Supplementary Info for: “Peer Review and Competition in the Art Exhibition Game”

Stefano Balietti[1,2,3], Rob Goldstone[4] and Dirk Helbing[5]

1 Northeastern University, Network Science Institute
2 Harvard Institute for Quantitative Social Science
3 D’Amore-McKim School of Business
4 Indiana University, Department of Psychological and Brain Sciences
5 ETH Zurich, Computational Social Science
∗ **E-mail**: s.balietti@neu.edu

Detailed description of the lab experiment

The experiment was divided into four stages. Stage 3 was subdivided into 4 steps which were repeated for 30 rounds each. The experiment is schematically illustrated in Fig. 1. In total, the experiment lasted approximately 1 hour and 15 minutes.

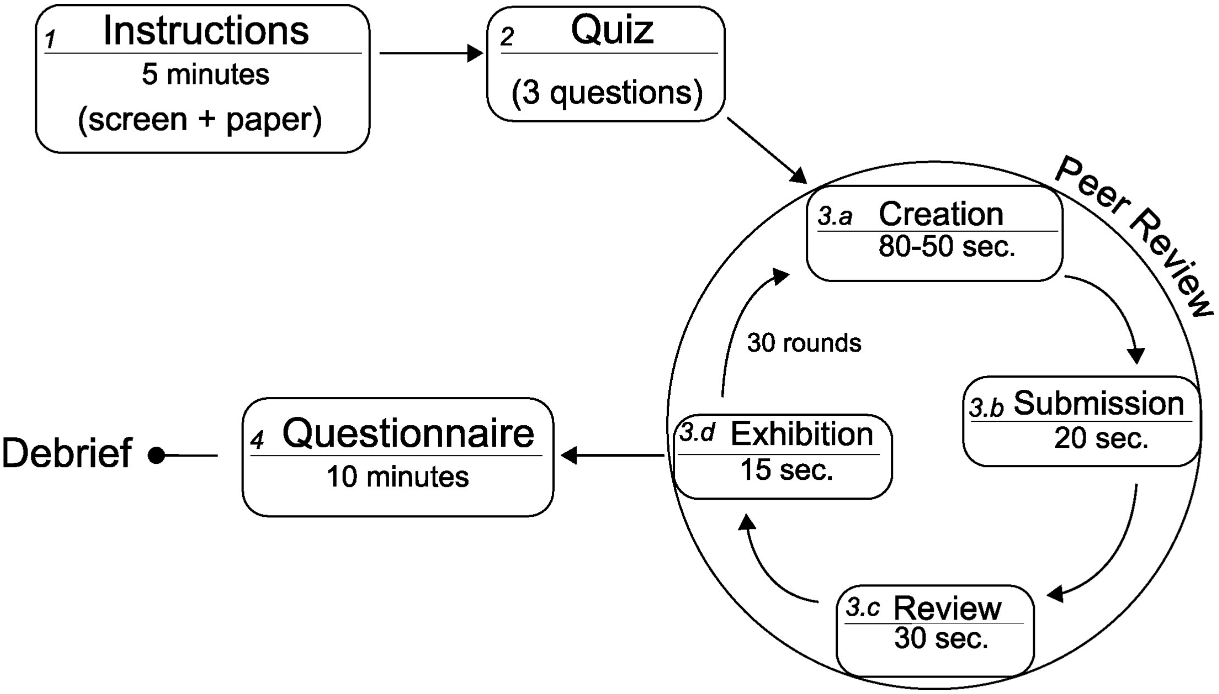


Figure 1. Game workflow.

Each experimental session began with 9 participants, who were randomly selected from a pool of 12 subjects; the chosen participants were then randomly assigned to a computer. Upon arrival at their computer each participant found a written sheet of instructions on the table detailing the game rules for the condition to which they had been assigned. The text of the instructions is reported below.[1]

[1] Headings marked with an asterisk were displayed only for the respective treatment condition.

Introduction

- You just became part of a community of 9 painters (including you). The community will collectively decide which paintings can be displayed in 3 (three) artistic exhibitions (named A, B, and C). All the three exhibitions have the same value for the members of the community.
- Always keep in mind: In this game you will act both as an artist and as an art critic.

The game

- The game is divided in rounds, and each round is divided in three steps:
- Step 1: Each artist creates 1 (one) painting, and chooses 1 (one) exhibition to which he / she submits the work of art.
- Step 2: Each artist reviews 3 (three) of the 8 (eight) paintings produced by the other artists. The review is done by assigning a rating on a scale from 0 (zero) to 10 (ten).
- Step 3: All the paintings that received an average review-score greater than 5 (five) are put on display in the exhibition to which they were submitted. If your painting is displayed you will receive a payoff.

Who reviews what* [CHOICE]

- When you choose one exhibition, you increase your chances to become a reviewer for that exhibition in the next round.
- If your painting is also published, your chances will be even higher.
- You will always review 3 paintings, and you can never review your own painting.

Who reviews what* [RANDOM]

- At each round you will be randomly assigned to review 3 paintings, and you can never review your own painting.

Payoff calculation* [COM]

- You will receive a fixed monetary compensation of 10 (ten) CHF, plus a variable amount based on your performance in the game.
- If one of your painting is published you will receive a sum equal to (3 / N) CHF, where N is the number of artists who published with you in the same exhibition.

N	1	2	3	4	5	6	7	8	9
Payoff	3	1.5	1	0.75	0.60	0.5	0.43	0.37	0.33

- If your painting is not published you will receive 0 (zero) CHF.
- The maximum amount of money you can win is 30*3 + 10 = 100 (hundred) CHF.

Payoff calculation* [non-COM]

- You will receive a fixed monetary compensation of 10 (ten) CHF, plus a variable amount based on your performance in the game.
- For each painting of yours that is displayed in one of the exhibitions you will be paid 2 (two) CHF.
- For each painting of yours that is not put on display you will receive 0 (zero) CHF.
- The maximum amount of money you can win is 30*2 + 10 = 70 (seventy) CHF.

Other rules

- You have a limited amount of time to complete each step of the game. The time left will be displayed on the screen.
- At each new round, when creating a new painting, you will continue from your previous submission.
- It is possible to copy paintings from the past exhibitions.

Termination

- The game will end either after 30 (thirty) rounds, or after 1 hour of playing.
- In case of misconduct, cheating or disturbing behavior during the game, you will not be paid any money.

After everybody was seated and welcomed with a standardized message, the computer screens were turned on and displayed four large images which illustrated the graphical interface used during the experiment (see Fig. 2). From that moment on, participants had 5 minutes to study the instructions before proceeding to the quiz stage.

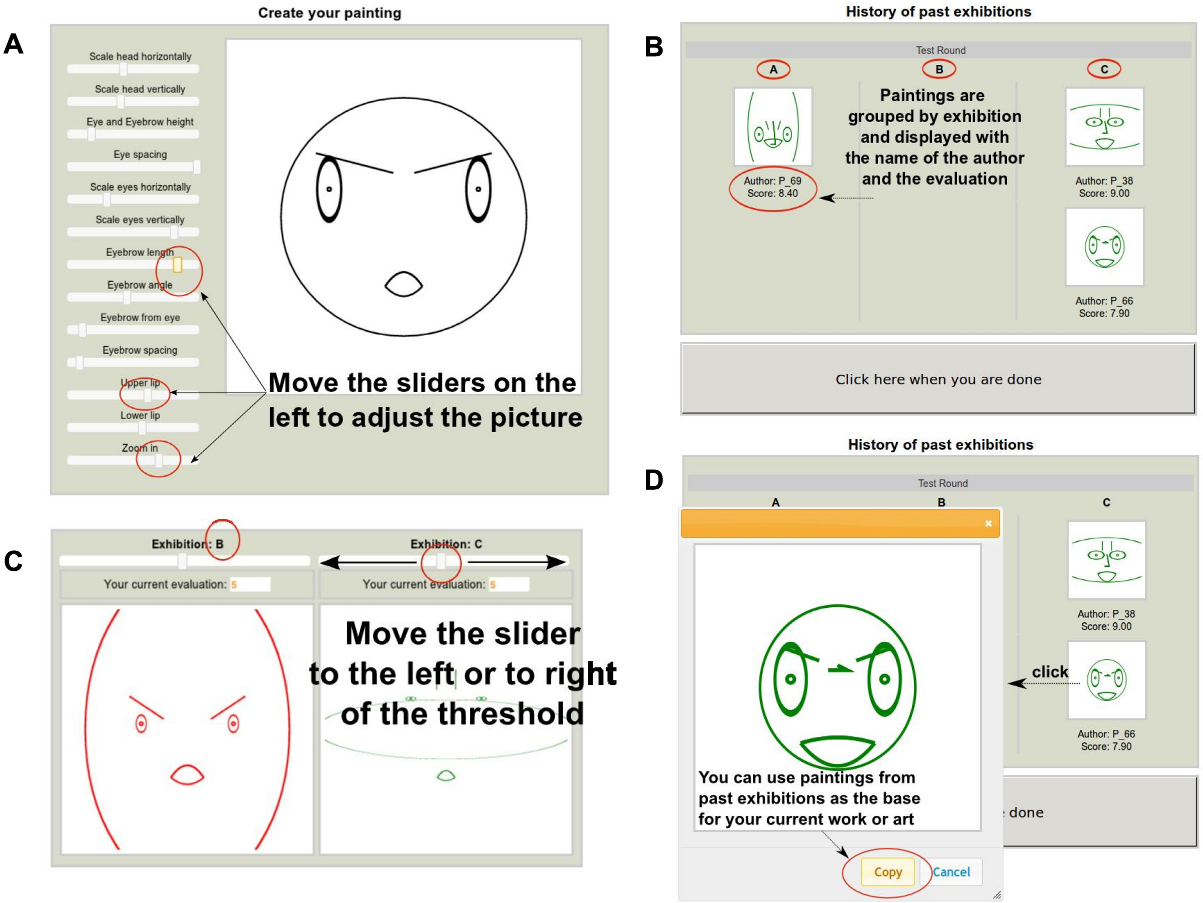
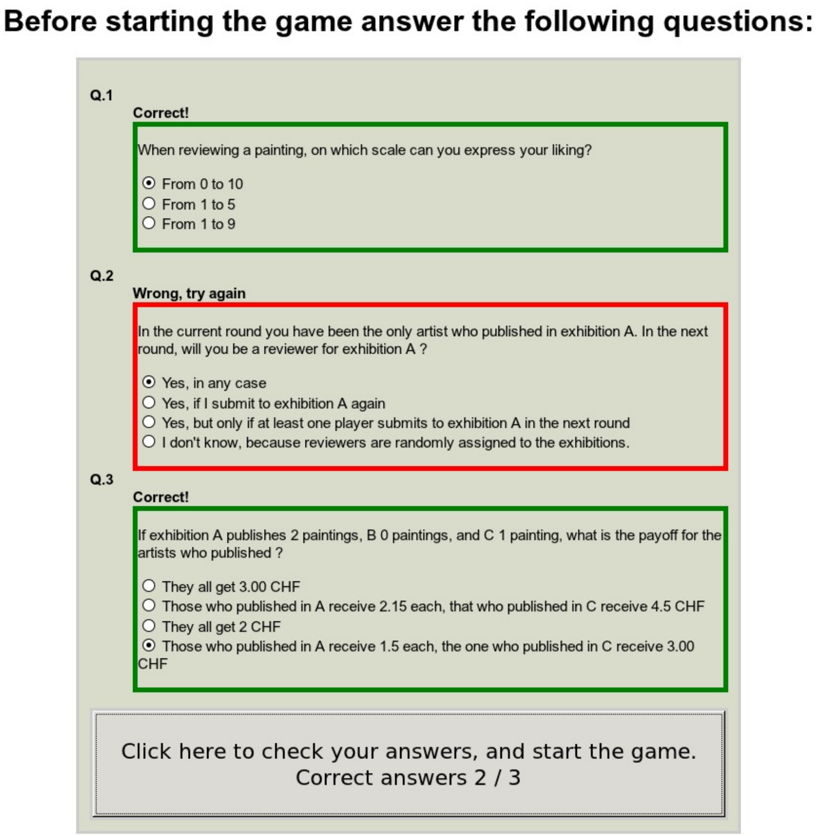

A
Create your painting
Scale head horizontally
Scale head vertically
Eye and Eyebrow height
Eye spacing
Scale eyes horizontally
Scale eyes vertically
Eyebrow length
Eyebrow angle
Eyebrow from eye
Eyebrow spacing
Upper lip
Lower lip
Zoom in
Move the sliders on the left to adjust the picture
B
History of past exhibitions
Test Round
A
B
C
Paintings are grouped by exhibition and displayed with the name of the author and the evaluation
Author: P_69
Score: 8.40
Author: P_38
Score: 9.00
Author: P_66
Score: 7.90
Click here when you are done
C
Exhibition: B
Exhibition: C
Your current evaluation: 5
Your current evaluation: 5
Move the slider to the left or to right of the threshold
D
History of past exhibitions
Test Round
A
B
C
Author: P_38
Score: 9.00
click
Author: P_66
Score: 7.90
You can use paintings from past exhibitions as the base for your current work or art
done
Copy
Cancel

Figure 2. Illustrative images displayed during the instruction stage. The four large images were shown sequentially in the order A,B,D,C. Each image highlighted one important feature of the graphical interface used during the experiment.

