Peer review and competition in the Art Exhibition Game

Stefano Balietti[a,b,c,1], Robert L. Goldstone[d], and Dirk Helbing[e]

[a]Network Science Institute, Northeastern University, Boston, MA 02115; [b]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138; [c]D'Amore-McKim School of Business, Northeastern University, Boston, MA 02115; [d]The Percepts and Concepts Laboratory, Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405; and [e]Computational Social Science, Department of Humanities, Social and Political Sciences, Eidgenossische Technische Hochschule (ETH) Zürich, 8092 Zurich, Switzerland

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved June 1, 2016 (received for review March 4, 2016)

To investigate the effect of competitive incentives under peer review, we designed a novel experimental setup called the Art Exhibition Game. We present experimental evidence of how competition introduces both positive and negative effects when creative artifacts are evaluated and selected by peer review. Competition proved to be a double-edged sword: on the one hand, it fosters innovation and product diversity, but on the other hand, it also leads to more unfair reviews and to a lower level of agreement between reviewers. Moreover, an external validation of the quality of peer reviews during the laboratory experiment, based on 23,627 online evaluations on Amazon Mechanical Turk, shows that competition does not significantly increase the level of creativity. Furthermore, the higher rejection rate under competitive conditions does not improve the average quality of published contributions, because more high-quality work is also rejected. Overall, our results could explain why many ground-breaking studies in science end up in lower-tier journals. Differences and similarities between the Art Exhibition Game and scholarly peer review are discussed and the implications for the design of new incentive systems for scientists are explained.

peer review | competition | creativity | innovation | fairness

Competitive incentives are an essential tool to manipulate effort and performance of human groups in many real-life situations (1, 2). Sport tournaments with huge prizes, goal-contingent rewards for employees, and lavish end-of-career bonuses for corporate CEOs are a few examples. However, the literature on incentives and rewards offers mixed evidence of how effective competitive incentives are in improving individual performance (3). In particular, external (monetary) incentives might crowd out intrinsic motivation, which results in no effect, or even a negative effect on individual effort (4–6). Moreover, intrinsic motivation might not only mediate effort, but might actually be necessary to achieve creative performance (7). Similarly, competitive pressure can reduce the performance of professional athletes, causing them to "choke under pressure" (8, 9). Finally, when higher interests are at stake, competition can also directly lead to negative consequences, such as uncooperative behavior and even sabotage (10, 11).

In this paper, we test the effect of competition in a peer review system. Peer review is a self-regulating system where individuals with similar competence (peers) assess the quality of each other's work. Peer review is widely used by governmental agencies and health care professionals, and it is one of the cornerstones of science. Scholarly peer review is a truly complex system: it involves many actors engaged in multiple roles encompassing various feedback loops (12). Thus far, its inherent complexity and the restricted access to data have made it difficult to investigate peer review. Empirical studies have documented that the review process has low levels of inter-referee agreement (13, 14), lacks reliability (15), and might be prone to biases (16, 17). There is also an increasing number of studies based on computer simulations. These investigations have shown, for example, that peer review is susceptible to the influence of selfish motives and "gaming behavior" (18–20).

Here, we translate scholarly peer review into an artistic context by developing a novel experimental setup called the "Art Exhibition Game." To the best of our knowledge, this is the first laboratory experiment examining the effect of competition on peer review that allows for the study of the evolution of the behavior of both referees and creators simultaneously. The setup was designed to mimic some of the relevant characteristics of journal peer review in a minimalistic and well-controlled experimental paradigm: competition and cooperation, a high-dimensional space of possible creative products, different outlets for publication, and a sequence of rounds of production and evaluation with feedback. The game defines a situation in which stakeholders take the diversity, innovation, and intrinsic appeal of works into account, and face a fundamental choice between copying aspects of the solutions of their peers, or innovating on their own. The experiment abstracted from other features. An exact copy of scientific peer review was not intended.

The study was conducted in two steps. First, a simplified peer review system was created in the laboratory, whereby the degree of competition could be varied. Second, the results were externally validated through an online experiment using Amazon Mechanical Turk. Within the paradigm of the Art Exhibition Game, we were interested in testing the following questions:

i) Innovation: Does competition promote or reduce innovation? Tournaments and competitive incentives are known to increase effort and individual contributions (1, 2). However, creativity can be stifled by external factors, such as expectation of evaluation and monetary rewards (7, 21).

Significance

Competition is an essential mechanism in increasing the effort and performance of human groups in real life. However, competition has side effects: it can be detrimental to creativity and reduce cooperation. We conducted an experiment called the Art Exhibition Game to investigate the effect of competitive incentives in environments where the quality of creative products and the amount of innovation allowed are decided through peer review. Our approach is general and can provide insights in domains such as clinical evaluations, scientific admissibility, and science funding. Our results show that competition leads to more innovation but also to more unfair reviews and to a lower level of agreement between reviewers. Moreover, competition does not improve the average quality of published works.

