# Peer Evaluation Tournaments

Martin Dufwenberg[*] Katja Görlitz[†] Christina Gravert[‡]

April 13, 2023

**Abstract:** Experts (e.g., academics) or team members (e.g., in firms) are often best at evaluating each other. Peer evaluation tournaments could be useful for revealing performance. However, rampant opportunities for cheating may throw a wrench in the process, unless, somehow, players have a preference for honest reporting. In a recent paper, Dufwenberg and Dufwenberg (2018) offer a theory of perceived cheating aversion, which we argue can be naturally extended to our multi-player setting with subjective performance evaluations. Players trade off desire to win and dislike of being identified as a cheater. We derive a set of predictions, and test these in a controlled laboratory experiment.

*Keywords: psychological game, cheating, tournaments, laboratory experiment*
*JEL codes: C91*

# 1 Introduction

How can we reliably judge performance in situations where output is complex, multi-dimensional or involves team work? While a 100m track record can be objectively measured with a stopwatch, deciding whom to award the Nobel prize, promote to CEO or how much to pay for a piece of art requires expert knowledge. For a reliable evaluation, the quality of the contribution needs

[*] Department of Economics, University of Arizona; University of Gothenburg; CESifo; email: martind@eller.arizona.edu

[†] University of Applied Labour Studies; e-mail: katja.goerlitz@hdba.de

[‡] University of Copenhagen and CEBI; e-mail: cag@econ.ku.dk. The activities of CEBI are financed by the Danish National Research Foundation, Grant DNRF134.

to be judged by individuals who are themselves knowledgeable about the subject area. This often means that the ones who are best at evaluating are also the ones being evaluated. Peer evaluation tournaments are a common solution to this informational problem. For example, most of the scientific process is build on peer evaluations. Scientists working on related topics evaluate which papers will end up published in top journals, whose research topic will receive funding and which applicants will be hired.[1]

Despite their benefits, winner-takes-all tournament incentives combined with the subjective nature of the evaluation creates incentives for dishonesty or deliberate sabotage. Scientific expert evaluators compete for the same journal space and grants as those whom they evaluate and consultants are rivals for the same promotions or bonuses that are given out based on their feedback. Given the strong incentives for dishonesty, it is perhaps surprising that these institutions are so prevalent in science and business. For these institutions to work, the desire to cheat needs to be counteracted by, for example, the desire to be seen as an honest person. In this paper, we model the tradeoff between the desire to win by cheating in a peer evaluation tournament and the disutility from being perceived as a cheater by peers. We then test the model predictions in a novel lab experiment.

Our modeling approach builds on the recent theory by Dufwenberg and Dufwenberg (2018) (D&D) according to which people suffer "perceived cheating aversion" in proportion to how much others believe that they cheat. We extend the D&D model to a multi-player tournament setting in which players can cheat by over- and underreporting their delivered quality to win the tournament. The other players can observe the true quality of the output of each player (with some noise), but the submitted assessment is confidential and only the winner of the tournament can be observed. Each player suf-

[1] The private sector also makes use of peer evaluation such as consultancies that use 360 degree feedback for their promotion decisions. 360 feedback apps, like Culture Amp, claim to have over 2000 major companies as their customers. Expert knowledge in the private sector usually comes e.g. from team members who work together and are, thus, able to evaluate their colleagues' performance.

fers perceived cheating aversion to the extent that, conditional on winning, the other players think he should not have won given their observation of his true quality. The model predicts that although everyone cheats, players with lower quality will cheat less and thus win less often, but conditional on cheating, cheat by a higher amount. We also introduce a source of ambiguity. Players might cut the other players some slack given the subjective nature of the quality ratings. This ambiguity can, however, be exploited by the players to hide some of their cheating. We, thus, predict that reducing ambiguity reduces cheating.

We test our model predictions in a novel laboratory experiment. A group of five players compete in a creativity task, coming up with original uses for a piece of paper. They then evaluate their own and each others' quality. Based on their assessments the winner is determined. The players do not observe the ranking of everyone else, but only who the winner is. This feature makes individual cheating possible and not directly verifiable. The nature of the task, evaluating creative output, leaves room for ambiguity. In a between-subjects treatment, we manipulate the perceived ambiguity regarding whether or not someone cheated, by providing a payment-irrelevant objective quality ranking to which the winner can be compared.

Our experimental setting is stark. If the players are entirely selfish, then the peer-evaluation institution that we study would be, essentially, useless. The incentives to over-report ones own performance, and to under-report that of others, would be so strong as to rule out any positive correlation between quality and reward. One the other hand, if the predictions of our theory are supported then this would provide some measure of hope that the institution that we study is useful.

Our paper contributes to the literature on sabotage in tournaments. Sabotage can either take the form of deliberately decreasing another person's output (Lazear, 1989; Carpenter et al., 2010), inflating one's own performance (Cadsby et al., 2010; Harbring and Irlenbusch, 2011; Conrads et al.,

2014) or both (Charness et al., 2014).[2] It is not surprising that sabotage occurs in tournaments, however, none of the studies find significant amounts of complete sabotage, despite high potential monetary gains from cheating. Understanding what counteracts the desire to cheat in tournaments is crucial in designing better mechanisms to reduce sabotage further. Our theoretical model can help interpret the behavior observed in this mostly empirical literature. We also add to the knowledge on the usefulness of 360 degrees feedback.[3]

Section 2 presents the model and derives testable predictions. Section 3 presents the experiment. Section section 4 reports results. Section 5 concludes.

# 2 The Model

The model in Dufwenberg and Dufwenberg (2018) focuses on the popular experimental "die roll paradigm" introduced by Fischbacher and Föllmi-Heusi (2013) (F&FH), and before we extend these ideas to peer evaluation tournaments is is useful to recall what F&FH and D&D did.

F&FH run experiments where subjects are asked to privately roll and report the outcome of a die-roll, and they get paid in proportion to how high a number they the report. Neither full honesty (each number is equally likely to be reported) nor full selfishness (everyone reports the highest-paying number) is observed, but something "in between," with higher numbers being more likely to be reported although all reports occur with positive probability.