The quiz stage consisted of three multiple-choice questions, as shown in Fig. 3. The first question, which was the same for every treatment condition, concerned the scale used to rate images. The remaining two questions, which were specific to the treatment condition, covered the way reviewers are assigned, and how the monetary payoff for publication is calculated. Participants were allowed to answer each question multiple times, and visual feedback informed them whether their guess was correct or not. The average number of attempted answers to the quiz questions is shown in Fig. 4. No significant difference was found between competitive and non-competitive conditions. The question concerning reviewer assignment had significantly more incorrect answers under the CHOICE condition, although the size of the effect was very small. The quiz lasted 4 minutes on average.

After the quiz, the Peer Review Loop – Stage 3 in Fig. 1 – began. Firstly, participants were presented with the "creation" interface consisting of 13 sliders, each of which manipulated a particular feature of a



Before starting the game answer the following questions:

Q.1

Correct!

When reviewing a painting, on which scale can you express your liking?

● From 0 to 10
○ From 1 to 5
○ From 1 to 9

Q.2

Wrong, try again

In the current round you have been the only artist who published in exhibition A. In the next round, will you be a reviewer for exhibition A ?

● Yes, in any case
○ Yes, if I submit to exhibition A again
○ Yes, but only if at least one player submits to exhibition A in the next round
○ I don't know, because reviewers are randomly assigned to the exhibitions.

Q.3

Correct!

If exhibition A publishes 2 paintings, B 0 paintings, and C 1 painting, what is the payoff for the artists who published ?

○ They all get 3.00 CHF
○ Those who published in A receive 2.15 each, that who published in C receive 4.5 CHF
○ They all get 2 CHF
● Those who published in A receive 1.5 each, the one who published in C receive 3.00 CHF

Figure 3. Illustration of the Quiz interface. Three questions needed to be answered correctly by all participants before beginning the Peer Review stage.

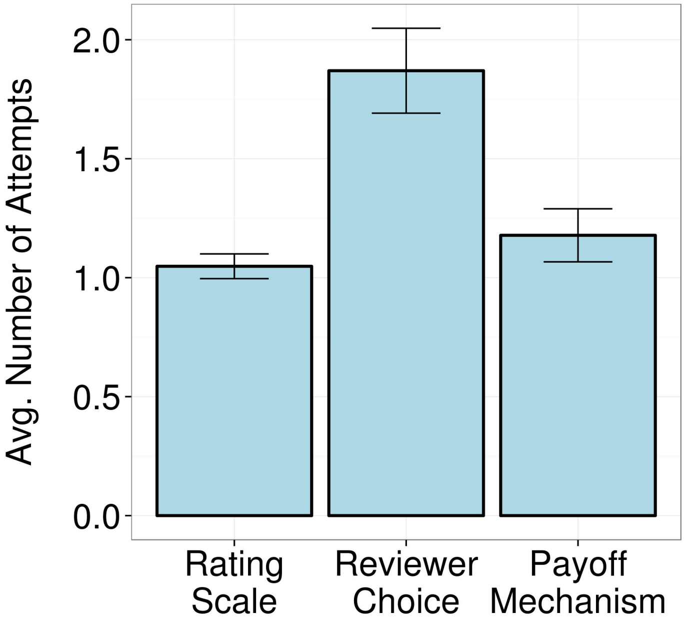


Figure 4. Average number of attempts to answer the questions of the quiz correctly.

modified Chernoff Face [1]. Table 1 illustrates all of the parameters available to manipulate the modified Chernoff Face, and the range of each parameter. The initial position of each slider was random. The sliders were named according to the characteristic they controlled, and were ordered as follows: *Scale head horizontally, Scale head vertically, Eye and Eyebrow height, Eye spacing, Scale eyes horizontally, Scale eyes vertically, Eyebrow length, Eyebrow angle, Eyebrow from eye, Eyebrow spacing, Upper lip, Lower lip, Zoom in.* Fig. 5 shows the complete creation interface. Even though at first sight the interface might appear complex because it encompasses many sliders, participants reported to have no major difficulties using it in the final questionnaire.

A sample of the images produced by participants during the experiment is shown in Fig. 6. Readers who would like to try the same interface used in the experiment can do so at this link: `http://nodegame.org/games/artex/chernoff`.

In addition to the features which can be modified through the "creation" interface, each computer was assigned two fixed and unalterable features: (i) one color ("red", "green", "black"), and (ii) a fictitious name of an American painter. The color was applied to the lines of the modified Chernoff Face, and it was shared by groups of three peers. It was introduced to test whether participants would use it as a means of group identification [9] to discriminate against out-group members in the review process. As shown in the supplementary data analysis section, the data provides limited support for this hypothesis. The fictitious name of an American painter was used to strengthen the participants' perception of being involved in an artistic context. To this extent, unfamiliar artist names, not associated with any positive or negative valence, were chosen. The complete list of fixed features which were associated with each computer is shown in Table 2.

To create a new image, the participants had 80 seconds in the first round, 60 seconds in the second round, and 50 seconds in all of the remaining rounds. The extra time in the first two rounds was intended

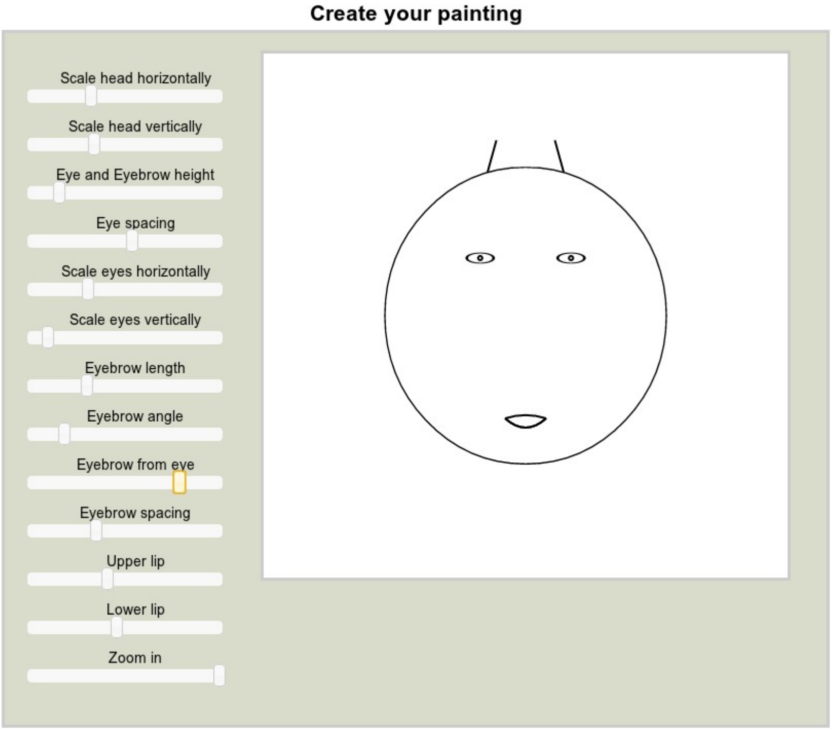


Figure 5. Illustration of the creation interface. 13 Sliders controlled the features of a modified Chernoff Face. As a slider moved, the image was updated in real time, giving immediate feedback to participants.

Table 1. Chernoff Face features used to create the parametric images. Some features were locked to a default value, and not displayed at all in the interface.

n.	Feature	Min	Max	Step	Default	Locked
1	Scale head horizontally	0.001	2	0.001	1	No
2	Scale head vertically	0.001	2	0.001	0.4	No
3	Eye and Eyebrow height	0	2	0.01	0.4	No
4	Eye spacing	0	40	0.01	10	No
5	Scale eyes horizontally	0.01	4	0.01	1	No
6	Scale eyes vertically	0.01	4	0.01	1	No
7	Eyebrow length	0	50	0.01	10	No
8	Eyebrow angle	-3.14	3.14	0.01	-0.5	No
9	Eyebrow from eye	0	50	0.01	3	No
10	Eyebrow spacing	0	50	0.01	5	No
11	Upper lip	-60	60	0.01	-2	No
12	Lower lip	-60	60	0.01	20	No
13	Zoom in	10	100	0.01	30	No
14	Pupil radius	1	9	0.01	1	Yes
15	Scale pupils horizontally	0.2	2	0.01	1	Yes
16	Scale pupils vertically	0.2	2	0.01	1	Yes
17	Mouth height	0.2	2	0.01	0.75	Yes
18	Mouth width	2	100	0.01	20	Yes
19	Nose height	0.4	1	0.01	0.4	Yes
20	Nose length	0.2	30	0.01	15	Yes
21	Nose width	0	30	0.01	10	Yes
22	Eye radius	0	30	0.01	10	Yes

Table 2. Immutable features associated with a computer position.