Author contributions: S.B., R.L.G., and D.H. designed research; S.B. performed research; S.B. contributed new reagents/analytic tools; S.B. analyzed data; and S.B., R.L.G., and D.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: s.balietti@neu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1603723113/-/DCSupplemental.

ii) Fairness: Does competition reduce or improve the fairness of the reviews? Although it is known that competitive tournaments can discourage cooperation and, in the worst case, even lead to sabotage (10, 11), few studies have looked into how competition affects psychological biases.

iii) Validity: Does competition improve or hamper the ability of reviewers to identify valuable contributions? Scholarly peer review has been shown to have good predictive validity for journal articles, but less so for grant proposals (17, 22). However, no study seems to have tested the interaction between competition and validity.

Materials and Methods

Lab Experiment. We invited 144 individuals to the Decision Science Laboratory (DeSciL) at ETH Zurich. We conducted 16 sessions, each involving 9 participants, using the experimental software nodeGame (nodegame.org).

Participants were asked to produce a parametric drawing on the computer screen using an interactive interface comprising a number of individually configurable elements. The interface allowed participants to create modified versions of Chernoff faces (23). However, participants were neither asked, nor confined by the interface, to create images that resembled faces. On the contrary, they were free to experiment with different shapes and sizes, enabling the creation of other objects and even completely abstract art. In fact, the total number of combinations available amounted to the astronomical number of 5.2×10^{43} possible creative outcomes.

Participants were then asked to submit their finished artwork to one of three possible exhibitions. Each image was reviewed by three peers on a scale from 0 to 10, and each peer rated three images (self-assessment was not permitted). Artworks that received an average score above 5.0 were published in the exhibition of choice and displayed to all participants, and their creators were given a monetary reward. The game was repeated for 30 rounds, and the final earnings of each participant were proportional to the total number of artworks which they published. The duration of each round was limited. More time was allocated to the first two rounds. After round 1, participants also had the possibility to copy previously published images by simply clicking on them in the history of past exhibitions.

Participants were also assigned to three groups based on a feature that could not be controlled: color (green, black, red). This immutable feature was used to check for in-group or familiarity bias.

We studied two treatment conditions: (*i*) level of competition and (*ii*) reviewer choice. First, we defined a baseline, where participants received a fixed individual reward for each of their artworks published in any of the exhibitions. Under this condition, reviewers could potentially accept all of the pieces of art submitted, which is why we label it as noncompetitive (non-COM). Conversely, under competitive conditions (COM), a fixed overall reward was divided among all participants who managed to publish in the same exhibition in the same round. We also varied how the submissions were assigned to reviewers. Under the random condition (RND), reviewers were randomly assigned to submissions. Under another condition (CHOICE), participants who submitted and published artworks in one exhibition were more likely to be chosen as reviewers for that exhibition. In the analysis that follows, however, we will solely report the findings for the competitive and noncompetitive treatment conditions (COM vs. non-COM), as the RND and CHOICE conditions did not significantly affect our results (i.e., we do not distinguish them here). Further details are summarized in *SI Appendix*.

Online Experiment. To assess the validity of laboratory peer review and to categorize the images which participants created, we recruited 620 additional individuals from Amazon Mechanical Turk (https://www.mturk.com). Each online participant was asked to rate 40 randomly selected images. Each image was rated on a scale from 0 to 10 on the basis of four criteria: (*i*) creativity, (*ii*) abstractness, (*iii*) interestingness as a face, and (*iv*) overall appeal or quality. Evaluations that took less than 1 s or longer than 50 s were discarded, leading to a total of 23,627 external reviews on each dimension. Online evaluations are devoid of strategic motives and, as shown by previous research in acceptability judgments (24), quality standards measured online are approximately equivalent to those in the laboratory. Therefore, in our analysis, we will treat averaged online evaluations as an independent quality standard. The online evaluation was conducted with the experimental software nodeGame (nodegame.org) (further details in *SI Appendix*).

Statement of Research Conduct. Both laboratory and online experiments were approved by the ETH Zurich DeSciL Review Board and conducted in accordance with the DeSciL Operational Rules. Every person who has signed up to the DeSciL's subject pool also gave his or her informed consent by agreeing to the terms and conditions of Universitäre Anmeldestelle für Studienteilnehmer (UAST). These terms and conditions are published on the UAST website (https://www.uast.uzh.ch/register). Further information is provided in *SI Appendix*.

Scholarly Peer Review and the Art Exhibition Game. The peer review process carried out in the Art Exhibition Game shares many common features with scholarly peer review but also has some important differences, which are discussed below.

First, real-world scholarly peer review practices vary largely across fields, history, and journal competitiveness. We did not attempt to recreate in a laboratory context an experiment that would encompass all possible facets of journal peer review. Instead, we aimed to achieve a purposeful idealization that eliminates many of the "contaminating" factors that would otherwise obscure the operation of the interactive dynamics involved in the production and evaluation of creative products.

Second, in artistic peer review, the range of creative possibilities does not include any answer that is a priori "correct." However, this does not imply that "anything goes." In fact, the best combinations are decided through social interactions that eventually lead to the development of a "taste" or "standard" that participants use to evaluate the quality of creative products. Indeed, very similar images can score highly or poorly in different experimental sessions or in the same session but in different rounds. It is important to note that, unlike in previous studies of diversity in artificial cultural markets (25), the participants in the Art Exhibition Game are not only passive evaluators but also produce similar kinds of artwork themselves, meaning that they constantly adapt their creative output to the feedback they receive during the review process.

