are consistently inflated, with evaluations exceeding own effort by $2.117 = -0.099 + 2.216$ on average when evaluations don't count, by $2.959 = -0.099 + .401 + 2.216 + 0.441$ when $\alpha = .5$, and by $4.488 = -0.099 + 1.067 + 2.216 + 1.304$ when $\alpha = 1$.

Figure 4AB shows scatter plots of effort-evaluation pairs. The panels on

the left shows own effort and self evaluations. The panels on the right show others' efforts and evaluations of others. Truthful evaluations lie on the 45 degree line, which is shown in green. The red lines in the left panels show linear regression lines for the model

$$\tilde{\varepsilon}_{it}^{i} = c + \beta e_{it} + \epsilon_{it}$$

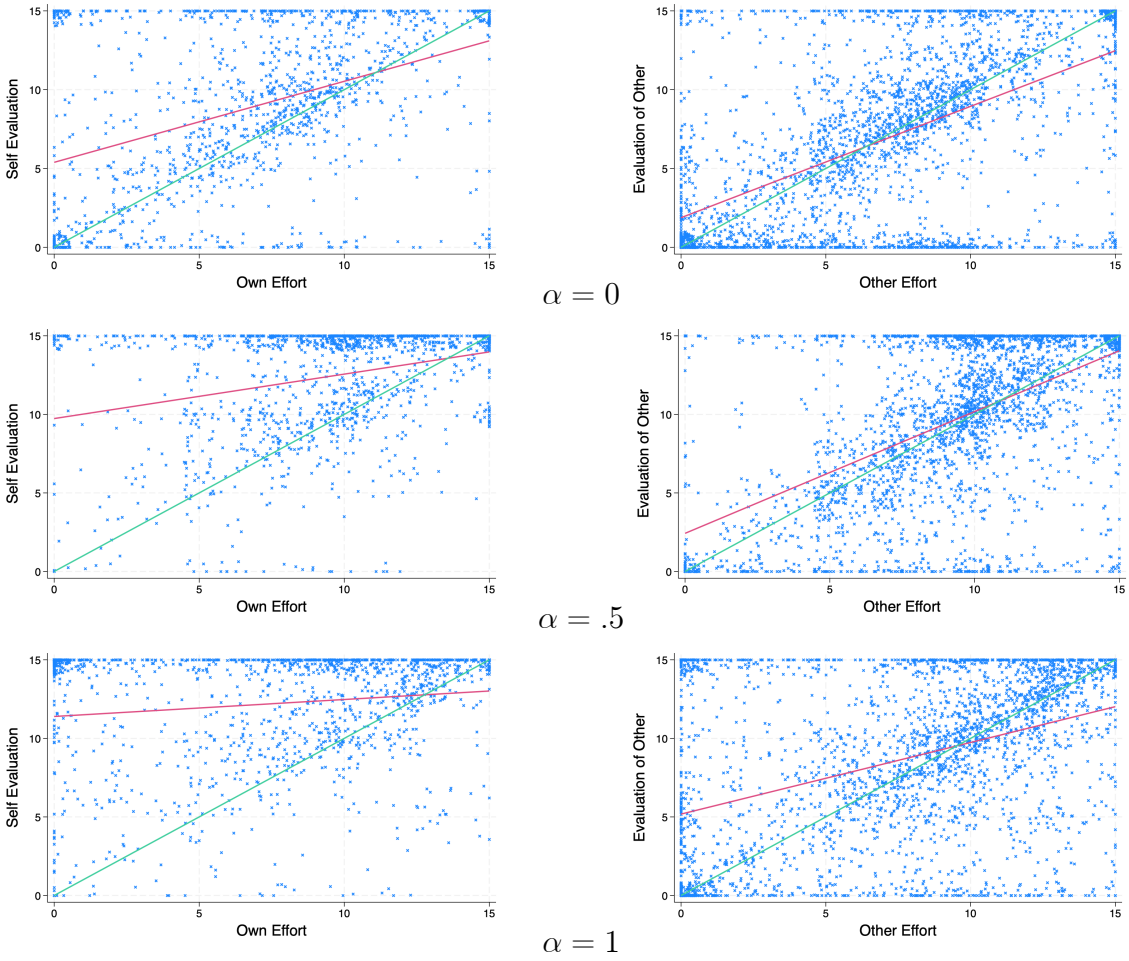
of players i's self evaluation given their own effort, while the red lines on the right panels side show linear regression lines for the model

$$\tilde{\varepsilon}_{it}^{j} = c + \beta e_{jt} + \epsilon_{it} \text{ for } i \neq j,$$

of players i's evaluation of player j's effort. The regression estimates for each treatment are shown in the top two lines of Table 6 below.