Position	Color	Fictitious Name
2	Green	Johnson
3	Green	Jackson
4	Green	Kerry
5	Red	Michealson
6	Red	McCotton
7	Red	Bradbury
8	Black	Howard
9	Black	Walbright
10	Black	Chestner

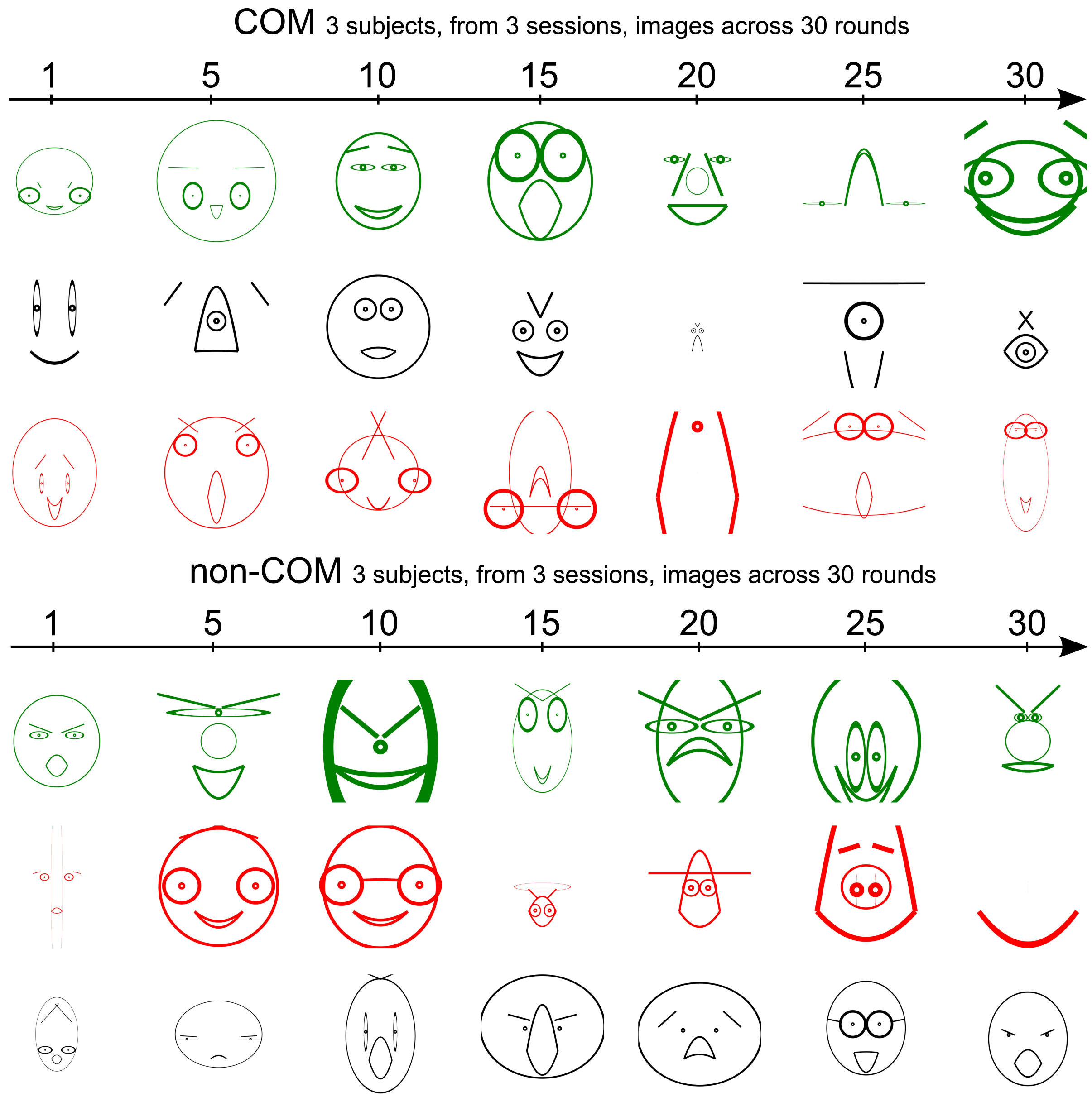
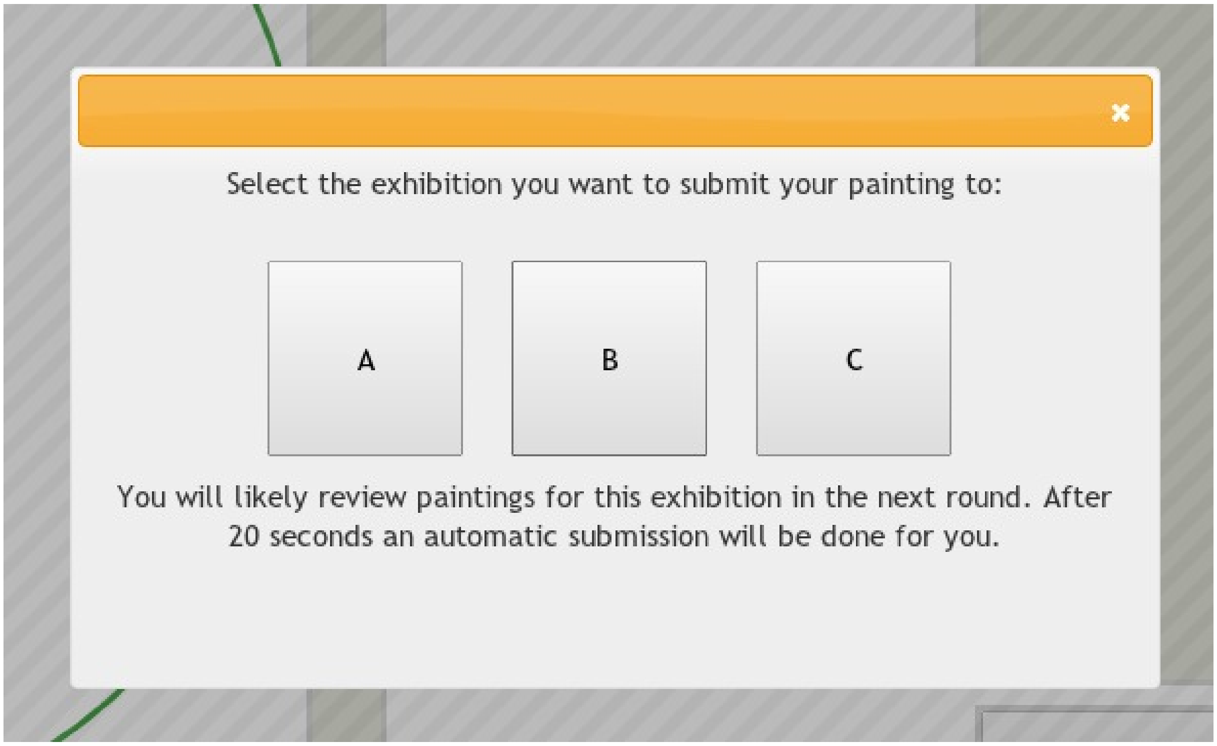

COM 3 subjects, from 3 sessions, images across 30 rounds
1
5
10
15
20
25
30
non-COM 3 subjects, from 3 sessions, images across 30 rounds
1
5
10
15
20
25
30

Figure 6. Sample of images created during the experiment. Each row shows the evolution of the images produced by a different participant across the whole experiment. Images from rounds 1,5,10,15,20,25,30 are displayed. Each participant belonged to a different session.

to help participants familiarize themselves with the interface.

After creating an image, or when the time had expired, the submission step would begin. A new pop-up window enabled the participant to select 1 of 3 exhibitions ("A","B","C") as shown in Fig. 7. If no decision was taken after 20 seconds, the system would submit the image automatically. In the first round, the exhibition was chosen randomly, and in all subsequent rounds, the previous choice was maintained. We flagged all cases of automatic submission.



Select the exhibition you want to submit your painting to:

You will likely review paintings for this exhibition in the next round. After 20 seconds an automatic submission will be done for you.

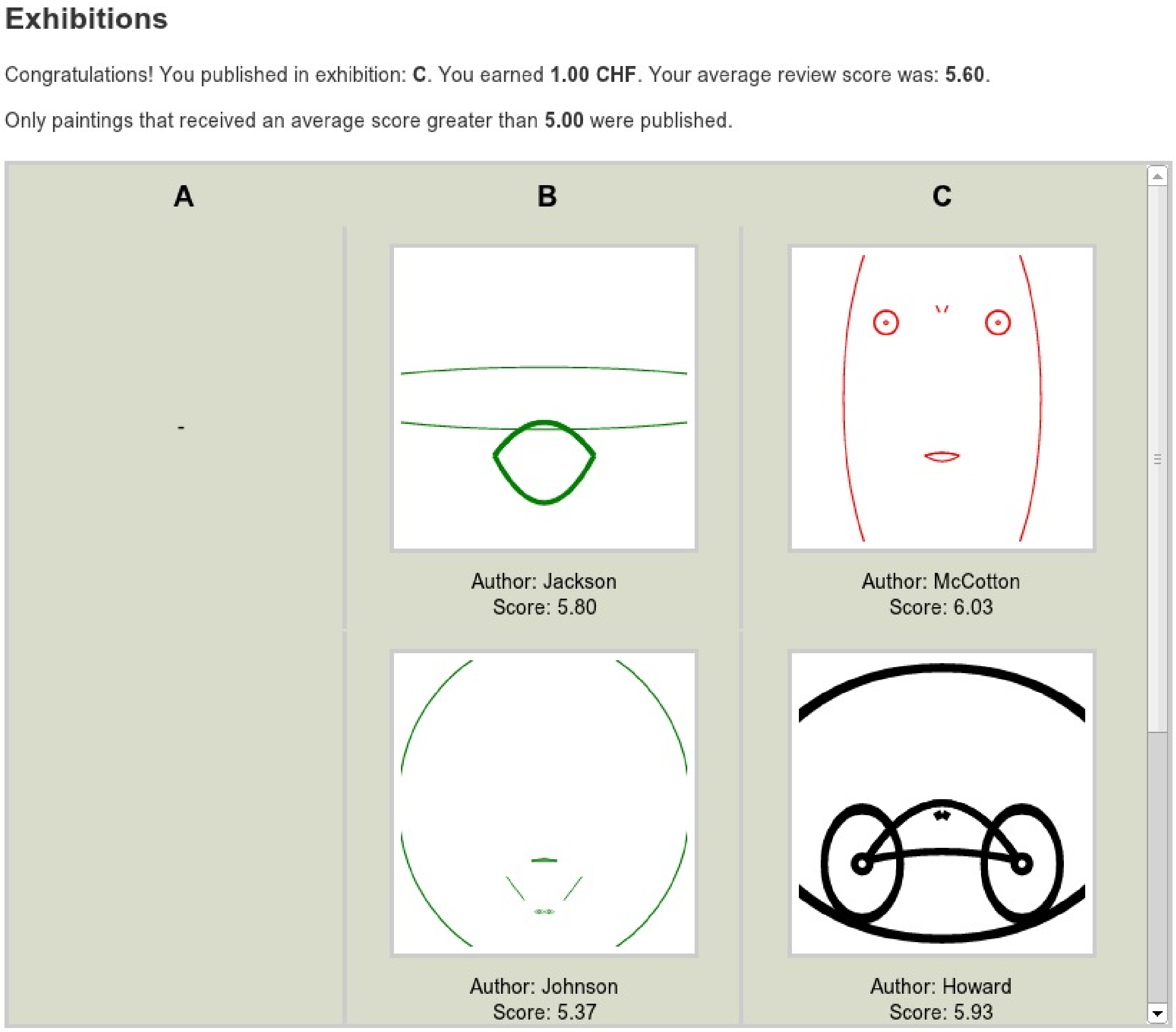
Figure 7. Illustration of the Submission interface. Image taken from the CHOICE condition.

Next, the "evaluation" stage would begin. Each individual was asked to evaluate the images produced by the other participants. Each participant evaluated three images, and each image was evaluated by three participants. The three images were displayed simultaneously, together with the name of the exhibition to which they have been submitted, but without the name of the author. The whole procedure aimed to reproduce a form of double-blind peer review. In their evaluations, participants were asked to *"Express a quality judgement on the paintings just made by other players"* and to *"Move the slider on a scale from 0 to 10, where 0 (zero) means complete dislike, 5 (five) means that you are indifferent, 10 (ten) means complete like."* All review sliders were initially set in the middle of their range (5.0), and increments (or decrements) of as little as 0.1 were possible. We flagged any sliders which were not moved at all. The overall time allocated for reviewing three images was 20 seconds, or less if all participants clicked the "continue" button before the time expired.