Third, scientists need more time to create and evaluate creative products, and much dedication of time and effort is required to foster the expertise needed to produce them. Moreover, the stakes (in terms of monetary gratification and implications for career and reputation) are much larger in scholarly peer review than in the Art Exhibition Game. However, all these simplifications are necessary to conduct a laboratory experiment that enables quantitative assessment of the similarity between solutions at varying levels of competition.

Finally, a critic might contend that scientific creativity and artistic creativity are two completely different processes. Although this remains an open question, there are many clues suggesting the opposite (26). Indeed, making a scientific contribution requires "imagination, intuition, synthesis, and a sense of aesthetics" (27). For instance, mathematical proofs and models are often not only evaluated for correctness, but also for their beauty, which often derives from their simplicity and parsimony. Moreover, neuroscientists have analyzed the neural basis for creativity and found that talented artists and scientists show similar patterns of brain activation in the association cortex and the socioaffective processing areas (28). Furthermore, several prominent creativity researchers have argued for a unified model of human creativity. For example, Teresa Amabile, who performed several laboratory and field experiments eliciting creative behavior of children and adults, proposes that "there is one basic form of creativity, one basic quality of products that observers are responding to when they call something 'creative,' whether they are working in science or the arts" (29, p. 32). Following another approach, Dean Keith Simonton interprets scientific creativity as a quasi-random, combinatorial process. He shows that quality can be estimated as a probabilistic function of quantity; the same Poisson-distributed pattern can be found in classical music, scientific publications and technological patents alike. Therefore, he concludes that "creativity must operate according to the same generic stochastic process in both the arts and the sciences" (30, 31). In the words of Nobel laureate Herbert Simon: "there is no reason to believe that the creative process in the arts is different from the creative process in the sciences" (32).

Results

In this section, we report the results of the laboratory and online experiments organized around the three research questions that we wanted to address.

Does Competition Promote or Reduce Innovation? In our experimental design, participants generate creative artifacts in the form of parametric images. This approach allowed us to uniquely map each created image onto a multidimensional parameter vector. In this way, we can quantitatively measure the difference between the images at both the individual and group level. To study

the creative behavior of participants, we introduce three measures: (*i*) innovation—the average Euclidean distance from the parameter vectors of images published in the previous round; (*ii*) diversity, i.e., diversity of products, or divergence—the average Euclidean distance between the parameter vectors of images submitted in the same round; and (*iii*) personal change—the Euclidean distance between the parameter vectors of two consecutive submissions by the same participant. The trend of these three measures over the 30 rounds of the experiment is illustrated in Fig. 1*A*. The levels of innovation and diversity increase markedly until the 20th round, after which the slope flattens but remains positive nevertheless. This result is striking because previous experiments in opinion formation have typically found that people tend to converge over rounds of social exchange (33–36). Our experiment finds the opposite to be true, due to the possibility of innovation. In fact, the Art Exhibition Game encourages increasingly diverse art because artworks are thought to be assessed for their creativity and novelty, as evidenced by the responses of participants to the questionnaire conducted after the experiment (*SI Appendix*).

When the global trend of innovation is disaggregated by the level of competition (Fig. 1 *B* and *C*), we observe that competitive conditions promote higher levels of diversity and innovation, about a 20% increase relative to noncompetitive conditions ($P < 0.001$). Given that the level of personal change is the same across conditions, these results suggest that, under competitive conditions, participants purposely aim to distinguish themselves from other peers in the same group. To check this, we computed the relative diversity of each image. This index is defined as the difference between (*i*) the average distance from images created in the same round in other sessions (between-diversity) and (*ii*) the average distance from images created in the same round in the same session (within-diversity). Values above zero indicate that forms of social influence, such as imitation, are at play. Artworks produced in this fashion are more similar to other artworks within the same group than they are to artworks produced in different groups. Values below zero suggest that social influence has a negative valence, meaning that participants are actively trying to differentiate themselves from others. Strikingly, the latter is the case for competitive conditions, where the relative diversity index is negative for all but the last four rounds (*SI Appendix*, Fig. S15); in noncompetitive conditions, conversely, relative diversity is always positive, meaning that a certain degree of social imitation is at play.

To get a richer picture of the behavior of the laboratory participants as creators, we now complement our analysis with the results of the online experiment on Amazon Mechanical Turk. As shown in Fig. 2, similarly to diversity and innovation, creativity also increases over time. Interestingly, there is a clear trend toward more abstract art as the experiment progresses. However, the two treatment conditions seem to incentivize different qualities: whereas noncompetitive conditions promote the creation of images resembling faces, competition fosters artistic abstractness and distinctiveness within a group. A Welch two-sample *t* test is significant for abstractness ($P < 0.001$), appeal as a face ($P = 0.001$), and overall score ($P < 0.001$). In contrast, neither treatment condition was found to have a significant effect on the level of creativity of the participants. Interestingly, noncompetitive artworks have a higher overall rating than competitive ones. However, this is related to the fact that participants under noncompetitive conditions produce more face-like images, which tend to have higher overall ratings than abstract art. These results hold throughout the whole experiment and there are no differences in convergence toward the final rounds.