Considering the each rows of Table 6, it is evident that the red lines are closer to the 45 degree lines for evaluations of others (right panels) than for self evaluations (left panels), i.e., evaluations of others are more truthful than self evaluations. Considering the columns of the figure, it is evident that the red lines are farther from the 45 degree lines as α is larger, i.e., evaluations (both self and of others) are most truthful when $\alpha = 0$, they are less truthful

when $\alpha = .5$ (middle row), and are least truthful when $\alpha = 1$ (bottom row).



(a) Own Effort (b) Other Effort

Figure 4AB: Scatter Plots of Effort and Evaluations
Top $\alpha = 0$, Middle $\alpha = .5$, and Bottom $\alpha = 1$

The regression results in Table 7(a) show the responsiveness of evaluations to effort. The top left entry of the table, for example, shows that if player i increases their effort by one unit, then other players increase their evaluation

of i by .706 units and player i increases their self evaluation by .514 units. The table reveals several general features of evaluations: (i) Evaluations of others are more responsive to others' efforts than self evaluations are to own effort, for all treatments, whether one considers the first 10 periods, the last 10 periods, or all 20 periods. (ii) Comparing the third column to the first two, we see that evaluations (both self and of others) are least response to effort when $\alpha = 1$ (e.g., in the first row we have both $.456 < .706$ and $.456 < .775$). (iii) Finally, comparing the first 10 to the last 10 periods, the responsiveness of evaluations to effort is little changed for $\alpha = 0$ and $\alpha = .5$, but is lower in the last 10 rounds for $\alpha = 1$. Notably, in the $\alpha = .5$ treatment, evaluations of others are highly responsive to effort and they remain so through all 20 rounds.

	Treatment		
	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
Panel A: All 20 periods			
Others	$1.879 + 0.706e$	$2.427 + 0.775e$	$5.180 + 0.456e$
Self	$5.391 + 0.514e$	$9.735 + 0.283e$	$11.394 + 0.108e$
Panel B: First 10 periods			
Others	$1.874 + 0.738e$	$2.653 + 0.736e$	$4.319 + 0.530e$
Self	$5.719 + 0.502e$	$8.650 + 0.354e$	$10.395 + 0.157e$
Panel C: Last 10 periods			
Others	$1.929 + 0.659e$	$2.124 + 0.821e$	$5.984 + 0.391e$
Self	$5.200 + 0.506e$	$11.131 + 0.181e$	$12.376 + 0.063e$

Table 7(a): Regressions of Evaluations on Effort

These results show that evaluations are responsive to effort, even though not fully truthful.

6 Truthiness

We have seen than evaluations respond to effort, but are less than fully truthful. In this section we characterize the subgame perfect equilibrium of the Peer Evaluation game for general linear evaluations functions. In this fashion, we investigate the extent to which behavior is consistent with the play of a subgame perfect equilibrium in which evaluations are responsive to effort, but not fully truthful.

Suppose that each player i evaluates j according to the evaluation function

$$\tilde{\varepsilon}_i^j(e_1, \ldots e_{j-1}, e_j, e_{j+1}, \ldots, e_N) = \kappa + \beta e_j,$$

where κ is a constant and β is the responsiveness of evaluations to effort. Given effort e_i, player i's median evaluation is $\kappa + \beta e_i$, and thus i's payoff function is

$$\pi_i(e_1, \ldots, e_N, \tilde{e}_1^i, \ldots, \tilde{e}_N^i) = (1-\alpha) f(e_1 + \ldots + e_N) + \alpha f(N(\kappa + \beta e_i)) - e_i.$$

Proposition 3 characterizes equilibrium.