In D&D's theory nature randomly draws $x \in \{0, ..., n\}$, $n \geq 1$, with probability $\pi_x \in (0, 1)$, $\sum_x \pi_x = 1$.[4] A decision maker (DM) observes $x$ and

[2] Gangadharan et al. (2020) survey studies of antisocial behavior in the workplace.

[3] While the previous literature sheds light on psychological or management aspects of 360 degree feedback (Beehr et al., 2001; Atkins and Wood, 2006; Buckingham and Goodall, 2015), less is known from an economic perspective (see Sliwka (2020) for a review of the literature on the economics of incentives in firms that includes subjective performance evaluations).

[4] F&Hs's setup is the special case where $n = 5$ and $\pi_x = \frac{1}{6}$ for all $x$.

is asked to report it, but the report is non-verifiable; DM can report any $y \in \{0, ..., n\}$ and is then paid $y$ units of money. An audience observes $y$, not $x$. A (behavior) strategy for DM is a function $s : \{0, ..., n\} \to \Delta\{0, ..., n\}$. Let $p(x'|y) \in [0, 1]$ be the probability the audience assigns to $x = x'$ given $y$, with $\sum_{x'} p(x'|y) = 1$. DM suffers from *perceived cheating aversion* to the extent that the audience believes he is cheating; DM's utility at $(x, y)$ equals

$$y - \theta \cdot \sum_{x'<y} (p(x'|y) \cdot (y - x')) \tag{1}$$

where $\theta \geq 0$ measures sensitivity to perceived cheating. The second term reflects how much DM is perceived to cheat and how much he suffers. (1) is independent of $x$; DM cares about his image, not cheating per se. (1) depends on the audience's beliefs, via $p(x'|y)$, generating a psychological game Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009). D&D explore equilibria and their (excellent) fit with data.[5]

**Peer-evaluation tournaments:** We adapt D&D's notion of perceived cheating aversion to our setting, which however is so different that many new modeling decisions must be made. Consider a tournament with $N > 1$ active players, not a single DM. That $N$-some constitute each others' audience. The counterpart to D&D's $x$ is now the players' true "qualities," observed by all. The counterpart to D&D's $y$ is the reports of own and others' qualities that players submit. However, $i$'s payoff now depends on all reports, not just $i$'s own. And $i$'s co-players do not observe $i$'s report, but merely who won.

We initially make some extreme assumptions regarding players' strategy sets and choices which allow us to generate key intuitions easily. The experi-

[5] See Abeler et al. (2019) for a survey of the (more than a hundred) experiments that were conducted with the die-roll paradigm. They also discuss various theoretical approaches, and conclude that D&D's theory (along with another approach, represented by Gneezy et al. (2018); Khalmetski and Sliwka (2019) is one of the (few) that are consistent with the data.

mental design will involve some differences the impact of which we comment on later (Remarks 1-2).

Let $(x_i)_{i\leq N} \in \mathbb{R}^N_+$ be the profile of players' qualities, where we choose indices such that $x_j \geq x_i$ if $j > i$. The players observe $(x_i)_{i\leq N}$ (with some "doubts," discussed wrt $\varepsilon$ below), and then each $i$ simultaneously files a report $(y_{ij})_{j\leq N} \in \mathbb{R}^N_+$, where $y_{ij}$ is $i$'s report of $j$'s quality. A single winner is selected based on who got the highest overall reported quality ($= \max_i \Sigma_{j\leq N} y_{ji}$); if there are ties a winner is selected at random. We normalize payoffs so that the winner's prize equals 1, while the others get 0.

If the players were motivated solely by desire to win, the game wouldn't have a equilibrium (since there is no upper bound on reports). However, we assume that the players are also motivated by perceived cheating aversion bestowed on the winner. To get at that, first note that a meaningful notion of how much $i$ cheats can be defined as follows

$$\max\{y_{ii} - x_i, 0\} + \frac{1}{N-1} \cdot \sum_{j\neq i} \max\{x_j - y_{ij}, 0\} \tag{2}$$

The first term reflects how much $i$ over-reports own quality (if a positive amount). Of course, rather than cheat that way, $i$ could achieve the same effect by instead under-reporting each of the others just as much; the second term reflects that. We assume that each $i$ derives disutily in proportion to how much cheating the others *believe* that $i$ *intentionally* engages in, and solve for equilibrium. Three further assumptions bear on the analysis:

- We (initially) look for equilibria such that $i$ only cheats by over-reporting own quality, not by under-reporting others' qualities. [This is not essential, as we explain shortly after Proposition 2 below.]

- We focus on equilibria such that each $i$ engages only two choices: to be honest meaning to file a report $(y_{ij})_{j\leq N} = (x_j)_{j\leq N}$ for all $j$, or to cheat to a "common target" $T > N \cdot x_N$ by filing a report $(y'_{ij})_{j\leq N}$ such that $y'_{ij} = x_j$ for all $j \neq i$ while $y'_{ii} = T - (N-1) \cdot x_i$. He may use a mixed strategy, randomizing across those two choices. Let $c_i \in [0,1]$ be the probability with which $i$ cheats. [We briefly acknowledge other types of equilibria, in footnote 5 below.]

- When assessing $i$'s cheating, there are many sources of ambiguity that we so far neglected.[6] As a catch-all for this, we introduce $\varepsilon \in (0,1)$. Let $\widehat{c}_i \in [0,1]$ be each player other-than-$i$'s (point) belief about $c_i$.[7] We assume that if $j \neq i$ observes that $i$ wins then every other player calculates the conditional probability that $i$ cheated as

$$\frac{(1-\varepsilon)\cdot\widehat{c}_i}{(1-\varepsilon)\cdot\widehat{c}_i + \frac{1_{\{j|x_j=x_N\}}(i)}{|\{j|x_j=x_N\}|}\cdot\Pi_{j\leq N}(1-\widehat{c}_j)+\varepsilon} \tag{3}$$

  If $\varepsilon \to 0$, (3) tends to the true cheating probability; any $j$ with $x_j = x_N$ wins without cheating with probability $\Pi_{j\leq N}(1-c_j)$ and the second term in the denominator reflects others' corresponding belief ($1_{\{j|x_j=x_N\}}$ is an indicator function for $x_i = x_N$). The presence of $\varepsilon$ in (3), biases the calculation. The idea: others cut $i$ some slack, assigning probability $\varepsilon$ to the possibility that $i$ isn't knowingly cheating but rather made his realized choice for any other reason.