Finally, we can learn more about the imitation patterns in our experiment by analyzing how participants across the two conditions made use of the "copy from past exhibition" feature (*SI Appendix*, Fig. S10). Participants in noncompetitive conditions copied significantly more images than in competitive conditions (χ^2, $\chi = 12.326$, $P < 0.001$). This finding is in line with our argument that participants in competitive conditions are purposely trying to differentiate themselves. However, if we look more closely into the type of images that are copied, participants under competitive conditions make a completely different use of the possibility to copy artwork. In fact, they overwhelmingly copy more of their own previous images, rather than images from other participants ($\chi = 34.75$, $P < 0.001$). As reported in the final questionnaire, copying was mainly used when they "wanted to change [. . .] drastically, to save some time."

In sum, competition causes participants to purposely be more innovative, more diverse and more abstract. This behavior is likely driven by the participants' belief that, under competitive conditions, diversity is the characteristic that is valued most by reviewers.

Does Competition Reduce or Improve the Fairness of the Reviews? A fair evaluation should not be affected by biases or strategic motives, but both may actually occur. Although a bias may be unintentional, a strategic evaluation refers to a case where the reviewer intentionally aims to derive personal benefit.

First, we examine the presence of potential biases in the reviews. The setup of the Art Exhibition Game allows us to scrutinize biases in the reviews not only ex-ante, but also ex-post (17). An ex-ante bias is found if any particular subgroup of creators consistently

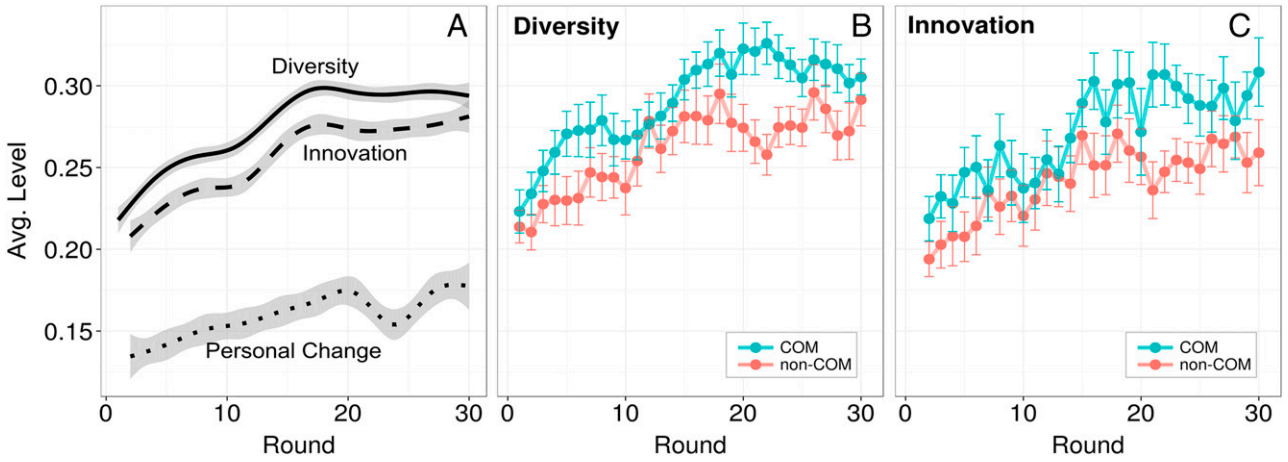

Fig. 1. Outcome measures for creative products. Diversity, innovation, and personal change increase over rounds (*A*). Competition (COM) fosters even higher levels of diversity (*B*) and innovation (*C*) than the noncompetitive condition (non-COM). Error bars show 95% CIs.

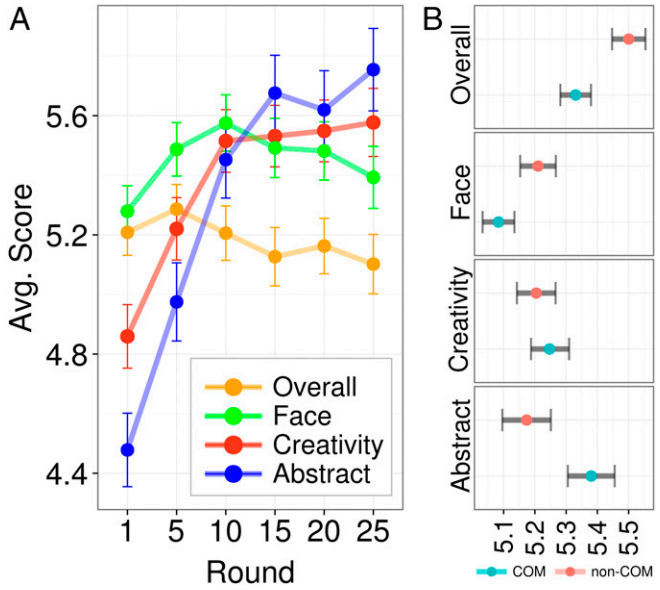

Fig. 2. External review scores of the creative products according to independent reviewers recruited through Amazon Mechanical Turk. Images become more creative and abstract over rounds (*A*), Under competitive conditions (COM), participants create more abstract art, whereas under noncompetitive conditions (non-COM), more images resembling faces are created (*B*). Error bars show 95% CIs.