Proposition 3: *(i) Assume that* $(1-\alpha)f'(0) + \alpha N \beta f'(N\kappa) - 1 > 0$. *If* $\alpha > 0$, *then there is a unique subgame perfect equilibrium* $(e^*, \tilde{\varepsilon}^*)$ *in which evaluations are given, for* $\beta \geq 0$, *by*

$$\tilde{\varepsilon}_i^j(e_1, \ldots e_{j-1}, e_j, e_{j+1}, \ldots, e_N) = \kappa + \beta e_j.$$

i.e., in which $\tilde{\varepsilon}_i^{j*}(e_1, \ldots, e_N) = e_j$ *for each* i, j, *and* $(e_1, \ldots, e_N)$. *Equilibrium is symmetric. Each player's effort is* e^*, *which is positive and the unique solution to*

$$(1-\alpha) f'(Ne^*) + \alpha N \beta f'(N(\kappa + \beta e^*)) = 1.$$

In the experiment, $f(x) = Ax - Bx^2$, where $A = 1.15$ and $B = .01$.

Equilibrium effort, denoted by $e^*(\kappa, \beta)$, is[23]

$$e^*(\kappa, \beta) = \frac{15 - 115\alpha + \alpha\beta(345 - 18\kappa)}{18\alpha\beta^2 + 6 - 6\alpha}.$$

Table 7(b) identifies the subgame perfect equilibrium efforts for the evaluation functions reported in Table 7(a). Table 7(b) shows, when $\alpha = .5$ and $\alpha = 1$, that equilibrium efforts for the estimated evaluation functions (8.834 and 3.955, respectively) are lower than the equilibrium efforts (10.833 and 13.611, respectively) obtained under truth-telling. Hence the incentive for subjects to provide effort is weakened by their failure to provide truthful evaluations. Nonetheless, the incentives to provide effort remain substantial when $\alpha = .5$. In this case, the subgame perfect equilibrium efforts for the estimated evaluation functions are remarkably close to the observed mean efforts: in the first 10 rounds, 8.493 in equilibrium versus 9.292 observed; in the last 10 rounds 9.202 in equilibrium versus 9.602 observed. These observed efforts are substantially higher than those observed when evaluations

[23] Assuming $15 - 115\alpha + \alpha\beta(345 - 18\kappa)$ is positive. Equilibrium effort is 0 otherwise.

don't count, *viz.*, 7.507 and 5.968.

	Treatment		
	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
Panel A: All 20 periods			
Evaluation Function	$1.879 + 0.706e$	$2.427 + 0.775e$	$5.180 + 0.456e$
SPE e^*	2.500	8.834	3.955
Mean e	6.738	9.447	7.745
Panel B: First 10 periods			
Evaluation Function	$1.874 + 0.738e$	$2.653 + 0.736e$	$4.319 + 0.530e$
SPE e^*	2.500	8.493	8.237
Mean e	7.507	9.292	7.676
Panel C: Last 10 periods			
Evaluation Function	$1.929 + 0.659e$	$2.124 + 0.821e$	$5.984 + 0.391e$
SPE e^*	2.500	9.202	0.000
Mean e	5.968	9.602	7.812

Table 7(b): SPE Efforts for Non-truthful Evaluation Functions

The nearly-complete breakdown of truth-telling when $\alpha = 1$ further reduces incentives to provide effort. In the case, the subgame perfect equilibrium efforts for the estimated evaluation functions is 8.237 in the first 10 rounds but drops to zero in the last 10 rounds. Subjects contribute only slightly higher effort than they contributed when $\alpha = 0$.

Observed effort substantially exceeds the subgame perfect equilibrium levels when there are no evaluations, when $\alpha = 0$, and when $\alpha = 1$. In these treatments, it appears that subjects receive a utility boost to contributing; observed effort can not rationalized by considerations of monetary payoffs.

TRUTH-TELLING AND WELFARE

The Peer Evaluation game has many equilibria, which differ in the respon-

siveness of evaluations to effort. While truth-telling is a compelling social norm, different behavioral rules and norms may develop in different sessions. We see, indeed, that there is considerable heterogeneity in the truthfulness of evaluations across sessions. In this subsection we show that effort and welfare are higher in sessions in which evaluations are more truthful. In other words, truth-telling is a welfare-enhancing norm.