We are ready to define $i$'s utility, focusing on numbers relevant to an equilibrium with the structure described via the above bullets. First, we assume that if $i$ does not win then there is no perception that he cheats, and his utility simply equals his material payoff of 0. Second, conditional on winning, $i$'s utility is defined as follows:

$$1-\theta\cdot\frac{(1-\varepsilon)\cdot\widehat{c}_i}{(1-\varepsilon)\cdot\widehat{c}_i + \frac{1_{\{j|x_j=x_N\}}(i)}{|\{j|x_j=x_N\}|}\cdot\Pi_{j\leq N}(1-\widehat{c}_j)+\varepsilon}\cdot(T-N\cdot x_i) \tag{4}$$

[6] Maybe there is noise in how players evaluate quality, others' & own. Maybe $i$ mistakenly believes his quality is higher than it is, so that he isn't knowingly cheating even if he gives himself a high score. Or maybe $i$ has high quality and it is $j$ who is mistaken.

[7] If it looks restrictive that all players have the same beliefs about $i$, this will soon be justified by our focus on equilibrium.

which should be read as “material payoff minus a pang of perceived cheating.” To explain the second term, walk through its factors in reverse order:

- $T - N \cdot x_i = [T - (N-1) \cdot x_i] - x_i$, so this factor reflects how much $i$ over-reports own quality. The amount of cheating is judged in proportion.
- The middle factor (=(3)) is the probability that $j \neq i$ assigns to a winning $i$ being a cheater.[8]
- $\theta \geq 0$ measures sensitivity to perceived cheating.

**Definition:** An $T$-*equilibrium* is a strategy profile $(c_i)_{i \leq N}$ in which all players use cheating-to-$T$ strategies as described above. Moreover, each player is maximizing his expected utility using that strategy rather than any other, and given that $(\widehat{c}_i)_{i \leq N} = (c_i)_{i \leq N}$.

**Proposition 1:** *If $\theta > 0$ then a $T$-equilibrium $(c_i)_{i \leq N}$ always exists.*

**Proof:** The proof is constructive. It cannot hold that $c_i = 0$ for any $i$. To see this, note that $c_i = 0$ would imply that (3) equals 0 regardless of $i$'s choice. In that case $i$ could unilaterally gain by cheating enough to win with impunity, regardless of $\theta$.

For each $i$, it must hold that the utility from winning equals that of not winning, which equals 0 by assumption. To see this, note first that since $c_i > 0$ for all $i$, and since there is cheating to a common target $T$, each player wins with positive probability. If $i$'s utility of winning were lower than 0 then $i$ would have a unilaterally profitable deviation by reporting $y_{ii}$ so low that winning were impossible. If instead $i$'s utility of winning were higher than 0 then $i$ would have a unilaterally profitable deviation to report $y_{ii} > T - (N-1) \cdot x_i$.[9] Thereby $i$ would win for sure rather than with

[8] Since $i$'s utility depends on other's beliefs $(\widehat{c}_i)$, we again have a psychological game.

[9] Note: Others don't observe $y_{ii}$ but merely who won, so the perceived cheating of winner $i$ is independent of $y_{ii}$.

probability lower than 1. Using (4), we get:

$$1-\theta\cdot\frac{(1-\varepsilon)\cdot\widehat{c}_i}{(1-\varepsilon)\cdot\widehat{c}_i+\frac{1_{\{j:x_j=x_N\}}(i)}{|\{j:x_j=x_N\}|}\cdot\Pi_{j\leq N}(1-\widehat{c}_j)+\varepsilon}\cdot(T-N\cdot x_i)=0. \quad (5)$$

Now, fix $\theta > 0$ and consider player $N$. Make (5) hold for $i = N$ by selecting $\widehat{c}_N = c_N > 0$ and $T > N \cdot x_i$, appropriately. (Note that this can be done in infinitely many ways.) Then consider each $i < N$. Given $T$ as just determined, make (5) hold by selecting $\widehat{c}_i = c_i \in (0, c_N]$, appropriately. (There is a unique way to do this.) Since the lower is $i$ the higher is $T - N \cdot x_i$, it follows that the lower is $x_i$ the lower is the appropriate $c_i$.

To verify that the strategy profile just constructed is indeed an $\varepsilon$-equilibrium, note that each player is indifferent between cheating to the common target and reporting honestly (getting zero utility in either case), and in fact also as regards using any other strategy since the involved experiences (winning while being perceived as a cheater, or not winning) remain the same and always involve zero utility. ■

At the cost of mathematical complexity, we could have defined a more general notion of equilibrium such that the $T$-equilibrium would be a special case.[10] However, it is natural to focus on $T$-equilibria for two reasons: First, they are simple to describe. Second, they exhibit the potential of peer evaluation tournaments to reveal information in a systematic way. $T$-equilibria are not unique; as seen in the proof of Proposition 1 the involved value of "$T$" is not unique. However, all $T$-equilibria share several striking properties:

**Proposition 2:** *Let $(c_i)_{i\leq N}$ be a $T$-equilibrium and $(w_i)_{i\leq N}$ the associated probabilities with which each player i wins. The following is true for all i and j:* (i) $c_i > 0$ and $w_i \in (0, 1)$. (ii) If $x_j > x_i$ then $c_j > c_i$ and $w_j > w_i$. (iii) If $x_j > x_i$, $y_j \neq x_j$, $y_i \neq x_i$ then $y_i - x_i < y_j - x_j$.

[10] This would be an adaptation of Battigalli & Dufwenberg's equilibrium notion to our setting. The adaptation is that our players have infinite rather than finite pure strategies.