receives a differential evaluation in their reviews. However, it might be that such a subgroup is actually performing differently, and, in such a case, a lower (higher) score should be considered an accurate evaluation and not a bias. In fact, only an ex-post analysis of the performance of a rated item that is independent of the initial evaluation can resolve the issue. We can perform such an ex-post analysis by using the evaluation scores of "overall appeal or quality" of the online reviewers on Amazon Mechanical Turk.

In our design, every participant is assigned an immutable feature, a color, that he or she cannot change throughout the experiment. According to social identity theory, even minimal differences such as colors may suffice to generate an in-group bias (37). We ran hierarchical regressions with session and subject as random effects to estimate the effect of the artwork color on ratings. Color proved to be insignificant ($P > 0.05$). Hence, our analysis shows that no ex-ante color bias exists for the two conditions: competitive and noncompetitive. Furthermore, ex-post analysis confirms that no color performs significantly better or worse than the others ($P > 0.05$). Therefore, we conclude that competition does not have a significant effect on psychological biases in our setup (more information in *SI Appendix*).

Let us now consider eventual strategic motivations in the reviews. As described in *Materials and Methods*, the reward for publication in the competitive environment is divided among all of the participants who published in the same exhibition. Therefore, participants who submit to the same exhibition in the same round are defined as "direct competitors." Fig. 3 shows the distribution of laboratory review scores disaggregated by the level of competition and by direct vs. nondirect competitors. The results show a dramatic difference across conditions (Kolmogorov–Smirnov test, $D = 0.3521$, $P < 0.001$). Noncompetitive sessions exhibit a roughly symmetrical distribution, which is approximately equivalent for both direct and nondirect competitors. Competitive sessions, however, exhibit a markedly different distribution for direct competitors. In fact, the distribution is no longer symmetrical, exhibiting a very pronounced peak for very low scores and very few high scores in comparison. This result suggests that participants under competitive pressure use their power as referees to unfairly reduce the review scores of their direct competitors. To capture this behavior, we define an extremely low review score as one which is less than 0.5 of 10. If an extremely low score is given to the direct competitor of a participant, we term this as an Asymmetric Strategic Selective (A.S.S.) review. We can then obtain the A.S.S. index for each reviewer by calculating the average fraction of A.S.S. reviews that he or she made throughout an experimental session. More formally, we define the A.S.S. index for reviewers as

$$A.S.S.(i) = \frac{1}{30}\sum_{r=1}^{30} A.S.S.(i,r),$$

that is, the average number of A.S.S. reviews assigned by reviewer i, given $n(r) = 0...3$ opportunities available throughout each of the $r = 1...30$ rounds of an experimental session. An A.S.S. review is in turn defined as

$$A.S.S.(i,r) = \begin{cases} \frac{1}{n(r)}\sum_{j=1}^{n(r)} A.S.S.(i,j), & \text{if } n(r) > 0 \\ 0, & \text{otherwise} \end{cases},$$

$$A.S.S.(i,j) = \begin{cases} 1 & \text{if } r(ij_d) < t \\ 0 & \text{if } r(ij_d) \geq t \end{cases},$$

where $r(ij_d)$ is the score that reviewer i assigns to an image of a direct competitor j_d, and t is a demarcation threshold set to 0.05 (0.5/10). We purposely chose a very low threshold to limit the rate of false positives.

As *SI Appendix*, Fig. S17 *A* and *B* shows, competitive sessions produce considerably more A.S.S. reviews (Wilcoxon rank sum test, $W = 1{,}400{,}126$, $P < 0.001$) and A.S.S. reviewers ($W = 3{,}024$, $P < 0.001$). Moreover, the number of A.S.S. reviews increases over time, pointing to a retribution cycle that takes place during the experiment (*SI Appendix*, Fig. S17*C*).