We measure the truthfulness of a group by the difference between the average individual grade in the group, i.e.,

$$I = \frac{f(3m_1(\tilde{e}^1)) + f(3m_2(\tilde{e}^2)) + f(3m_3(\tilde{e}^3))}{3},$$

and the group grade $G = f(e_1 + e_2 + e_3)$. The difference $I - G$ is exactly zero when evaluations are truthful and effort choices are symmetric. The difference tends to be positive if evaluations are inflated relative to effort, and it is negative if evaluations are below actual efforts.[24]

Table 8 reports session averages of effort, individual grades, and group grades, when $\alpha = .5$ and $\alpha = 1$. The results are most striking for $\alpha = 1$. Evaluations were most truthful in Session 5 (i.e., $I - G$ is closest to zero), and this session has the highest average effort (of 12.49). Evaluations were least truthful in Session 1, and this session had the lowest average effort (of 3.62). In this treatment greater truthfulness is perfectly correlated with

[24]When evaluations are truthful but effort choices are asymmetric, the difference is negative: since f is concave, then $(f(3e_1) + f(3e_2) + f(3e_3))/3 < f(e_1 + e_2 + e_3)$ for any any e_1, e_2, and e_3. Hence, when evaluations are truthful, the difference $I - G$ is either zero or negative.

higher effort.

	Effort e_i		Individual (I) $f(3m_i(\tilde{e}^i))$		Group (G) $f(e_1+e_2+e_3)$		$I-G$	
Session	$\alpha=.5$	$\alpha=1$	$\alpha=.5$	$\alpha=1$	$\alpha=.5$	$\alpha=1$	$\alpha=.5$	$\alpha=1$
1	9.53	3.62	26.439	24.447	24.323	10.759	2.116	13.688
2	10.44	5.79	24.930	22.347	25.440	16.357	−0.510	5.990
3	9.91	9.96	27.883	26.260	25.254	25.244	2.629	1.016
4	9.66	10.16	26.746	26.497	24.688	25.620	1.492	0.877
5	7.97	12.49	25.433	28.955	21.486	28.934	3.947	0.021
6	10.09	4.85	25.947	23.538	25.357	14.061	0.590	9.477
Average	9.60	7.81	26.230	25.341	24.425	20.162	1. 805	5.179

Table 8: $\alpha > 0$, Average Effort and Grades, rounds 11-20

We test for each value of α whether greater truthfulness in a session is associated with higher effort in the session. Let E^s denote average effort in session s and let T^s denote the difference $I-G$ in session s. When $\alpha = .5$, we have $(E^1, T^1) = (9.53, 2.116)$, $(E^2, T^2) = (10.44, -0.510)$, and so on to $(E^6, T^6) = (10.09, 0.590)$. We replace each entry with its rank R, thereby obtaining the rank pairs $(R(E^1), R(T^1)) = (2, 4)$, $(R(E^2), R(T^2)) = (6, 1)$, and so on. The Pearson correlation coefficient r of the six rank pairs $\{(R(E^s), R(T^s))\}_{s=1}^6$ is given by[25]

$$r = 1 - \frac{\sum_{s=1}^6 (R(E^s) - R(T^s))^2}{6^2 - 1}.$$

Under the null hypothesis that E and T are independent, then each rank pair (i, j), for $i, j \in \{1, \ldots, 6\}$, is equally likely. Let F denote the *c.d.f.* of the Pearson correlation coefficient of the rank pairs under this null hypothesis.[26]

[25] See p. 529 of Mood, Graybill, and Boes (1974) for this formula.

[26] It is straight forward to compute F numerically.

At the 5% significance level we reject the null hypothesis of independence if $F(r) < .025$ or $F(r) > .975$. When $\alpha = .5$, the correlation coefficient of the ranks is $r = -0.829$ (with $F(-0.829) = \frac{21}{720} = .029167$). When $\alpha = 1$, the correlation coefficient of the ranks is $r = -1$ (with $F(-1) = \frac{1}{720} = .001389$). Thus, in both cases we reject the null of independence in favor of the alternative that greater truthfulness (i.e., lower values of T) is associated with higher effort. When evaluations matter, effort and welfare are higher in sessions in which evaluations are more truthful.