**Proof:** (i) It was seen at the start of the proof of Proposition 1 that $c_i > 0$ for all $i$. Since all players cheat to the common target $T$, *all* players win with strictly positive probability, implying $w_i \in (0, 1)$ for all $i$. (ii) It was seen in the penultimate paragraph of the proof of Proposition 1 that the higher is $x_i$ the higher is $c_i$. Since all cheating is to the common target $T$, the higher is $c_i$ the higher is $w_i$. (iii) This follows directly from the construction of $c_i$ and $c_j$, with cheating to the common target $T$. ■

In words: (i) All players cheat, and all players win, with positive probability. Regardless of $\theta$, no one can be fully trusted! (ii) The higher is a player's quality the more likely he is to cheat. The higher is a player's quality, the more likely he is to win! (iii) Conditional on cheating, the lower is a player's quality the more he will cheat. Players with lower qualities do not cheat as often as others, but when they do, watch out!

In a $T$-equilibrium, cheating involves only over-reporting of own quality. Countless other equilibria could be constructed that also (or instead) involve under-reporting of others' qualities. To see this, note that the marginal effect to every player of $i$ adding an amount $\Delta$ to his reported own score is the same as that of $i$ deducting $\Delta$ from the reported score of every other player. *Mutatis mutandis*, all equilibria thus constructed would share the properties highlighted in Proposition 2.[11]

Our final result addresses the impact of ambiguity. Namely, the higher is $\varepsilon$ the more likely $i$ is deemed to not cheat even if he wins. This shelters $i$ from others opprobrium, to some degree, so the more likely is $i$ to cheat:

**Proposition 3:** *Fix $T$ s.t. $(c_i)_{i \leq N}$ is a $T$-equilibrium if $\varepsilon = \delta > 0$ while $(c_i')_{i \leq N}$ is a $T$-equilibrium if $\varepsilon \in (0, \delta)$. It holds that $c_i' < c_i$ for all $i$.*

[11] Had we defined a broader notion of equilibrium, where not all players cheat to the same value(s) of $T$ then it need neither be the case that higher-quality players win more often nor that lower-quality players cheat in larger quantities. One can show that $w_i > 0$ would always be implied though.

**Proof:** This follows immediately from inspecting (5). ■

Two remarks pave the way for our experimental tests:

**Remark 1:** If one modifies the above games to incorporate upper bounds on reports, the main intuitions captured by Propositions 1-3 largely remain. To see this, fix a $T$-equilibrium as described. Consider a modified game such that reported qualities $y_{ij}$ cannot exceed $M > 0$. Obviously, as long as $M \geq T$ the strategy profile that was an equilibrium of the original game remains an equilibrium in the new game. However, it may seem that a problem occurs if instead $M < T$. This is, however, a mirage in the sense that one can redefine units, "making $T' < M$ the new $T$," as follows: Redefine qualities $(x_j)_{j \leq N}$ and sensitivity $\theta$ as $(x'_j)_{j \leq N}$ and $\theta'$ such that $x'_j = \frac{T'}{T} \cdot x_j$ for all $j$ and $\theta' = \frac{T}{T'} \cdot \theta$. This re-creates the old $T$-equilibrium such that $T$ changes to $T'$, but the involved strategies $(c_i)_{i \leq N}$ are exactly the same as before. $\theta$ is higher, yes, but this is just matching the quality adjustments, just like the value of money is invariant to currency conversions.[12]

**Remark 2:** The $T$-equilibria looked at so far presume that there are no integer constraints on reports. Further adjustments must be made in their presence to maintain the spirit of the predictions made. For example, and in anticipation of the experiment below, let $N = 5$ and suppose that it must hold that $y_{ij} \in \{0, 1, ..., 10\}$ for all $i, j$. Suppose that, following the previous remark, we describe qualities as fractions of 1, so that $x'_i \in [0, 1]$, with $x'_N = 1$. We can now construct an equilibrium resembling the previous ones by replacing "honest reporting" with "no over-reporting." For example, let the "no over-reporting" choices for any $i$ such that $x_i < x_5$ involve $y_{ij} = 0$

[12] The arguments made here presume that there are no integer constraints on reports. Further adjustments must be made in their presence. However, as long as there is a clear yardstick for what report corresponds to "honest" reporting of $x'_N$, as regards winning probabilities the equilibrium can be recreated by letting any $i$ such that $x'_i < x'_N$ submit a lower report than $N$ would.

for all $j$, while for $i$ such that $x_i = x_5$ it involves $y_{ii} = 1$ and $y_{ij} = 0$ for $j \neq i$. Cheating, on the other hand, would be to some $T \in \{2, 3, ..., 10\}$ (so $y_{ii} = T$ if $x_i < x_N$ and $y_{ii} = T - 4$ if $x_i = x_N$). Unlike in the non-discrete case, $\theta$ may now have to pass a higher bound than just $\theta > 0$. And, in the example, the lower is $T$, the higher $\theta$ must be.

Let us wrap up this section by summarizing the spirit underlying any $T$-equilibrium: Players either report honestly, and likely don't win, or they exaggerate, and have a decent shot of winning. They are indifferent between these two modes of behavior, because in order to win the degree of cheating needed is just high enough that (conditional on winning) the sweetness of the material prize is exactly counterbalanced by pangs via others' suspicion that one cheated. In the next section, we test whether this story, as told by the notion of $T$-equilibrium and Propositions 2 & 3, is empirically relevant.

# 3 Experimental Design

We now turn to the experiment that tests the theoretical predictions of our model. Participants compete in groups of five for a prize. Each player performs a task that all other players (including the player him or herself) rate on a scale from 0 to 10. The person with the highest total score in the group wins. To make peer evaluation meaningful and to introduce some uncertainty $\varepsilon$, we chose a creativity task that is subjective, leaving scope for cheating when rating one's competitors. We used the "unusual uses task" (Guilford, 1967) where players should come up with as many unconventional uses for a piece of paper (e.g. make a hat, dry wet shoes, insulate a house) that they could think of. The experiment was run in sessions consisting of ten players. Within each session, participants were randomly assigned to their competitors. They never knew with whom of the other players they compete for the prize.