Furthermore, competition directly affects the level of consensus among referees (*SI Appendix*, Fig. S18). In fact, under competitive conditions, the level of agreement between reviewers decreases steadily with each round, such that the average level during the last three rounds was just 0.47. This value is consistent with the empirical level of 0.5 found in the literature (38, 39), but significantly lower than the level of consensus under noncompetitive conditions ($W = 13{,}602$, $P < 0.001$). Moreover, under competitive

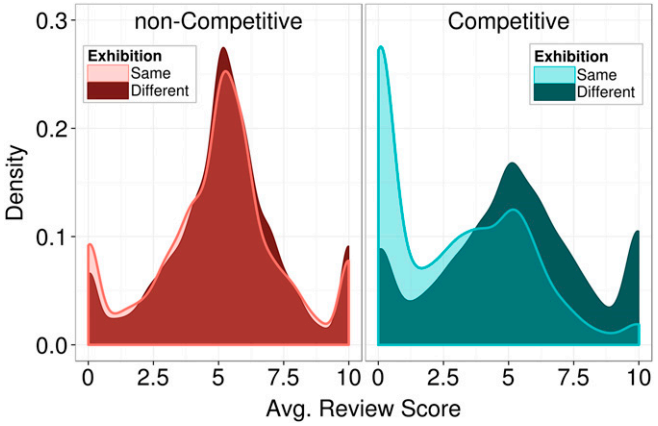

Fig. 3. Under competitive conditions, direct competitors in the same exhibition receive significantly lower review scores (close to zero), indicating self-interested behavior and gaming of the review system. Plots show 1D kernel estimate.

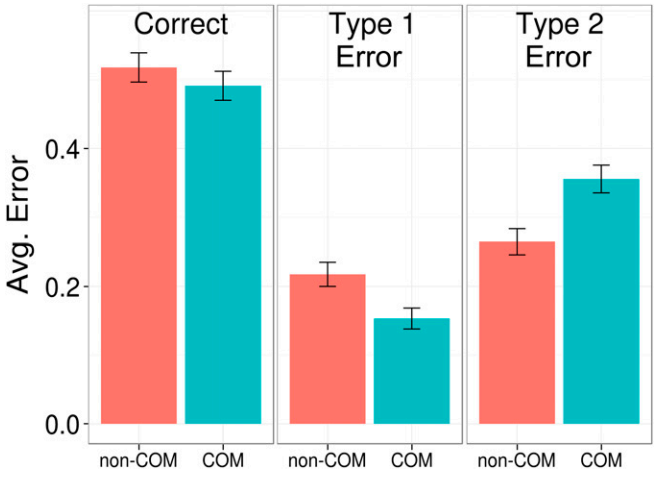

Fig. 4. Competition reduces type 1 errors, i.e., publication of low-quality items, but also introduces more type 2 errors, i.e., the rejection of high-quality items. Error bars show 95% CIs.

conditions, reviewers are even less consistent than predicted by a null model with shuffled reviews (*SI Appendix*, Fig. S20). This result provides further evidence that a substantial amount of "gaming of the review system" is taking place (40), to the point that the actual differences in the creative output of the participants can become less important than the "luck of the referee draw" (41).

In summary, our results show that peer review in the Art Exhibition Game does not show any ex-ante or ex-post biases. However, competition incentivizes reviewers to behave strategically, which reduces the fairness of evaluations and the consensus among referees.

Does Competition Improve or Hamper the Ability of Reviewers to Identify Valuable Contributions? To answer this question, we first analyze the rejection rates across the two conditions. Under noncompetitive conditions, the average review score was 5.21 ± 0.03, whereas with competition, it was only 4.30 ± 0.03. These numbers translate to an average rejection rate of 45% for noncompetitive conditions and of 65% for competitive conditions, just a little lower than actual rejection rates of top journals like *Nature*, *Science*, or PNAS. To understand whether the higher rejection rate led to an increase in the average quality of the published artworks, we performed a similar analysis as for fairness. We used the evaluation scores of overall appeal or quality of the independent Amazon Mechanical Turk reviewers and compared them to the outcomes of the peer reviews in the laboratory experiment. Using this method, we can compute the type I and type II error rates of our peer review experiment.

The results displayed in Fig. 4 show that, on average, both conditions perform about as well in selecting what to publish. However, noncompetitive conditions produced about 40% more type I errors, meaning that a greater amount of low-quality artwork (with average external evaluation ≤ 5.0) was published, whereas competition produced about 34% more type II errors, meaning that a greater amount of high-quality artwork (with average external evaluation > 5.0) was rejected.

Finally, we compared the average quality of published and rejected artwork under each condition to find out whether one condition would be able to discriminate high- from low-quality work better than the other. Our results show that even though published artworks are of higher quality than rejected artwork on average, the difference is not significant, with a thin spread of about 0.08 ± 0.03 for both conditions.

In summary, competition leads to higher rejection rates, but not to a higher average quality of published artworks. In fact, competition increases the rejection rates of both low-quality and high-quality contributions.

Discussion

We designed a novel experimental setup called the "Art Exhibition Game" to investigate the effects of competitive incentives under peer review. Our setup is unique because it allows us to simultaneously study both the evolution of creative output by authors and the behavior of reviewers. Our results clearly indicate that competition is a double-edged sword: whereas on the one hand, competition fosters innovation and diversity of products, on the other hand, it also leads to more unfair reviews and to a lower level of agreement between reviewers. Furthermore, competition does not improve (nor worsen) the validity of the outcomes of peer review. In fact, under competition the rejection rate increases by 20%, meaning that both more low-quality and more high-quality work is eliminated. Whether this is good or not is a difficult question. On the one hand, it depends on the philosophical standpoint taken, e.g., whether one is an elitist or populist, a follower of Popper or Feyerabend. On the other hand, it also depends on the nature of the problem at study and on the type of solution that is sought. In fact, it is known that when more solutions are admitted, participants in tournaments tend to lower their effort (42); this is consistent with our finding that participants in noncompetitive conditions tend to be less innovative on average. However, at the same time, the likelihood of finding profound or radical solutions also increases due to "parallel search paths" (42). If we look at the top 10 highest rated images (according to the external Amazon Mechanical Turk reviewers), we find that participants under noncompetitive conditions created 7 of the 10 most appealing images, 7 of the 10 most beautiful faces, 5 of the 10 most creative images, and 5 of the 10 most abstract images. These numbers are remarkable given the low level of innovation under noncompetitive conditions.