By contrast, when $\alpha = 0$, then the correlation coefficient of the ranks is $r = .600$ (with $F(.600) = .913$). In this case, E and T are positively correlated, although not statistically significantly so. When evaluations don't matter, the truthfulness of evaluations has no impact on effort.

7 Conclusion

This paper shows how the evaluation norms people follow – truth-telling, meritocratic, or collusive – affects equilibrium effort and welfare, when peer evaluations are used to provide incentives to contribute to the public good. It demonstrates that a well-designed peer evaluation system is a practical and effective means of raising effort and welfare. It also demonstrates the value of truth-telling as a social norm, showing that groups that follow more meritocratic norms, and are closer to being truthful, achieve higher levels of welfare.

The present paper studies three-person groups. We conjecture that the median peer evaluation will be less noisy and provide clearer incentives in larger groups. Consequently, in larger groups, evaluation may be even more effective in incentivizing effort and enhancing welfare.

Our analysis also applies to settings where effort is observable but not verifiable, i.e., it is not possible to write enforceable contracts that are contingent on effort.

8 Appendix

All proofs are provided in this Appendix.

Claim 1: When $\alpha > 0$, then welfare coincides with the average of the players' payoffs if both (i) all the players choose the same effort and (ii) evaluations are truthful.

Proof: Let $(e, \tilde{\varepsilon}) = ((e_i, \tilde{\varepsilon}_i)_{i=1}^N)$ be a strategy profile such that $e_1 = \ldots = e_N$ and suppose evaluations are truthful. Since evaluations are truthful, then $m(\tilde{e}^i) = e_i$ and the payoff of player i is

$$\pi_i = \alpha f(e_1 + \ldots + e_N) + (1-\alpha) f(Ne_i) - e_i = f(e_1 + \ldots + e_N) - e_i.$$

Thus the average of the players payoffs is

$$\frac{1}{N}\sum_{i=1}^{N} \pi_i = f(e_1 + \ldots + e_N) - \frac{1}{N}\sum_{i=1}^{N} e_i = W(e_1, \ldots, e_N).$$

□

Claim 2: It is a subgame perfect equilibrium for each player to free-ride on the efforts of others, choosing effort $e^{FR}(\alpha)$ and giving every player an evaluation of c in every subgame e.

Proof: Suppose $(1-\alpha) f'(0) \geq 1$. Since $N f'(N\bar{e}) < 1$ by assumption, then $f'(N\bar{e}) < 1$ and thus $(1-\alpha) f'(N\bar{e}) < 1$. And since f' is strictly decreasing, then there is a unique e^{FR} such that $(1-\alpha) f'(Ne^{FR}) = 1$. If $(1-\alpha) f'(0) < 1$, then let $e^{FR} = 0$.

We now show that $e_i = e^{FR}$ and $\tilde{\varepsilon}_i^j(e_1, \ldots, e_N) = c$ for all i, j, and $(e_1, \ldots, e_N) \in \mathbb{R}^{N+}$ is a subgame perfect equilibrium. Clearly player i's median evaluation is c, regardless of their effort e_i. Player i's optimal effort is the solution to

$$\max_{e_i} (1-\alpha) f(e_i + (N-1)e^{FR}) + \alpha f(Nc) - e_i.$$

The first order necessary condition is that $(1-\alpha)f'(e_i+(N-1)e^{FR})=1$, which holds for $e_i=e^{FR}$ by the definition of e^{FR}. The first order condition is sufficient since f is concave. If $e^{FR}=0$, then $(1-\alpha)f'(e_i)-1<0$, and hence $e^{FR}=0$ is optimal. $\square$

Proof of Proposition 1: Assume $\alpha>0$. Let $(e^*,\tilde{\varepsilon}^*)$ be a subgame perfect equilibrium in which evaluations are truthful. Since evaluations are truthful, given effort e_i and self evaluation $\tilde{e}_i^i$, then Player i's median evaluation is

$$m\{\tilde{\varepsilon}_1^i(e),\ldots,\tilde{\varepsilon}_{i-1}^i(e),\tilde{e}_i^i,\tilde{\varepsilon}_{i+1}^i(e),\ldots,\tilde{\varepsilon}_N^i(e)\}=e_i$$

and is therefore independent of their self evaluation. Player i's payoff is

$$(1-\alpha)f(e_i+\sum_{j\neq i}e_j^*)+\alpha f(Ne_i)-e_i.$$

In equilibrium, e_i^* satisfies the Kuhn-Tucker condition

$$(1-\alpha)f'(e_i^*+\sum_{j\neq i}e_j^*)+\alpha Nf'(Ne_i^*)-1\leq 0, \tag{1}$$

where the inequality holds with equality if $e_i^*>0$. $1-\alpha+\alpha N=1+\alpha(N-1)$ is true