**The experimental set-up** All participants were informed that they took part in a tournament consisting of five players where the winner gets a 500 SEK (≈50 USD) prize and the losers received a 50 SEK (≈5 USD) show up fee. To guarantee anonymity, each subject got an ID number assigning them randomly to a group of five players who they would compete against for the prize. Next, the creativity task was introduced and explained in detail. Subjects were informed that it was in their best interest to perform well in the task to increase their chances of winning. Importantly, they were not informed about the scoring mechanism when performing the task to rule out that the scoring could affect the creative performance of the players.

After a 3 minute practice round on unusual uses for an old tire, the experimenter distributed the sheets for the incentivized task. Subjects had three minutes to come up with unusual uses for a piece of paper. When the time was up, the experimenter collected all answer sheets and handed out new instructions with the scoring rules. Subjects then received a copy of the answer sheet of each member of their group including their own and were asked to score each answer sheet on a score from 0 to 10. The instructions stated: "The winner of the 500 SEK will be determined by the following procedure: You will now evaluate your own answer and the answers of the other four players with whom you compete for the 500 SEK. Please evaluate the answers with respect to their originality. Originality is scored for each person on a scale from 0 to 10 where 0 indicates overall "not at all original answers" and 10 "very original" answers. For scoring, take into consideration i) the number of answers, ii) their degree of being unusual and iii) the number of different categories they come from. The other players in your group will also do the same scoring. The points given and the points received are kept anonymous by the research team as well as the information who are the players in your group. For each player, the research team will add up the points given to a TOTAL SCORE of a minimum of 0 and a maximum of 50. The person with the highest TOTAL SCORE out of your group will receive 500 SEK (including the show-up fee)." The ID of the winner was announced at the end of the experiment. Individual scores and the identity of the winner were kept anonymous by the experimenter. If there was a tie for the highest TOTAL SCORE, the winner was determined randomly which was the case one time only. After scoring, subjects filled out a questionnaire on general characteristics and demographics.

We summarize the process as follows:

1. General overview of the tournament setting
2. Instructions for the creativity task
3. Practice round "Old tire"
4. Incentivized round "Piece of Paper"
5. Information about the scoring rules that differed by *Treatment*
6. Scoring of all five answer sheets
7. Questionnaire
8. Announcement of winner

**Treatments** There is a baseline treatment and an objective-ranking treatment. Both treatments followed the process as outlined above and the scoring rules were the same. The only difference was that players in the baseline treatment only got the scoring rules, while players in the objective-ranking treatment obtained further information after the scoring rules were announced, but before they made their scoring. In particular, they were informed that we used the answers of more than 100 test persons who did the same task in a prior experiment to generate (what we called) an objective score for each participant. We also informed them that each player would be able to see these OBJECTIVE SCORES for each player in their group at the end of the experiment. The instructions emphasized again that winning is solely determined based on the TOTAL SCORE (composed of the scores given by one's players in the group and one's own assessment of the task). The objective-ranking treatment reduced the uncertainty about the objective creativity. With this treatment variation, we intended to reduce $\varepsilon$ from our model, which is defined as any ambiguity in regard to assessing a players cheating.[13]

**Calculating the objective rank** We created the objective ranking following Bradler et al. (2019). The objective creativity increases with i) the number of valid answers, ii) the number of distinct categories the ideas come from (e. g. ring, necklace, bracelet belong only to one category which is "jewelry") and iii) originality. Original ideas are those being mentioned less than 8% and very original answers were named less than 1%. Because Bradler et al. (2019) provided us with their participants' answers of the unusual uses task for a piece of paper, we could calculate for each answer how often it was mentioned. Because the objective score of Bradler et al. (2019) had no maximum of 10, we had to transform their scale to our OBJECTIVE SCORE ranging from 0 to 10 by keeping the distribution of original answers identitical.[14]

**Further information** The experiment was conducted at the experimental laboratory of the University of Gothenburg. It was a pen and paper experiment and all earnings were paid out in cash or via a direct payment app right after the experiment. We conducted 14 sessions with ten participants in each session, 70 participants per treatment. Treatments were assigned at the session level. The sessions lasted up to 60 minutes. Average earnings were 140 SEK. 57 percent of the participants were female. See the Appendix

[13] We conducted one additional treatment which is not suitable to test the theory, as individuals did not see the objective ranking in the end.

[14] Because this transformation did not change the distribution of the objective creative performance, by definition the results should be unaffected by using either of the two scales. We confirm this empirically when reestimating all results with the original objective score. Using the scale ranging from 0 to 10 is closer to our theoretical model which is why we decided to use it in our main analyses.

for the full instructions.

# 4 Results

**Testing proposition 1** In line with the general version of the model, the experimental design allows individuals to cheat by exaggerating their own quality or by underreporting the quality of their competitors. The strategy that maximizes each players' probability of winning is to give themselves the best possible score (10) and all the other players the worst possible score (0). However, this strategy of reporting 10-0-0-0-0 is only followed by 5.7 percent of the individuals which refers to 8 individuals only. This low amount of individuals playing the dominant strategy suggests that the majority of players were not solely motivated by winning the prize. In accordance with equation 2 from our model, we can further analyze cheating by approximating it by:

$$(y_{ii} - x_i) + \frac{1}{4}\sum_{j \neq i}(x_j - y_{ij}) \tag{6}$$

This proxy for cheating considers overreporting own quality in the first term and underreporting the other players' quality in the second term. In the case of honest reports of all players, this measure should be zero on average. If players cheat instead, it is larger than zero. Figure 1 shows that this measure differs statistically significantly from zero in both treatment. Even though we cannot directly test proposition 1, the presented empirical evidence is in line with proposition 1.