Our approach to the study of peer review is focused on abstracted art exhibitions. However, as previously explained in the main text and in *Materials and Methods*, we tried to mimic some of the essential features involved in scholarly journal peer review, so that some of our results may be transferable. Nowadays, many are concerned that the competitive pressure to "publish or perish" in academia is so high that it has distorted the incentives for scientists (43–45). Our results show that, even if outcomes of peer review are generally valid, competition increases editorial type II errors and encourages self-interested referees to behave strategically. Our results are consistent with other empirical studies that found that the most competitive journals in the field of medical sciences failed to accept some of the most cited articles, which later appeared in lower-tier journals (46). To identify mechanisms to mitigate the waste or delay of potential innovation in competitive peer review systems, additional research is needed. Future inquiry should focus on assessing the importance of the differences between our experimental setup and scholarly peer review; for example, variables such as timing, the size of the stakes at risk, and reputation could play an important role. Moreover, future investigation could also explore whether competition has a different effect on scientific teams, which are increasingly becoming the basic unit of research in most disciplines (47).

In conclusion, our work provides evidence of how competition can shape the incentives of both creators and reviewers involved in a peer review system, thereby altering its outcomes. In times where science is increasingly witnessing peer review rings, scientific fraud, and plagiarism (40, 45, 48, 49), the results of our study suggest a redesign the scientific incentive system such that sustainable forms of competition are promoted. For example, career schemes that tolerate early failure and focus on long-term success (50) could be the best way to guarantee high levels of responsible innovation.

ACKNOWLEDGMENTS. We thank M. Mäs, T. Kuhn, A. C. Balietti, Chris Riedl, and the members of the (European) Cooperation in Science and Technology Action TD1306 New Frontiers of Peer Review for useful discussion and comments. S.B. and D.H. acknowledge support by the European Commission through European Research Council Advanced Investigator Grant "Momentum" (324247).