We first establish that the equilibrium is symmetric, i.e., that $e_i^*=e_k^*$ for all $i\neq k$. Suppose to the contrary that there are i and k such that $e_i^*>e_k^*$. Since f' is decreasing, then $f'(Ne_i^*)<f'(Ne_k^*)$. Hence $\alpha>0$ implies

$$(1-\alpha)f'(e_i^*+\sum_{j\neq i}e_j^*)+\alpha Nf'(Ne_i^*)-1<(1-\alpha)f'(e_k^*+\sum_{j\neq k}e_j^*)+\alpha Nf'(Ne_k^*)-1\leq 0,$$

where the weak inequality holds by (1). Hence $e_i^*=0$, which contradicts $e_i^*>e_k^*$. Thus in equilibrium $e^*=e_1^*=\ldots=e_N^*$.[27]

[27] When there is no ambiguity, we abuse notation by writing e or e^* for a player's effort and equilibrium effort, respectively.

Zero total effort, i.e., $e^* = 0$, is not an equilibrium since $f'(0) > 1$ implies $(1-\alpha+\alpha N)f'(0) - 1 > 0$, which contradicts (1).[28] Thus in any equilibrium we have $e^* > 0$ and

$$(1-\alpha)f'(Ne^*) + \alpha N f'(Ne^*) - 1 = 0. \tag{2}$$

Next we establish that equilibrium is unique. As already noted, $(1-\alpha)f'(0) + \alpha N f'(0) - 1 > 0$. Furthermore, we have

$$(1-\alpha)f'(N\bar{e}) + \alpha N f'(N\bar{e}) - 1 < 0,$$

where the inequality holds since $f'(N\bar{e}) < Nf'(N\bar{e}) < 1$. Since

$$(1-\alpha)f'(Ne) + \alpha N f'(Ne) - 1$$

is continuous, by the Intermediate Value Theorem there is a solution to (2). The solution e^* is unique since $(1-\alpha)f'(Ne) + \alpha N f'(Ne) - 1$ is strictly decreasing in e as $f'' < 0$.

Finally, we show that $(e^*, \tilde{\varepsilon}^*)$ as given in Proposition 1 is an equilibrium. Since

$$(1-\alpha)f(e_i + \sum_{j \neq i} e_j^*) + \alpha f(Ne_i) - e_i$$

is concave in e_i, the first order condition is sufficient and Player i's effort is optimal given e_{-i}^*. Further, in any subgame $(e_1, \ldots, e_N)$, since the other players are evaluating i truthfully, then player i's median evaluation is independent of their evaluation strategy $\tilde{\varepsilon}_i$, and thus it is optimal for i's self evaluation to be truthful, i.e., $\tilde{\varepsilon}_i^*(e) = e_i$. Likewise, i's payoff does not depend on their evaluation of j and thus it is optimal for i to evaluate j truthfully.

Proof of (ii). Assume $\alpha = 0$. Let $(e^*, \tilde{\varepsilon}^*)$ be a subgame perfect equilibrium.

[28] It would be weaker to assume directly that $(1-\alpha+\alpha N)f'(0) - 1 > 0$.

We show that total effort $e_1^* + \ldots + e_N^*$ satisfies

$$f'(e_1^* + \ldots + e_N^*) = 1.$$

Suppose that

$$f'(e_1^* + \ldots + e_N^*) < 1.$$

Player i's payoff given effort e_i is

$$f(e_i + \sum_{j \neq i} e_j^*) - e_i.$$

In equilibrium, each player i's equilibrium effort e_i^* satisfies the Kuhn-Tucker condition

$$f'(e_i^* + \sum_{j \neq i} e_j^*) - 1 \leq 0,$$

where the inequality holds with equality if $e_i^* > 0$. If $f'(e_1^* + \ldots + e_N^*) - 1 < 0$ then $e_i^* = 0$ for each i, which contradicts that $f(0) > 1$. Since f' is strictly decreasing and since there is an $\bar{e}$ such that $f'(N\bar{e}) < 1$, then equilibrium total effort is the unique solution to $f'(e_1^* + \ldots + e_N^*) = 1$.