**Testing proposition 2** Proposition 2 of our model gives us several testable predictions. First, it follows that since all players cheat, all players should win with positive probability. Figure 2 shows that individuals with the best objective quality (rank 5) wins with a probability of 38 percent and the second best player with 30 percent. The individuals with third and fourth highest

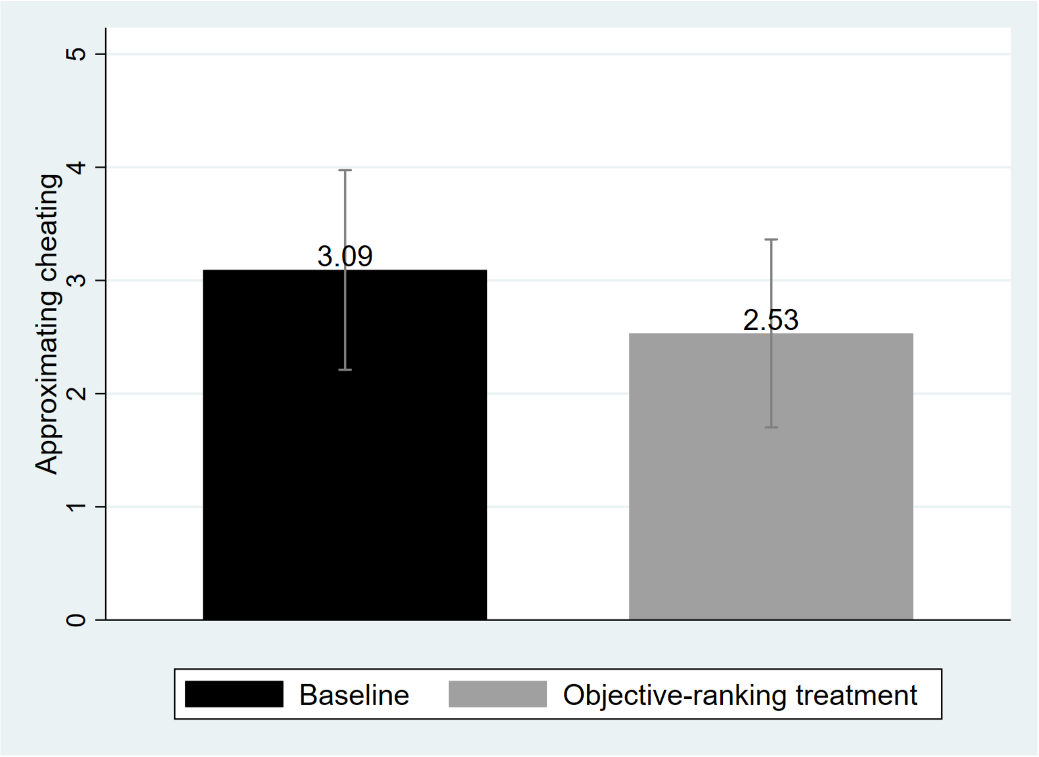

Figure 1: Approximating cheating by treatments

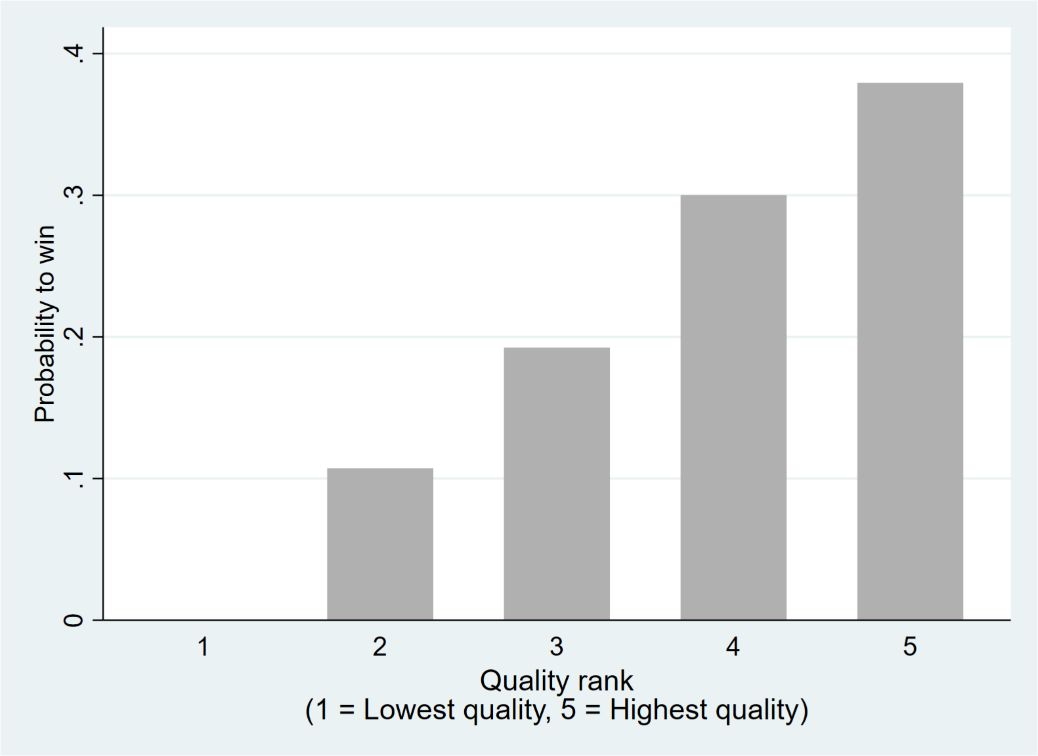

Figure 2: Winning probability by rank according to the objective score

quality win with 19 and 11 percent, respectively. The worst player (rank 1) never wins in our data. In sum, the data shows that also lower ranked players have a positive probability of winning, even though we cannot fully confirm our hypothesis.