After all the reviews were collected, the mean review score for each image was computed. All images with an average score above 5.0 were considered "accepted for publication," i.e. they were displayed in the exhibition to which they were submitted.[2] Each exhibition was displayed in a separate column, and

[2] Different decision rules for aggregating reviewers' judgments could have been used, however it has been empirically shown that the mean recommendation value across reviewers is the best predictor for final acceptance decisions by editors in scholarly peer review [3].

images were sorted by average rating, from highest to lowest. The name of the creator and the score which he/she had received were visible under the picture. Finally, above the exhibition panel, participants were also informed about their payoff for the round, as shown in Fig. 8. The exhibition was displayed for 15 seconds, or less if all participants clicked on the "continue" button before the time expired.



Exhibitions

Congratulations! You published in exhibition: **C**. You earned **1.00 CHF**. Your average review score was: **5.60**.

Only paintings that received an average score greater than **5.00** were published.

Author: Jackson
Score: 5.80

Author: McCotton
Score: 6.03

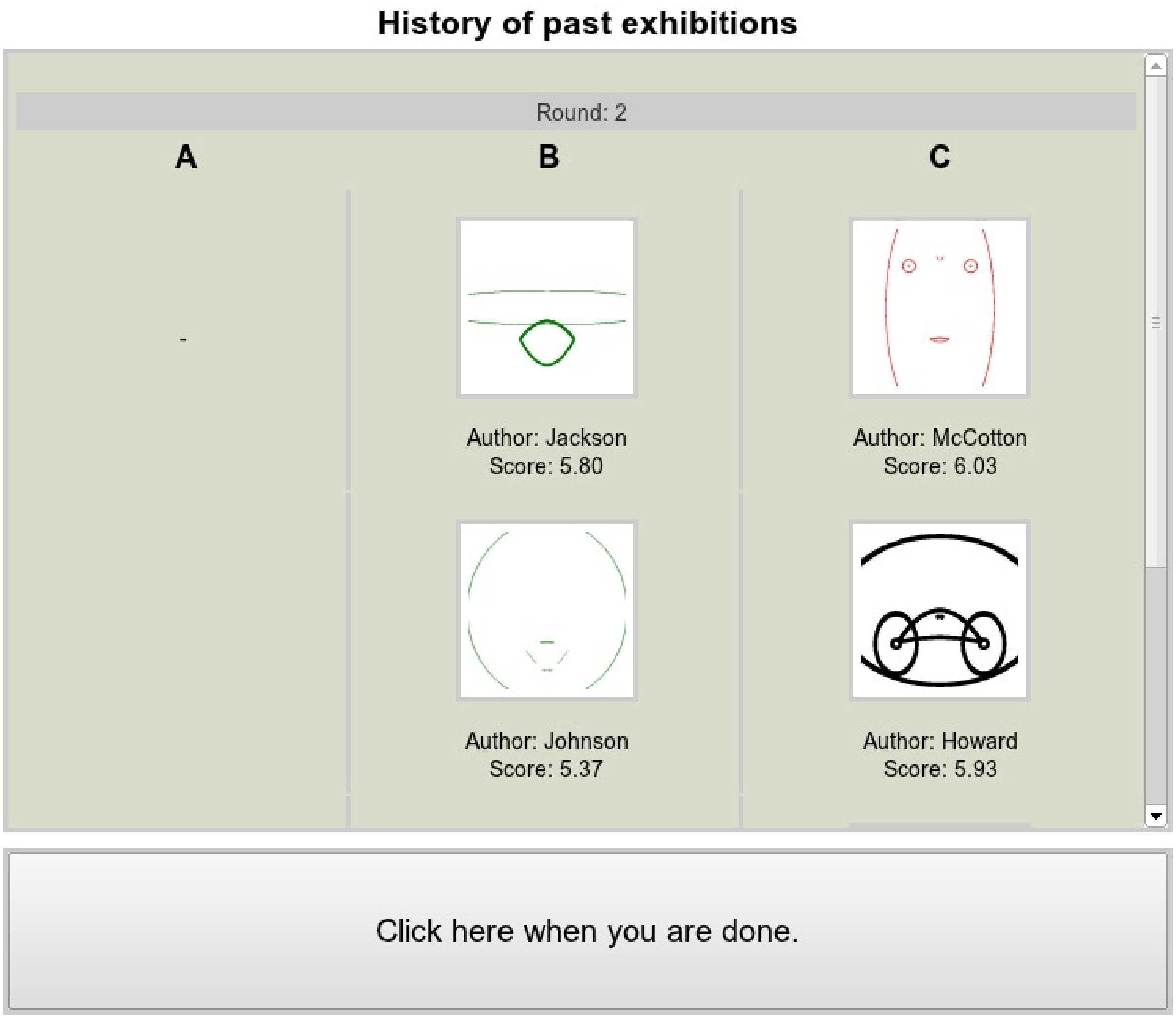
Author: Johnson
Score: 5.37

Author: Howard
Score: 5.93

Figure 8. Illustration of the Exhibitions interface.

At the beginning of each round $r > 1$, participants started to create a new image by modifying the version that was submitted in the previous round. The decision to use the previously submitted image as the initial starting point assured a certain degree of continuity in the development of the artworks of the creator. In addition to the usual "creation" interface, on the right-hand side of the screen a "history" panel was shown. This was sorted by round in descending order, displaying all previously exhibited images together with the name of the author and their score (see Fig. 9). To see the exhibitions that took place at the beginning of the experiment, participants needed to scroll down.

Participants could use the "history" panel to copy an image from a previous exhibition. In order to do so, they could click on an image, which opened a pop-up window with an enlarged version of the chosen



Author: Jackson
Score: 5.80

Author: McCotton
Score: 6.03

Author: Johnson
Score: 5.37

Author: Howard
Score: 5.93

Figure 9. Illustration of the History interface.

image. They could then decide to copy or cancel the operation by clicking the "copy" or "cancel" button. Around 20% of all the images submitted were modified versions of previously copied images, as shown in Fig. 10.

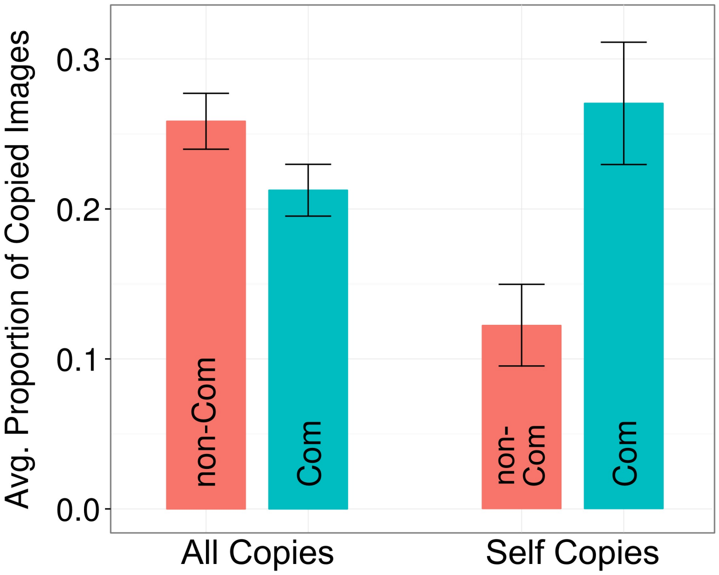


Figure 10. Average share of copied images per condition. (Left) The share of copied images is significantly higher under non-competitive conditions. However, the ratio of images that are copied from own past creations is significantly higher under competitive conditions (right). Error bars show 95% confidence intervals.

After 30 rounds in Stage 3 (see Fig. 1), participants would proceed to a questionnaire that lasted for a maximum of 10 minutes. The results of the questionnaire are reported at the end of this section.

Scope and context of the experimental design.

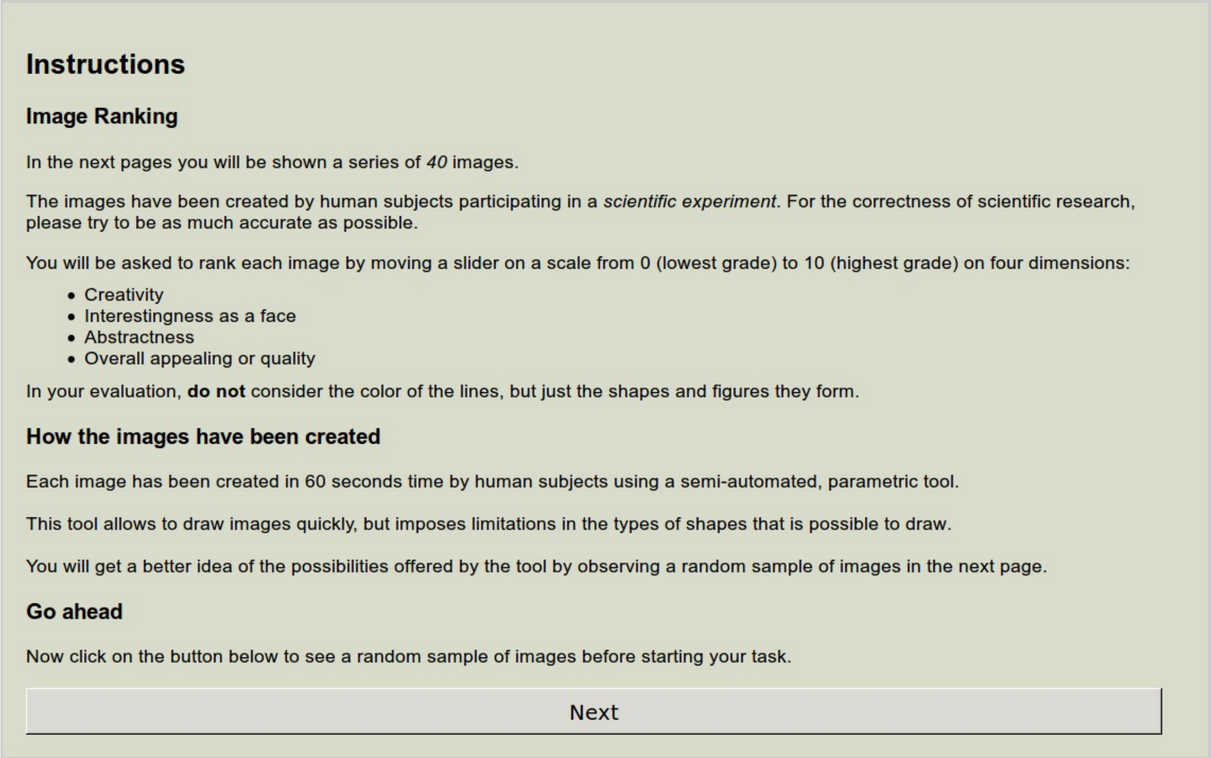
Our design innovates on previous experimental work on the psychology of creative cognition. For example, in the Geneplore model [2] participants are given different shapes and forms from which they construct recognizable objects. The objects they produce are subsequently evaluated by external judges. Although similar, our experimental design features a number of important differences from the Geneplore model. Firstly, our participants exclusively produced creative artifacts using a computer interface. Although this restricted creative freedom to some extent, inventions were still possible, such as using elements of the Chernoff face to create new features such as glasses, ears or a nose, and the creation of abstract art. Using parametric graphics introduces two important advantages: (i) the possibility to measure differences between the art works quantitatively, and (ii) the drastic reduction in the time necessary to generate a new creative artifact. Overall, it led to a more precise estimation of the effects of competition over an extended number of rounds. Finally, as our experimental design aimed to mimic peer review procedures, the judges were not external evaluators, but fellow participants in a creative process that is repeated for several rounds.

Detailed description of the online experiment

In order to independently evaluate the quality of all the images created in the lab, we recruited 620 additional participants from Amazon Mechanical Turk (AMT). Their task consisted of rating a random set of 40 images created in the lab. It lasted 10 to 12 minutes on average, and the participants earned 1.4 USD upon completion. This payment is relatively high by AMT standards and is in line with the US minimum hourly wage. As a result, the participation rate in the experiment was high.[3]

Upon accepting the task, participants from AMT were redirected to a dedicated server which hosted our online experiment. The interface was designed to be perfectly consistent with the graphical interface of the lab experiment in terms of color palette, font family and size.