Conversely, any strategy profile $(e^*, \tilde{\varepsilon}^*)$ such that $f'(e_1^* + \ldots + e_N^*) = 1$ is a subgame perfect equilibrium. The Kuhn-Tucker condition is sufficient since f is strictly concave. □

Proof of Proposition 2: Let $\alpha'' > \alpha'$. We show that $e^*(\alpha'') > e^*(\alpha')$. Suppose to the contrary that $e'' \equiv e^*(\alpha'') \leq e^*(\alpha') \equiv e'$. Then

$$1 = (1 - \alpha'' + \alpha''N) f'(Ne'') > (1 - \alpha' + \alpha' N) f'(Ne') = 1,$$

where the inequality holds since (i) $\alpha'' > \alpha'$ implies $1 - \alpha'' + \alpha''N > 1 - \alpha' + \alpha' N$, (ii) $e'' \geq e'$ and f' is decreasing implies $f'(Ne'') \geq f'(Ne')$, and (iii) $f' > 0$. This is a contradiction.

When $\alpha = 1$, then $e^*(0)$ solves

$$Nf'(Ne^*(1)) = 1,$$

and hence $e^*(0) = e^{WM}$. $\square$

Proof of Proposition 3: Assume $\alpha > 0$. Let $(e^*, \tilde{\varepsilon}^*)$ be a subgame perfect equilibrium in which

$$\tilde{\varepsilon}_i^j(e_1, \ldots e_{j-1}, e_j, e_{j+1}, \ldots, e_N) = \kappa + \beta e_j.$$

Given effort e_i and self evaluation $\tilde{e}_i^i$, then Player i's median evaluation is

$$m\{\tilde{\varepsilon}_1^i(e), \ldots, \tilde{\varepsilon}_{i-1}^i(e), \tilde{e}_i^i, \tilde{\varepsilon}_{i+1}^i(e), \ldots, \tilde{\varepsilon}_N^i(e)\} = \kappa + \beta e_i$$

and is therefore independent of their self evaluation. Player i's payoff is

$$(1-\alpha)f(e_i + \sum_{j \neq i} e_j^*) + \alpha f(N(\kappa + \beta e_i)) - e_i.$$

In equilibrium, e_i^* satisfies the Kuhn-Tucker condition

$$(1-\alpha)f'(e_i^* + \sum_{j \neq i} e_j^*) + \alpha N\beta f'(N(\kappa + \beta e_i^*)) - 1 \leq 0, \tag{3}$$

where the inequality holds with equality if $e_i^* > 0$.

We first establish that the equilibrium is symmetric, i.e., that $e_i^* = e_k^*$ for all $i \neq k$. Suppose to the contrary that there are i and k such that $e_i^* > e_k^*$. Since f' is decreasing, then $f'(Ne_i^*) < f'(Ne_k^*)$. Hence $\alpha > 0$ implies

$$(1-\alpha)f'(e_i^* + \sum_{j \neq i} e_j^*) + \alpha N\beta f'(N(\kappa + \beta e_i^*)) - 1 < (1-\alpha)f'(e_k^* + \sum_{j \neq k} e_j^*) + \alpha N\beta f'(N(\kappa + \beta e_k^*)) - 1 \leq 0,$$

where the weak inequality holds by (1). Hence $e_i^* = 0$, which contradicts

$e_i^* > e_k^*$. Thus in equilibrium $e^* = e_1^* = \ldots = e_N^*$.[29]

Zero total effort, i.e., $e^* = 0$, is not an equilibrium since $(1-\alpha)f'(0) + \alpha N\beta f'(N\kappa) - 1 > 0$. Thus in any equilibrium we have $e^* > 0$ and

$$(1-\alpha)f'(Ne^*) + \alpha N\beta f'(N(\kappa + \beta e_i^*)) - 1 = 0. \tag{4}$$