Second, the $T$-equilibrium implies that higher-quality individuals win more often. Figure 2 documents that this is generally the case. The probability of winning increases by objective rank. The probability of winning with the highest quality (rank 5) is significantly higher than the probability of winning with the two lowest ranks in both treatments (individuals $\chi^2$-tests comparing rank 5 to ranks 4-1 respectively: $p < 0.78$, $p < 0.41$, $p < 0.09$, $p < 0.03$ ). Overall, the experimental data supports the second implication from Proposition 2, because cheating is common given that the best quality

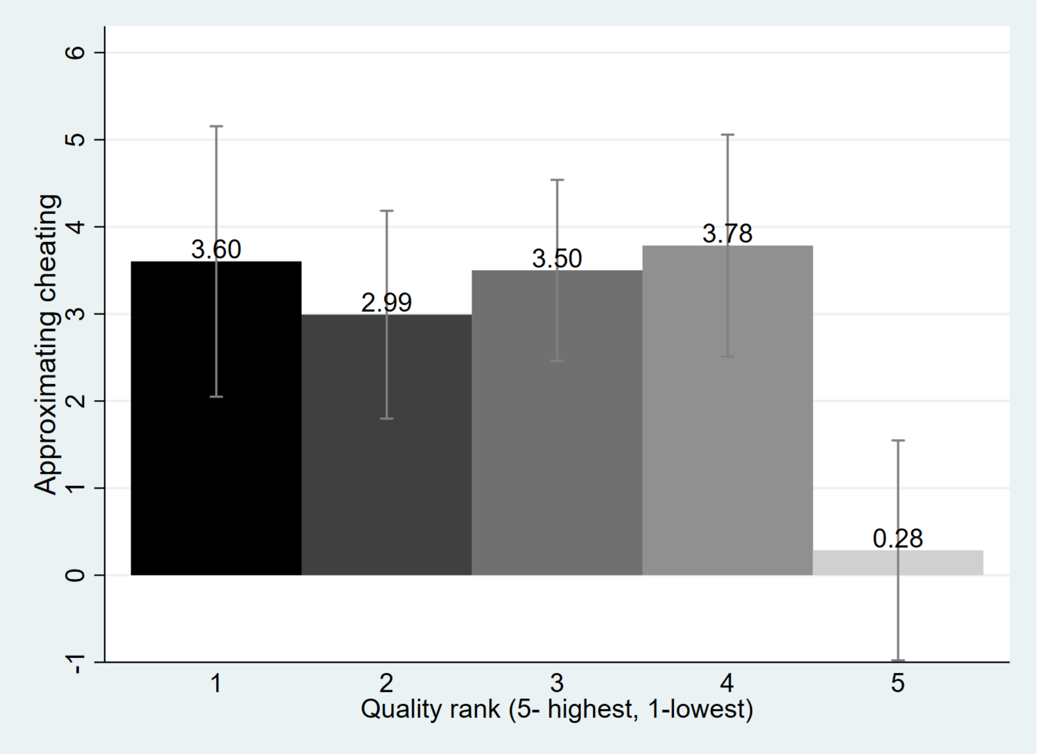

Figure 3: Winning probability by rank according to the objective score

player only wins in 1 out of 3 games, but they have the highest probability of winning compared to the lower ranked players.

Third, conditional on cheating, lower ranked players cheat to a larger extent. Figure 3 plots how our proxy for cheating (see equation 6) differs by the players' objective quality rank. Players with the highest rank have a much lower probability of cheating. This difference turns out to be statistically significant using a two-sided Wilcoxon rank sum test ($z = -4.10$, $p < 0.00$). Testing the hypothesis that the approximation of cheating is larger than the medium when comparing rank 5 with the other ranks, we perform the Pearson $\chi^2$, reinforcing that higher ranked players indeed cheat less of often ($\chi^2 = 8.03$, $p = 0.01$)

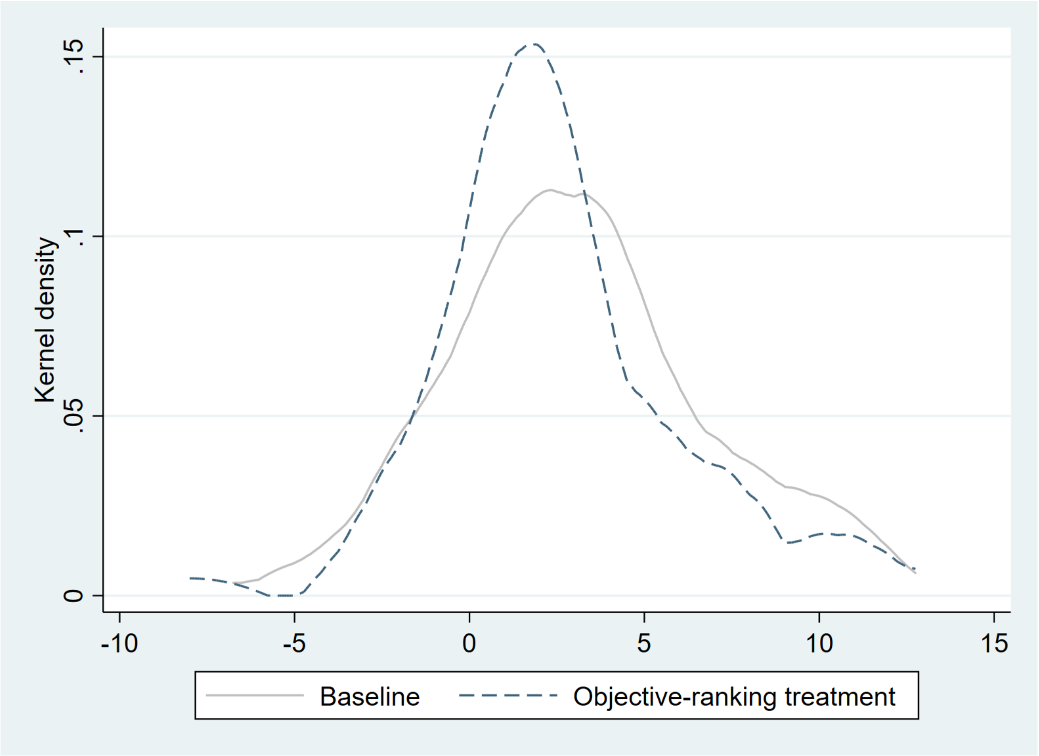

Figure 4: Cheating by treatment

**Testing proposition 3** Next, we test our hypothesis deriving from proposition 3, which implies that increasing the ambiguity of the evaluations increases cheating. The higher the ambiguity, the more slack the other players will cut their peers. In the context of our experiment, this means that if players are not aware during scoring that there will be objective ranks to compare their rankings to, then the model predicts higher amounts of cheating. Comparing the approximation of cheating by treatment status eliminates all differences of personality or other non-observed factors because of the random assignment of individuals to treatment groups.[15] Figure 1 has already shown that there is more cheating in the baseline treatment. While this difference is not statistically significantly ($z = 1.15$, $p = 0.252$, Wilcoxon rank sum, two-sided), the median is significantly larger in the baseline compared to the objective-ranking treatment ($\chi^2 = 3.46$, $p = 0.06$). Figure 4 illustrates the kernel density of approximated cheating, showing that the distributions differ by treatments. Overall, the provided evidence confirms the hypothesis derived from Proposition 3 that increasing the ambiguity increases cheating.[16]