Firstly, AMT participants were presented with instructions which explained the task, as shown in Fig. 11. Next, they were shown a random sample of 60 images created during the whole experiment. The information page remained open for at least 90 seconds and participants were instructed to view the images attentively. After 90 seconds had elapsed, they could click the continue button to start the classification task itself. This step is illustrated in Fig. 12.



Instructions

Image Ranking

In the next pages you will be shown a series of *40* images.

The images have been created by human subjects participating in a *scientific experiment*. For the correctness of scientific research, please try to be as much accurate as possible.

You will be asked to rank each image by moving a slider on a scale from 0 (lowest grade) to 10 (highest grade) on four dimensions:

- Creativity
- Interestingness as a face
- Abstractness
- Overall appealing or quality

In your evaluation, **do not** consider the color of the lines, but just the shapes and figures they form.

How the images have been created

Each image has been created in 60 seconds time by human subjects using a semi-automated, parametric tool.

This tool allows to draw images quickly, but imposes limitations in the types of shapes that is possible to draw.

You will get a better idea of the possibilities offered by the tool by observing a random sample of images in the next page.

Go ahead

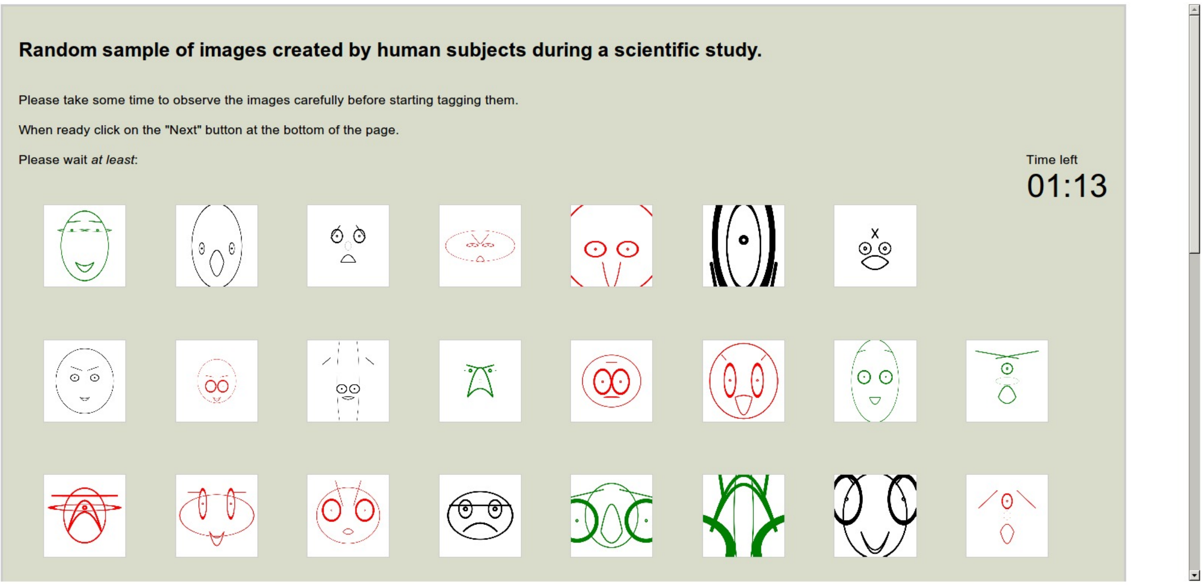
Now click on the button below to see a random sample of images before starting your task.

Figure 11. Illustration of the instructions page for online participants.

Finally, the participants reached the main evaluation task. They were prompted with the following text: “Rank the image above on a scale from 0 (lowest) to 10 (highest). In your evaluation, do not consider the color of the lines, but just the shapes and figures they form.”

They were asked to judge four characteristics: (i) *creativity*, (ii) *abstractness*, (iii) *interestingness as a face*, and (iv) *overall appeal or quality*. The order of these characteristics was randomly shuffled after

[3] Conducting the same task in the lab would have been more than 20 times more expensive and would have required considerably more time also.



Random sample of images created by human subjects during a scientific study.

Please take some time to observe the images carefully before starting tagging them.

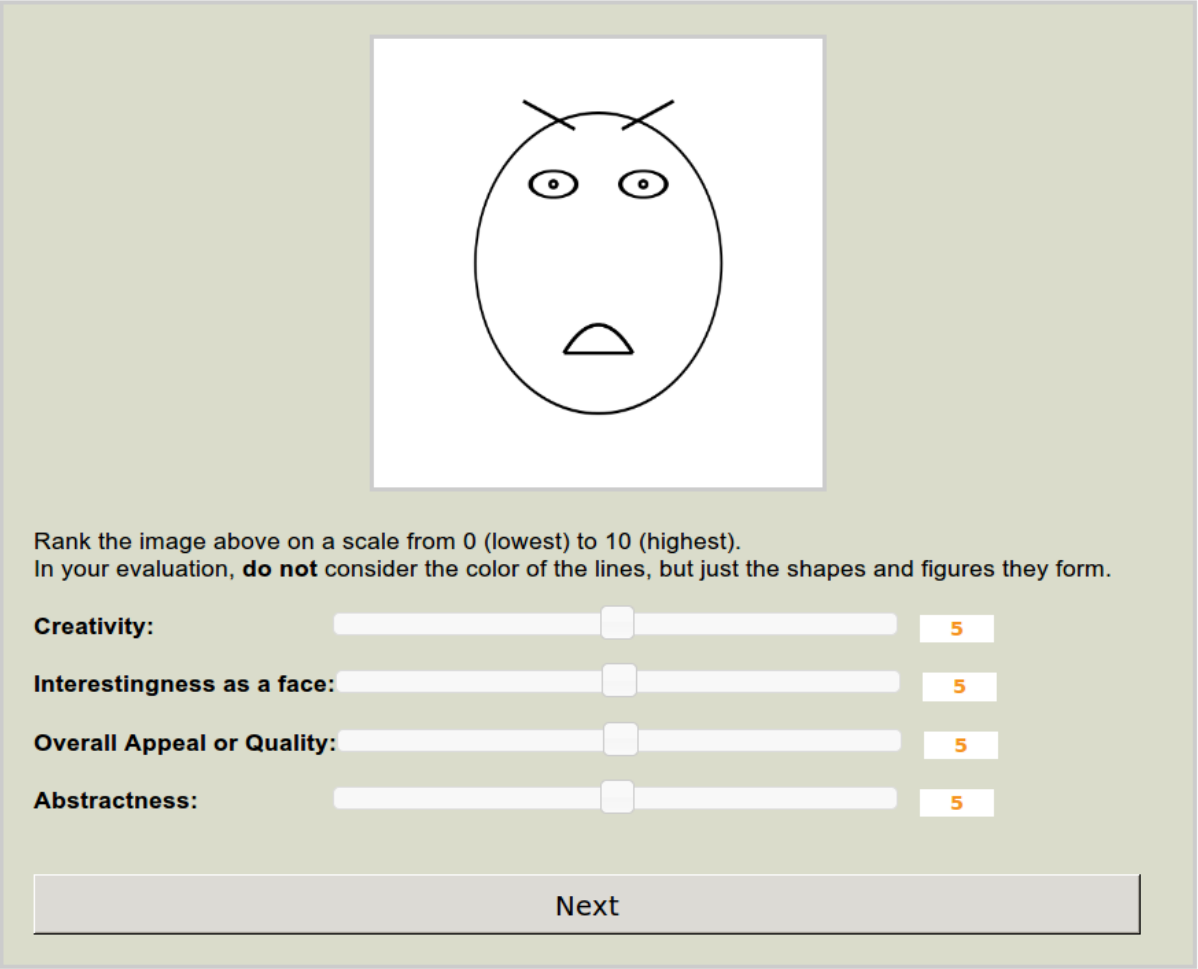
When ready click on the "Next" button at the bottom of the page.

Please wait *at least*:

Figure 12. Illustration of the information page containing a sample of images for online participants. The screen is truncated. Subjects scrolled down to see the remaining images, and to reach the continue button.

each image had been assessed. The procedure is illustrated on Fig. 13.

On average, it took 13 seconds to rank the four characteristics of an image. Evaluations that lasted longer than 50 seconds, or less than 1 second were discarded.



Rank the image above on a scale from 0 (lowest) to 10 (highest).
In your evaluation, **do not** consider the color of the lines, but just the shapes and figures they form.

Creativity: 5

Interestingness as a face: 5

Overall Appeal or Quality: 5

Abstractness: 5

Figure 13. Illustration of the evaluation page for online participants.

Statement of research conduct

The experiment (lab + online) is part of the ERC Advanced Investigator Grant "Momentum" (Grant No. 324247), which was evaluated and approved by the ETH Zurich Ethics committee (EK-2012-N-63). The experiment was conducted in accordance with the ETH Zurich Decision Science Laboratory (DeSciL) Operational Rules, which are approved by the review board and published on the DeSciL website (https://www.descil.ethz.ch/research/policies). The review board of DeSciL is called DeSciL Review Board, and its members are listed on the DeSciL website (https://www.descil.ethz.ch/people). All participants in our experiment were recruited from the subject pool maintained by the University Registration Center for Study Participants (UAST) of the University of Zurich and ETH Zurich. Every person who has signed up to this subject pool also gave his or her informed consent by agreeing to the terms and conditions of UAST. These terms and conditions are published on the UAST website (https://www.uast.uzh.ch/register).

Supplementary data analysis

Here, we report additional statistics pertaining to the results of the laboratory experiment.

Distribution of publications. Fig. 14 shows the global distributions of publications by level of competition. We can see that *more* publications are accepted under non-competitive conditions than in competitive conditions, i.e. 1,171 *vs* 769. This is not surprising, given that a substantial number of referees acted strategically when the degree of competition was increased. More interestingly, the two distributions differ in the amount of skewness, respectively 0.36 *vs* 0.69. A two-sided Kolmogorov-Smirnov test with bootstrap confirms that the difference is significant ($D = 0.4444, P < 0.0001$). However, if we disaggregate the data at the level of each single session and compute a measure of inequality, like the Gini coefficient, the two treatment conditions are *not* significantly different (average Gini coefficients of 0.115 and 0.126 respectively).

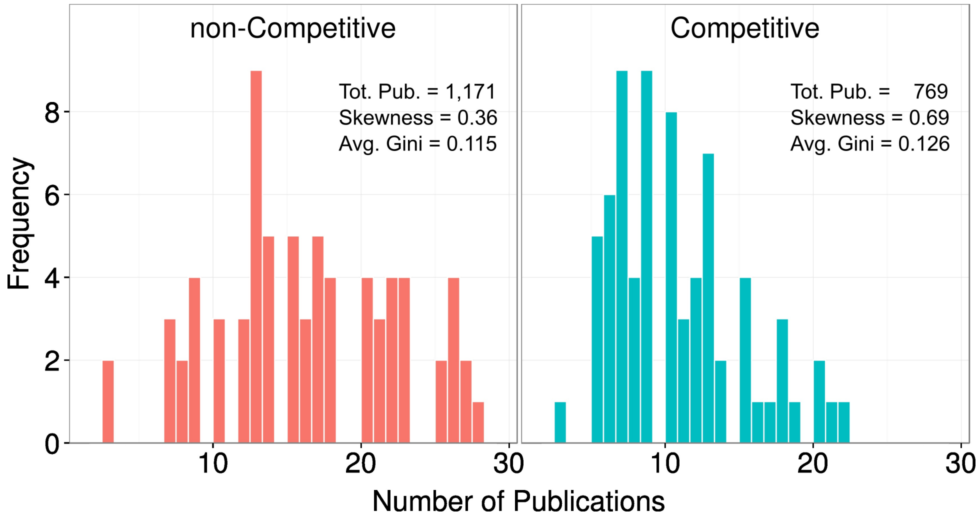


Figure 14. Distribution of publications per participant by level of competition. Under competitive conditions (COM)

Relative Diversity. The relative diversity of an image is defined as the difference between the average distance from images created in the same round in other sessions (between-diversity), and the average distance from images created in the same round in the same session (within-diversity). More formally the relative diversity of image i at round r is defined as:

$$D_{rel}(i,r) = D(i,\hat{i}) - D(i,\bar{i})$$

where D is the diversity index as defined in the main text, $\bar{i}$ are all other images created in the same session at the same round, and $\hat{i}$ are all other images in other sessions at the same round. The value of the relative diversity index over time is plotted in Fig. 15. 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; in non-competitive conditions, conversely, relative diversity is always positive, meaning that a certain degree of social imitation is at play.