Next we establish that equilibrium is unique. As already noted, $(1-\alpha)f'(0) + \alpha N f'(N\kappa) - 1 > 0$. Furthermore,

$$(1-\alpha)f'(N\bar{e}) + \alpha N\beta f'(N\bar{e}) - 1 < 0,$$

since $f'(N\bar{e}) < Nf'(N\bar{e}) < 1$. Since

$$(1-\alpha)f'(Ne) + \alpha N\beta f'(Ne) - 1$$

is continuous, by the Intermediate Value Theorem there is a solution to (4). The solution e^* is unique since $(1-\alpha)f'(Ne) + \alpha N\beta f'(Ne) - 1$ is strictly decreasing in e as $f'' < 0$.

Finally, we show that $(e^*, \tilde{\varepsilon}^*)$ as given in Proposition 3 is an equilibrium. Since

$$(1-\alpha)f(e_i + \sum_{j \neq i} e_j^*) + \alpha f(Ne_i) - e_i$$

is concave in e_i, the first order condition is sufficient and Player i's effort is optimal given e_{-i}^*. Further, in any subgame $(e_1, \ldots, e_N)$, since the other players give i the same evaluation $\kappa + \beta e_i^*$, then player i's median evaluation is independent of their evaluation strategy $\tilde{\varepsilon}_i$, and thus it is optimal for i's self evaluation to be $\tilde{\varepsilon}_i^*(e) = \kappa + \beta e_i$. Likewise, i's payoff does not depend on their evaluation of j and thus it is optimal for i to evaluate j truthfully. □

[29] When there is no ambiguity, we abuse notation by writing e or e^* for a player's effort and equilibrium effort, respectively.

References

[1] Abeler, J., Nosenzo, D., and C. Raymond (2019), "Preferences for Truth-telling," *Econometrica,* 87, pp. 1115–1153.

[2] Barron, K. and T. Nurminen (2020). "Nudging cooperation in public goods provision," *Journal of Behavioral and Experimental Economics*, 88, pp. 1-16.

[3] Bicchieri, C. (2006). "The Grammar of Society: The Nature and Dynamics of Social Norms," Cambridge University Press.

[4] Cason, T. and Gangadharan, L. (2015). "Promoting Cooperation in Nonlinear Social Dilemmas Through Peer Punishment," *Experimental Economics*, 18, pp. 66-88.

[5] Carpenter, J., Matthews, P, and Schirm, J. (2010). "Tournaments and Office Politics: Evidence from a Real Effort Experiment," *American Economic Review*, 100, pp. 504-517.

[6] Carpenter, J., Robbet, A., and Akbar, P. (2017). "Profit Sharing and Peer Reporting," *Management Science*, 64, pp. 4261-4276.

[7] Dufwenberg, M., Görlitz, K., Gravert, C. (2024). Peer Evaluation Tournaments. CEBI Working Paper Series.

[8] Fehr, E. and S. Gächter (2000). "Cooperation and Punishment in Public Goods Experiments," *The American Economic Review*, 90, pp. 980-994.

[9] Kandori, M. (1992). "Social Norms and Community Enforcement," *Review of Economic Studies*, 59, pp. 63-80.

[10] Kandori, M., Mailath, G, and R. Rob (1993). "Learning, Mutation, and Long Run Equilibria in Games," *Econometrica*, 61, pp. 29-56.

[11] Kandel, E. and E. Lazear (1992). "Peer Pressure and Partnerships," *Journal of Political Economy*, 100, pp. 801-817.

[12] Krupka, E. and Weber, R. (2013). "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?," *Journal of the European Economic Association*, 11, pp. 495-524.

[13] Ledyard, J. (1995). "Public Goods: A Survey of Experimental Research," in The Handbook of Experimental Economics, pp. 111-194, edited by John Kagel and Alvin Roth.

[14] Laury, S., and C. Holt. (2008). "Voluntary Provision of Public Goods: Experimental Results with Interior Nash Equilibria," Handbook of Experimental Economics Results, Volume 1, pp. 792-801.

[15] Nikiforakis, N. (2008). "Punishment and counter-punishment in public good games: Can we really govern ourselves?," *Journal of Public Economics*, 92, pp. 91-112.