# 5 Conclusion

Many high-stakes environments such as science and business rely on expert evaluators, who can adequately judge the contribution of an individual. Given the high level of expertise needed, the evaluators are also often competitors to the ones they are evaluating. This setting creates incentives to cheat. Depending on the peer tournament setting, this could happen by understating the performance of one's competitors or overstating one's own contribution. If individuals are completely selfish, peer tournaments would

[15] The treatment samples are balanced with regard to exogenous variables such as gender and or creative performance according to the objective score.

[16] This finding is mainly driven by men who report much higher values on cheating in the baseline only. As our model does not predict gender differences, we will not discuss this issue further.

be used strategically to win the competition by which they were unrelated to the actual performance of the persons who gets evaluated.

This paper proposes a psychological game theory model that can explain why cheating is not as prevalent, as one might expect. This is because the disutility from being perceived as a cheater counteracts the desire to cheat to win. Individuals differ in the extent of perceived cheating aversion, meaning how much they suffer when others believe they are cheating. Because most settings requiring peer tournaments have in common that the output is not (easily) observable, the model introduces ambiguity. The model predicts that the higher the ambiguity about peers' actual performance is, the higher gets cheating, because ambiguity also reduces others beliefs about cheating. Applied to the scientific process this would mean that if scientists care about their reputation among their peers, they will balance their desire to win with the potential shame of being perceived as a cheater.

We test the predictions of the model in a laboratory experiment in which five players compete in a winner-takes-all tournament. After completing a creativity task, we informed each person that evaluations determine the winner of the high stakes price. Each player had to evaluate the outputs of their four competitors and their own output. This setting allowed overreporting of one's own performance and understating competitors' performance. The two treatments differed in their degree of ambiguity. While the baseline conveyed no further information, the objective-ranking treatment informed each subject that we will create an objective score that each player can see at the end of the experiment. Again, we emphasized that this score had no impact on winning. The results support the predictions from the theory as maximum cheating in which an individual gives themselves the highest evaluation score and their competitors the worst is a rare event. Reducing the ambiguity about whether someone cheated, by providing an objective ranking, reduces cheating.

Our model and our empirical findings provide good news for expert evaluation systems. While cheating is observed, it is not as prevalent, as one might expect. If possible, additional objective criteria should be used to complement peer evaluations, as well as possible reputation losses made salient. In the scientific review process, the editor already serves as such objective authority. In business, managers could make peer evaluations public so that persons being evaluated serve as the audience that mitigates cheating.

# 6 Acknowledgements

We thank Christiane Bradler, Susanne Neckermann and Arne Jonas Warnke for providing us with their data from the unusual uses task.

# References

Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for truth-telling. *Econometrica 87*(4), 1115–1153.

Atkins, P. W. B. and R. Wood (2006). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence from 360-degree feedback programs. *Personnel Psychology 55*(4), 871–904.

Battigalli, P. and M. Dufwenberg (2009). Dynamic psychological games. *Journal of Economic Theory 144*(1), 1–35.

Beehr, T. A., L. Ivanitskaya, C. P. Hansen, D. Erofeev, and D. M. Gudanowski (2001). Evaluation of 360 degree feedback ratings: relationships with each other and with performance and selection predictors. *Journal of Organizational Behavior 22*(7), 775–788.

Bradler, C., S. Neckermann, and A. J. Warnke (2019). Incentivizing creativity: A large-scale experiment with performance bonuses and gifts. *Journal of Labor Economics 37*(3), 793–851.

Buckingham, M. and A. Goodall (2015). Reinventing performance management. how one company is rethinking peer feedback and the annual review, and trying to design a system to fuel improvement. *Harvard Business Review*, 1–10.

Cadsby, C. B., F. Song, and F. Tapon (2010). Are you paying your employees to cheat? an experimental investigation. *The BE Journal of Economic Analysis & Policy 10*(1).

Carpenter, J., P. H. Matthews, and J. Schirm (2010). Tournaments and office politics: Evidence from a real effort experiment. *American Economic Review 100*(1), 504–17.

Charness, G., D. Masclet, and M. C. Villeval (2014). The dark side of competition for status. *Management Science 60*(1), 38–55.

Conrads, J., B. Irlenbusch, R. M. Rilke, A. Schielke, and G. Walkowitz (2014). Honesty in tournaments. *Economics Letters 123*(1), 90–93.

Dufwenberg, M. and M. A. Dufwenberg (2018). Lies in disguise–a theoretical analysis of cheating. *Journal of Economic Theory 175*, 248–264.

Fischbacher, U. and F. Föllmi-Heusi (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association 11*(3), 525–547.

Gangadharan, L., P. J. Grossman, and J. Vecci (2020). Antisocial behavior in the workplace. *Handbook of Labor, Human Resources and Population Economics*.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). Psychological games and sequential rationality. *Games and Economic Behavior 1*(1), 60–79.

Gneezy, U., A. Kajackaite, and J. Sobel (2018). Lying aversion and the size of the lie. *American Economic Review 108*(2), 419–53.

Guilford, J. P. (1967). Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior 1*(1), 3–14.

Harbring, C. and B. Irlenbusch (2011). Sabotage in tournaments: Evidence from a laboratory experiment. *Management Science 57*(4), 611–627.

Khalmetski, K. and D. Sliwka (2019). Disguising lies—image concerns and partial lying in cheating games. *American Economic Journal: Microeconomics 11*(4), 79–110.

Lazear, E. P. (1989). Pay equality and industrial politics. *Journal of Political Economy 97*(3), 561–580.

Sliwka, D. (2020). Bonus plans, subjective performance evaluations, and career concerns. *Handbook of Labor, Human Resources and Population Economics*.