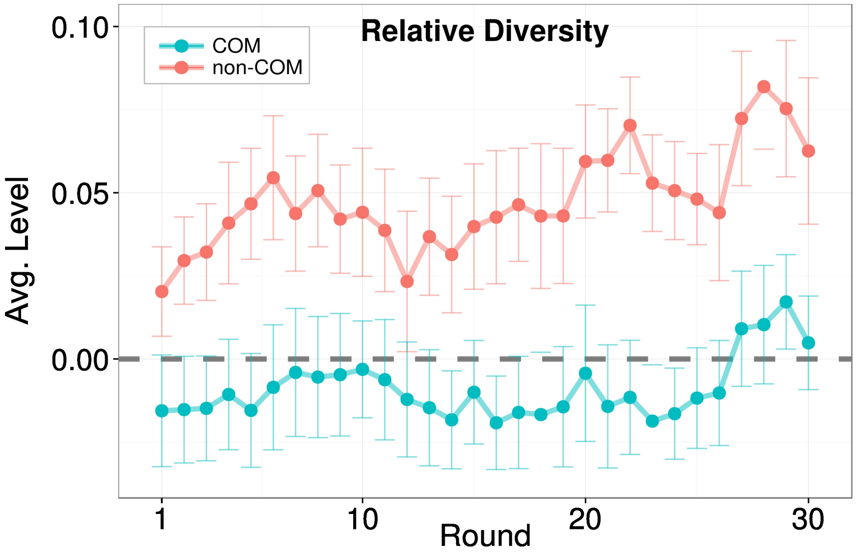


Figure 15. Relative diversity over time (rounds) by level of competition. The relative diversity of images in the competitive session is negative almost for almost all rounds (D), which means that participants are striving to be different from each other. Error bars show 95% confidence intervals.

Innovation and diversity regressions. In the main text we showed that competition fosters innovation and diversity. Here, we quantify this effect using statistical methods. We tested three regression models: (i) a linear model with heteroskedasticity-robust standard errors (variant HC1), (ii) a linear model with heteroskedasticity-robust standard errors (variant HC1) clustered on subjects, and (iii) a hierarchical model with session and subject as random effects. The results are summarized in Tables 3 and 4. In the model with robust standard errors, competition increases the level of innovation by 0.05 units ($P < 0.01$), and the level of diversity by 0.02 units ($P < 0.01$). In the case of innovation, for example, this accounts for more than 20% of the baseline measurement which constitutes a visible change in the appearance of the artworks. The hierarchical model with random effects on subject and session is much more conservative, and therefore the results are significant only at a 95% and 94% level for innovation and diversity, respectively.

Diversity premium. Next, we examine the mechanisms which promote higher levels of innovation under competitive conditions. One hypothesis is that the trend of increasing diversity is driven by a more pronounced preference for dissimilarity. If that is the case, we should find evidence of a "diversity premium" in the average review scores. Fig 16A shows the correlation between the level of innovation of an image and the average review score it received. Under both conditions, reviewers tend to give lower scores when the level of dissimilarity with previously published images is very low. However, when images are highly dissimilar, the situation is slightly different. Under non-competitive conditions, there is a small but significantly ($P = 0.0152$) positive relationship overall. Under competitive conditions, a positive relationship does not exist. This might be due to the strategic behavior of some of the referees. In fact, if we look at published and rejected images separately, we notice that images that were eventually published gained a higher premium for dissimilarity. On the other hand, we also find highly diverse images that did not get published and received low review scores. Hence, the strategic behavior of the referees

Table 3. Regression of innovation and diversity on the level of competition with heteroskedasticity-robust standard errors, variant HC1. Models labeled with a (C) extend the other regressions using heteroskedasticity-robust standard errors clustered on subjects. All regressions were implemented using the statistical software R.

	Innovation	Diversity	Innovation (C)	Diversity (C)
(Intercept)	0.23^{***}	0.25^{***}	0.23^{***}	0.25^{***}
	(0.002)	(0.002)	(0.002)	(0.005)
com1	0.05^{***}	0.05^{***}	0.05^{***}	0.05^{***}
	(0.003)	(0.002)	(0.008)	(0.007)
choice1	0.01^{***}	0.01^{*}	$0.017^{\cdot}$	$0.015^{\cdot}$
	(0.003)	(0.002)	(0.01)	(0.009)
com1:choice1	-0.04^{***}	-0.04^{***}	-0.04^{***}	-0.04^{***}
	(0.004)	(0.003)	(0.0013)	(0.011)
AIC	-9385.1	-11423.1	-	-
BIC	-9353.6	-11391.3	-	-
Log Likelihood	4697.5	5716.5	-	-
Num. obs.	4027	4280	4023	4284

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4. Multilevel regression of innovation and diversity on the level of competition. The regression uses subject and session as random effects. All regressions were implemented using the statistical software R.

	Innovation	Diversity
(Intercept)	0.23^{***}	0.25^{***}
	(0.02)	(0.02)
com1	0.05^{*}	$0.05^{\cdot}$
	(0.02)	(0.02)
choice1	0.02	0.02
	(0.02)	(0.02)
com1:choice1	-0.04	-0.04
	(0.03)	(0.03)
AIC	-10106.4	-12550.6
BIC	-10062.3	-12506.0
Log Likelihood	5060.2	6282.3
Num. obs.	4027	4284

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

might hide a higher premium for diversity under competitive conditions. Overall, however, there is not enough evidence to support this hypothesis.

We also explored another potential mechanism through which competition can promote diversity – as a response to a low review score. This implies that subjects change their image more if they receive a bad review score. Therefore, the lower acceptance rates under competitive conditions could lead to an increased level of diversity. Fig. 16B shows the correlation between the average review score in the previous round and the extent to which the participant altered their image. Overall, there is no significant difference in the level of personal change between competitive and non-competitive conditions. However, in the case of very low review scores ($0.0-2.5$), subjects under competitive conditions change comparably *more* than subjects under non-competitive conditions (Wilcoxon Rank Test with continuity correction $W = 7330, P = 0.01097$). This result is consistent with the hypothesis that the high number of low review scores observed under competitive conditions trigger higher levels of innovation and diversity.

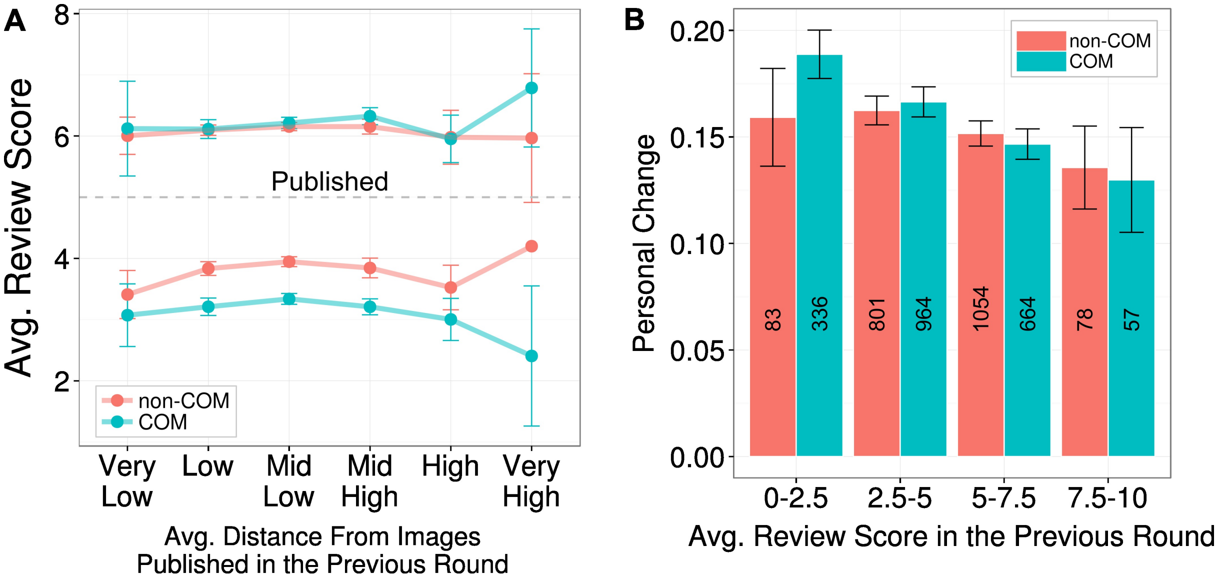


Figure 16. How average review scores affect innovation and personal change by level of competition. (A) Correlation between the level of innovation of an image and its average review score. The data is disaggregated by published and rejected images. Under both conditions, lower scores tend to be given when the level of dissimilarity with previously published images is very low. However, this effect is only significant under non-competitive conditions, and the size is relatively small. Moreover, there is no conclusive evidence of a premium for highly diverse images. The 6 classes of dissimilarity corresponds to the following 6 intervals: *"Very Low" = (0.00998 – 0.108), "Low" = (0.108 – 0.205), "Mid Low" = (0.205 – 0.303), "Mid High" = (0.303 – 0.401), "High" = (0.401 –0.499), "Very High" = (0.499 – 0.596).* (B) Correlation between the average review score in the previous round and the amount of change in the image by subject. Competition (COM) makes subjects more reactive to bad review scores. The number inside the bars indicate the number of items in the category. Error bars show 95% confidence intervals.

Asymmetric Strategic Selective (A.S.S.) reviews. As explained in the main text, an Asymmetric Strategic Selection (A.S.S.) review refers to a case where a very low score (below 0.5) is given to a direct competitor. The A.S.S. index is defined as the ratio between the number of times an A.S.S. review was given, compared to the number of opportunities to submit an A.S.S. review. For example, if a participant reviews 2 images from direct competitors in a particular round, there are 2 opportunities to submit A.S.S. reviews. If only 1 A.S.S. review is given, the A.S.S. index of the reviewer is 0.5. The A.S.S. index varies therefore between 0 and 1; within this interval we further define two categories: an average A.S.S. index below 0.5 is considered to be non-A.S.S, whereas levels above 0.5 are labeled as Asymmetric Strategic Selective (A.S.S.). As shown in Fig. 17A, and B the majority of reviewers fall into the non-A.S.S. category in both COM and non-COM conditions. However, under competitive conditions, we find a significantly larger number of A.S.S. reviews and A.S.S. reviewers than under non-competitive conditions. Furthermore, under competitive conditions, the average A.S.S. index rapidly increases over time, although both COM and non-COM start at the same level (see panel C).