1. Bull C, Schotter A, Weigelt K (1987) Tournaments and piece rates: An experimental study. *J Polit Econ* 95(1):1–33.
2. Bonner SE, Sprinkle G (2002) The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Account Organ Soc* 27(4):303–345.
3. Byron K, Khazanchi S (2012) Rewards and creative performance: A meta-analytic test of theoretically derived hypotheses. *Psychol Bull* 138(4):809–830.
4. Gneezy U, Rustichini A (2000) A fine is a price. *J Legal Stud* 29(1):1–17.
5. Gagné M, Deci E (2005) Self-determination theory and work motivation. *J Organ Behav* 26(4):331–362.
6. Dube J, Luo X, Fang Z (2015) Self-signalling and prosocial behavior: A cause marketing mobile field experiment. NBER working paper 21475. Available at www.nber.org/papers/w21475.
7. Amabile T (1998) How to kill creativity. *HBS Rev* Sep–Oct:77–87.
8. Baumeister R, Showers C (1986) A review of paradoxical performance effects: Choking under pressure in sports and mental tests. *Eur J Soc Psychol* 16(4):361–383.
9. Beilock SL, Carr TH (2001) On the fragility of skilled performance: What governs choking under pressure? *J Exp Psychol Gen* 130(4):701–725.
10. Harbring C, Irlenbusch B (2008) How many winners are good to have?: On tournaments with sabotage. *J Econ Behav Organ* 65(3):682–702.
11. Gürtler O, Münster J (2010) Sabotage in dynamic tournaments. *J Math Econ* 46(2):179–190.
12. Helbing D (2007) *Managing Complexity: Insights, Concepts, Applications* (Springer, New York).
13. Scott W (1974) Interreferee agreement on some characteristics of manuscripts submitted to the journal of personality and social psychology. *Am Psychol* 29(9):698–702.
14. Fiske D, Fogg L (1990) But the reviewers are making different criticisms of my paper! diversity and uniqueness in reviewer comments. *Am Psychol* 45(5):591–598.
15. Rothwell PM, Martyn CN (2000) Reproducibility of peer review in clinical neuroscience. Is agreement between reviewers any greater than would be expected by chance alone? *Brain* 123(Pt 9):1964–1969.
16. Mahoney M (1977) Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognit Ther Res* 1(2):161–175.
17. Mutz R, Bornmann L, Daniel HD (2015) Testing for the fairness and predictive validity of research funding decisions: A multilevel multiple imputation for missing data approach using ex-ante and ex-post peer evaluation data from the austrian science fund. *J Assoc Inf Sci Technol* 66(11):2321–2339.
18. Thurner S, Hanel R (2011) Peer-review in a world with rational scientists: Toward selection of the average. *Eur Phys J B* 84(4):707–711.
19. Squazzoni F, Gandelli C (2012) Saint matthew strikes again: An agent-based model of peer review and the scientific community structure. *J Informetrics* 6(2):265–275.
20. Thorngate W, Chowdhury W (2014) *Advances in Social Simulation* (Springer, New York), pp 177–188.
21. Amabile TM, Hennessey BA, Grossman BS (1986) Social influences on creativity: the effects of contracted-for reward. *J Pers Soc Psychol* 50(1):14–23.
22. Bornmann L, Mutz R, Marx W, Schier H, Daniel HD (2011) A multilevel modelling approach to investigating the predictive validity of editorial decisions: Do the editors of a high profile journal select manuscripts that are highly cited after publication? *J R Stat Soc Ser A Stat Soc* 174(4):857–879.
23. Chernoff H (1973) The use of faces to represent points in K-dimensional space graphically. *J Am Stat Assoc* 68(342):361–368.
24. Sprouse J (2011) A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behav Res Methods* 43(1):155–167.
25. Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854–856.
26. Sawyer R (2011) *Explaining Creativity: The Science of Human Innovation* (Oxford Univ Press, Oxford, UK).
27. Fang F, Casadevall A (2015) Competitive science: Is competition ruining science? *Infect Immun* 83(4):1229–1233.
28. Andreasen NC, Ramchandran K (2012) Creativity in art and science: Are there two cultures? *Dialogues Clin Neurosci* 14(1):49–54.
29. Amabile T (1983) *The Social Psychology of Creativity* (Springer-Verlag, Berlin).
30. Simonton D (1999) Creativity as blind variation and selective retention: Is the creative process darwinian? *Psychol Inq* 10(4):309–328.
31. Simonton DK (2009) Varieties of (scientific) creativity a hierarchical model of domain-specific disposition, development, and achievement. *Perspect Psychol Sci* 4(5):441–452.
32. Simon H (2001) Creativity in the arts and the sciences. *Kenyon Review* 23(2):203–220.
33. Latane B, Bourgeois M (2001) Handbook of social psychology. *Group Processes*, eds Tindale R, Hogg M (Blackwell, Malden, MA), Vol 4, pp 235–258.
34. Kenrick DT, Li NP, Butner J (2003) Dynamical evolutionary psychology: Individual decision rules and emergent social norms. *Psychol Rev* 110(1):3–28.
35. Mason WA, Conrey FR, Smith ER (2007) Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Pers Soc Psychol Rev* 11(3):279–300.
36. Lorenz J, Rauhut H, Schweitzer F, Helbing D (2011) How social influence can undermine the wisdom of crowd effect. *Proc Natl Acad Sci USA* 108(22):9020–9025.
37. Tajfel H, Turner J (1986) *Psychology of Intergroup Relations*, eds Worchel S, Austin L (Nelson-Hall, Chicago), pp 7–24.
38. Petty R, Fleming MA, Fabrigar L (1999) The review process at pspb: Correlates of interreviewer agreement and manuscript acceptance. *Pers Soc Psychol Bull* 25(2):188–203.
39. Thorngate W, Dawes R, Foddy M (2011) *Judging Merit* (Psychology Press, New York).
40. Ferguson C, Marcus A, Oransky I (2014) Publishing: The peer-review scam. *Nature* 515(7528):480–482.
41. Bornmann L, Daniel HD (2009) The luck of the referee draw: The effect of exchanging reviews. *Learn Publ* 22(2):117–125.
42. Boudreau K, Lacetera N, Lakhani K (2011) Incentives and problem uncertainty in innovation contests: An empirical analysis. *Manage Sci* 57(5):843–863.
43. Alberts B, Kirschner MW, Tilghman S, Varmus H (2014) Rescuing US biomedical research from its systemic flaws. *Proc Natl Acad Sci USA* 111(16):5773–5777.
44. Schekman R (2013) How journals like Nature, Cell and Science are damaging science. Available at www.theguardian.com/commentisfree/2013/dec/09/how-journals-nature-science-cell-damage-science. Accessed 15 April 2016.
45. Anderson MS, Ronning EA, De Vries R, Martinson BC (2007) The perverse effects of competition on scientists' work and relationships. *Sci Eng Ethics* 13(4):437–461.
46. Siler K, Lee K, Bero L (2015) Measuring the effectiveness of scientific gatekeeping. *Proc Natl Acad Sci USA* 112(2):360–365.
47. Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. *Science* 316(5827):1036–1039.
48. Fang FC, Steen RG, Casadevall A (2012) Misconduct accounts for the majority of retracted scientific publications. *Proc Natl Acad Sci USA* 109(42):17028–17033.
49. Butler D (2010) Journals step up plagiarism policing. *Nature* 466(7303):167.
50. Azoulay P, Graff Zivin J, Manso G (2011) Incentives and creativity: Evidence from the academic life sciences. *Rand J Econ* 42(3):527–554.