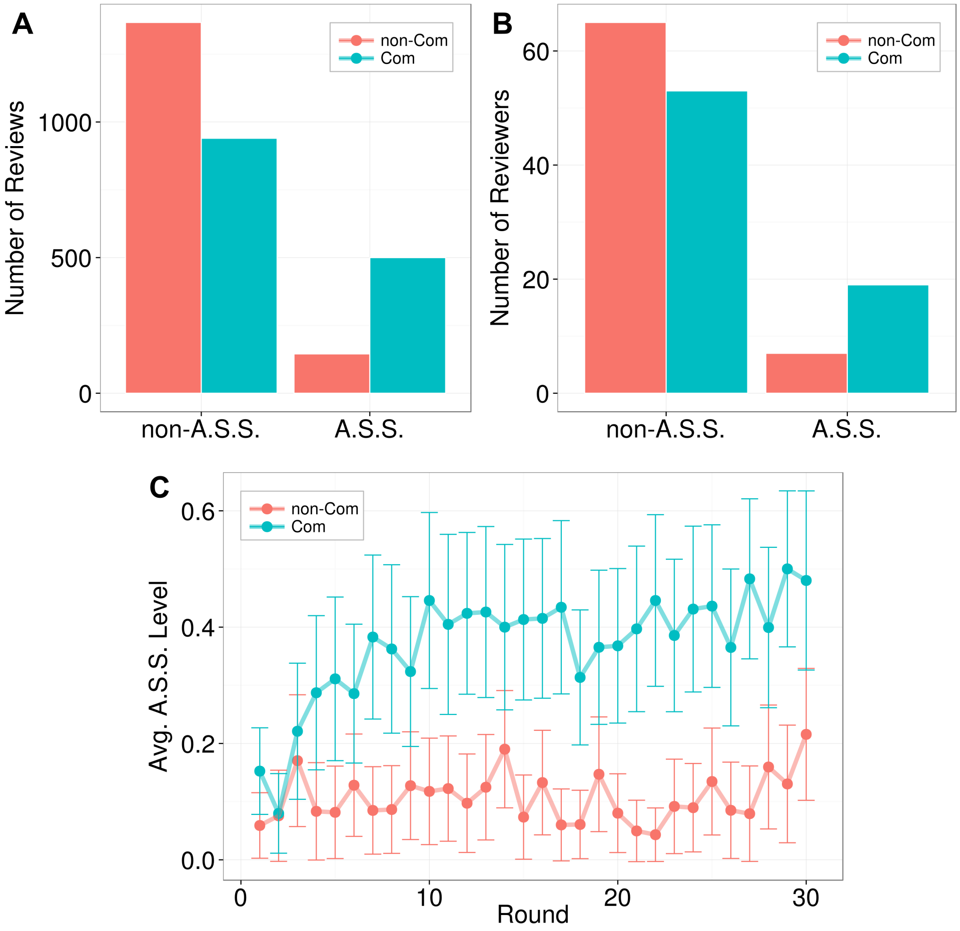


Figure 17. Competition increases the number of A.S.S. reviews and reviewers; such increase develops quickly, so that already after two rounds the average A.S.S. index of competitive conditions is significantly higher than that of non-competitive conditions. Error bars show 95% confidence intervals.

As a form a heuristic decision, A.S.S. reviews are expected to take less time to be executed than unbiased reviews. We recorded the duration of all evaluation rounds for each participant as measured on the server machine. The result of our analysis shows that A.S.S. reviews are on average about 1.8 seconds faster than unbiased reviews, i.e. about 11% faster. A Wilcoxon rank-sum test confirms that the difference is significant ($W = 275380.5, P < 0.001$). However, given that our measure of time is approximate, we need to take this result with a grain of caution. In fact, our server-side measure of time does not take into account situations in which a decision was taken quickly, but the participant did not click the "Continue" button, and instead just waited for the timer to expire. Regardless of these situations, decision times recorded on the server are slighter higher than if they would be if measured directly on the browser of the clients due to network latency and concurrent computing processes. Nonetheless, such bias is expected to be consistent across all sessions and conditions due to the highly controlled settings in which the experiment was run at the ETH Decision Science Laboratory (DeSciL). Further research is needed to target the issue of response times across reviewers' profiles.

Finally, it is a licit question to ask whether participants who constantly deliver A.S.S. reviews manage to accumulate a higher payoff at the end of the experiment. As expected, there is a trend such that self-interested reviewers do earn more on average. However, the difference is not statistically significant: their expected earnings per round are 0.716 ± 0.101 vs 0.646 ± 0.048 of other participants. Interestingly, self-interested reviewers also tend to publish slightly less: avg. publication per round $= 0.316 \pm 0.041$ vs 0.368 ± 0.023. However, this difference is not significant when controlling for session and player as random effect in a hierarchical regression model.

Consensus among referees. Consensus is formally defined as the ratio between the standard deviation of the reviews R_x given to an image x and the maximum standard deviation possible:

$$C(i) = 1 - \frac{\sigma(R_1, R_2, R_3)}{max(\sigma_R)}$$

As already articulated in the main text, competition causes a drop in the level of consensus among referees over rounds, as Fig. 18 shows.

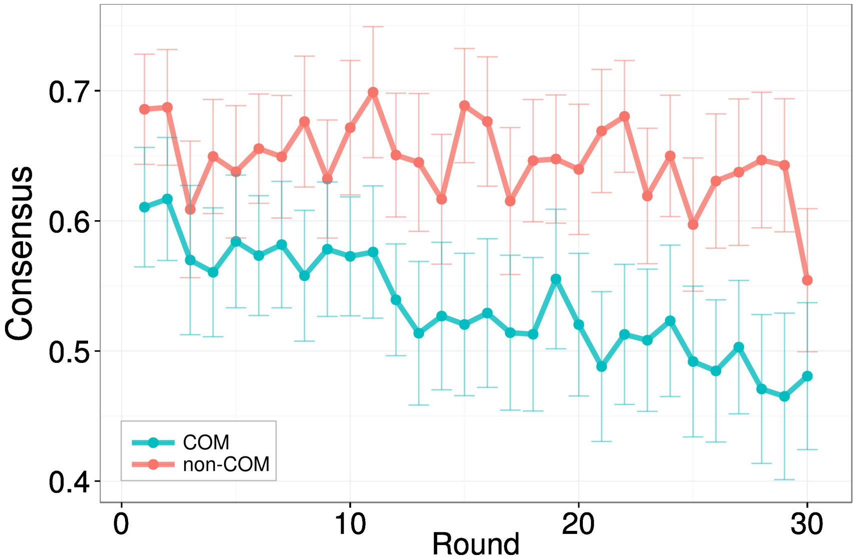


Figure 18. Consensus amongst referees over time (rounds) by level of competition. Under competitive conditions (COM) the level of consensus among referees steadily decreases over time.

The drop in consensus under competitive conditions can be better understood by looking at the distribution of review scores over time, displayed in Fig. 19. We can see how the width of the interquartile range increases under competitive conditions, while it remains more or less constant under non-competitive conditions. The figure suggests that competition causes a drop in consensus, evidenced by the increasing number of very low review scores.

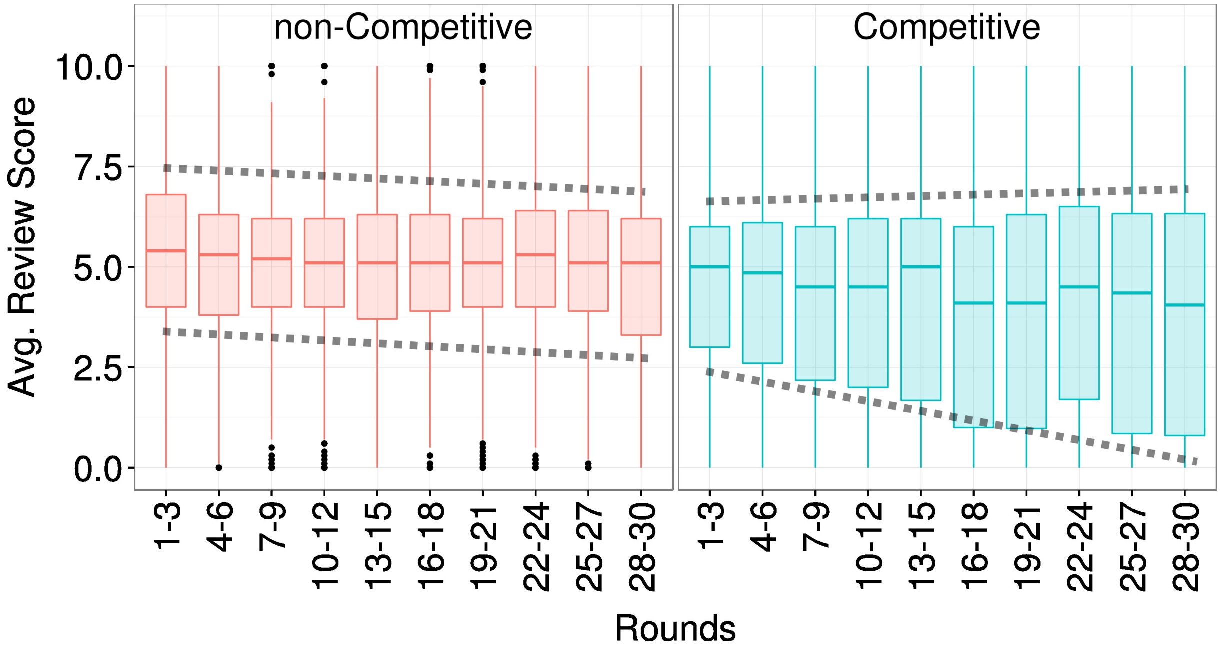


Figure 19. Distribution of review scores over time (rounds) by level of competition. Under competitive conditions, the number of low review scores surges over time. As a consequence, the spread between first and third quartile of the within-round distribution increases. The dotted lines bridge the quartiles for the first and last period (a vertical shift has been added for visibility).

In order to quantify the reduction in consensus caused by strategic reviews, we constructed the following null model. For each round, we randomly assigned the review scores to images created in the same round. Then we measured the consensus among referees again, and compared it to the original consensus index for the same round. We repeated the procedure 100 times. The difference in consensus between the original and null model for each round is shown in Fig. 20. There is less consensus in competitive sessions, indicating that referees are engaging in "gaming" behavior.

Moreover, we also computed the average intraclass correlation coefficient (ICC) per round across the two conditions. We found that our correlational measure of consensus is consistent with ICC, that has a baseline level of which 0.6, which is reduced by competition by 0.023 ($P < 0.001$).

In-group biases. In the detailed description of the lab experiment, we explained that participants were assigned to a group of three individuals whose images shared the same color. Here, we test if the in-group color was used as a criteria for discriminating against out-group members in the review scores. A Kolmogorov-Smirnov test with bootstrap on the cumulative distribution of review scores shows that there is no color bias under competitive conditions, whereas a small, but significant color bias exists under non-competitive conditions ($D = 0.0518, P = 0.003484$). However, the bias seems not to be persistent over time, evidenced by the fact that the difference is no longer significant after the first eight rounds (see Fig. 21). The bias under non-competitive conditions might be seen as an initial attempt to cooperate within the in-group. Alternatively, the lack of color bias under competitive conditions can be explained by theories of the psychology of competition which postulate that individuals may be most competitive with individuals who are more similar to them [4]. Hence, under competitive conditions participants do not exhibit the same color bias. In sum, combined with the analysis of the questionnaire responses, color bias seems to have played a marginal role in the review scores over the course of the whole experiment.

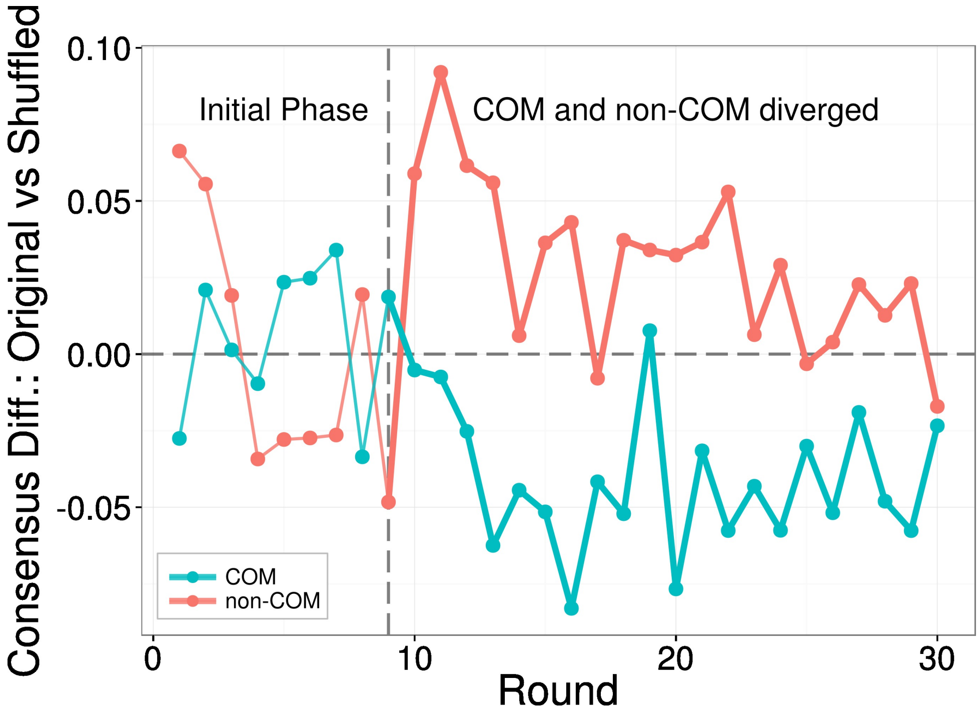


Figure 20. Difference in consensus between actual reviews and shuffled reviews. Actual consensus compared to a null model with shuffled reviews. There is *less* consensus in competitive sessions, indicating that the referees are engaging in "gaming" behavior.

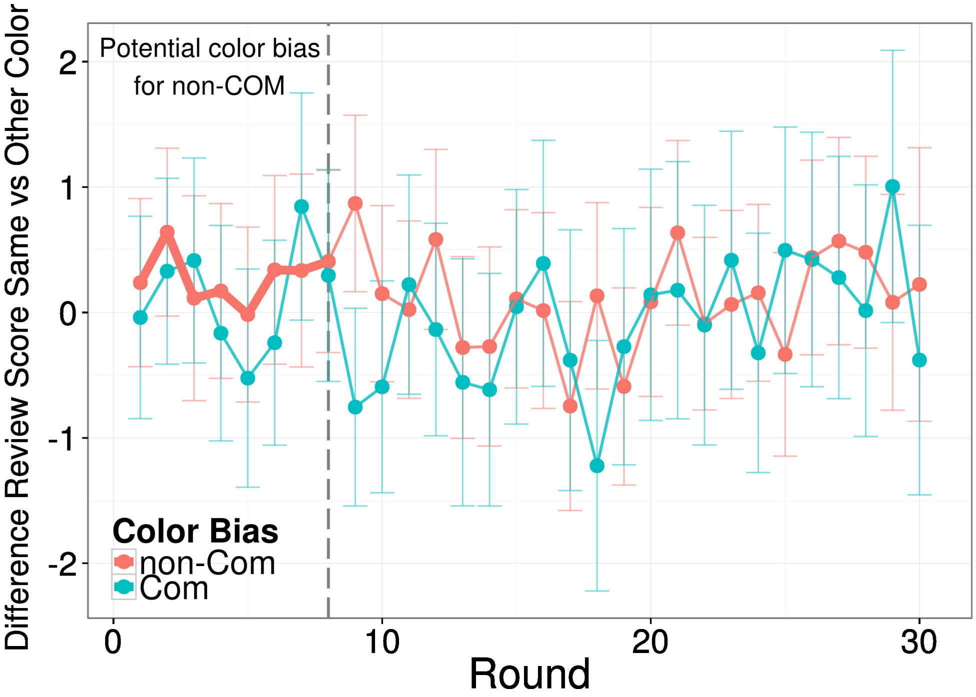


Figure 21. Color bias in review scores. The plot shows the average difference in review scores given by a reviewer to images of the same color, compared to other colors. Initially, a significant, but small-sized bias seems to exist in non-competitive sessions. However, after round 8 the difference is no longer significant.

Questionnaire Results.

Here, we present a summary of the responses to the final questionnaire, which participants completed after the lab experiment. The analysis is a useful way to gain further insight into the decision-making process of the participants, and it provides further evidence in support of the main arguments of the paper.

The questionnaire consisted of 8 questions about the nature of the game and the strategies that the participants used. The questionnaire allowed for both open-ended comments and multiple-choice answers, and on average it required about 8 minutes to complete. The multiple-choice questions prompted participants to express their level of agreement with a certain statement on the following discrete scale: *"Complete disagree" (0), "Quite disagree" (1), "Quite agree" (2), "Complete agree" (3).* The order of the options for the multiple-choice answers was randomized for each participant.

Although we have found some differences in the participants' responses depending on whether the setting was competitive or non-competitive, such differences were never significant. Therefore, only aggregated statistics are presented here.

The participants' understanding of how to use the "creation" interface (Fig. 5) is crucial to the validity of the results of the study. The majority of participants reported to have felt immediately comfortable using the "creation" interface . Some mentioned that they underwent a natural learning process during the first rounds, and that they subsequently understood how to use the interface well. Interestingly, many participants wished to have had the possibility to modify more parameters, whereas conversely, almost none reported that the interface had too many options (see Fig. 22).

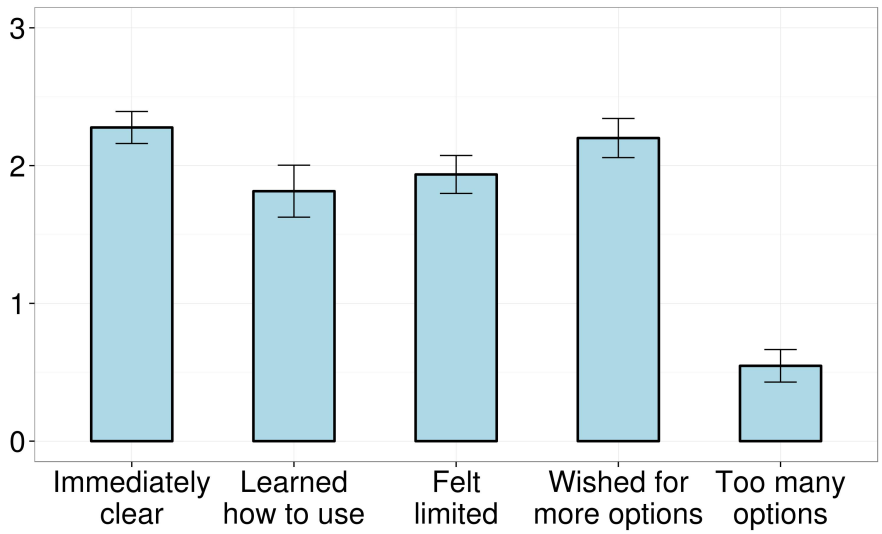


Figure 22. Average level of agreement with statements about the usability of the "creation" interface. From left to right the complete statements are: *(i) It was immediately clear how to use it; (ii) I felt a bit confused at the beginning, but then I understood well how to use it; (iii) I felt limited in the possibility of expressing my creativity; (iv) I would have liked to modify other parameters; (v) There were too many parameters to modify.* Error bars are 95% confidence intervals.

Secondly, we wanted to assess the motivations of participants when creating new images. Many participants reported that were motivated by a desire to express their creativity, as evidenced in comments such as: *"I wanted to do something creative"*, *"I always tried to come up with a new idea that no one had chosen before."* Moreover, even if the "creation" interface was based on a modified version of a Chernoff face [1], participants understood that they could go beyond producing images resembling faces. For example, some reported that they tried to *"make something original, to surprise the judges. Not necessarily a face."*, or that *"when a thought popped into my mind I just changed it, for example I did an angry bird once."* One participant went even further, explaining that his/her aim was to *"create smart, beautiful and touching drawings."* These results confirm that our paradigm encourages diversity and creativity, rather than convergence, in contrast to previous studies on social influence [5–8]. A second motivation which is evident in Fig. 23 is that participants frequently indicated that they were more inclined to change their image if their work had been rejected in the previous round. This behavior partially explains the higher level of innovation and diversity under competitive conditions, where the rejection rates are markedly higher.

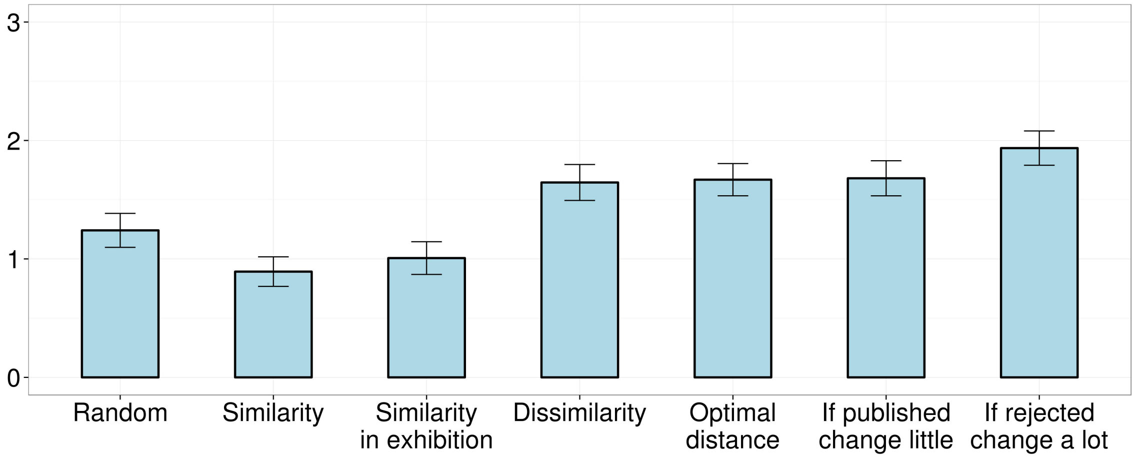


Figure 23. Average level of agreement with statements about the strategy used for creating images. From left to right the complete statements are: *(i) I was changing parameters randomly; (ii) I tried to be more similar to all the other images; (iii) I tried to be more similar to the other images that were displayed in the exhibition to which I submitted; (iv) I tried to be as different as possible to all the other images; (v) I tried to be a little different, but not too much from the other images; (vi) If my image was published, I changed little; (vii) If my image was not published, I changed a lot.* Error bars are 95% confidence intervals.

When asked about their criteria for reviewing other images, many participants stated that they followed *only* their own personal taste, as shown in Fig. 24. However, some also admitted that they considered whether the image was created by a direct competitor. This behavior is succinctly summarized by one of the comments: *"In principle... grading badly the others that were trying to post in the same gallery and be objective for the other galleries"*. Interestingly, the percentage of participants who admitted that they evaluated other works strategically is not significantly different under competitive and non-competitive conditions. This may indicate that the subjects perceived this behavior to be potentially socially undesirable and preferred not to report it. Conversely, one participant disclosed an even more sophisticated strategy: *"I gave 0 for paintings which were interesting, so I could copy and only a few people could see it (chance of win was higher in a second round)."*

When examining the propensity of participants to copy other images, there was a low level of agreement

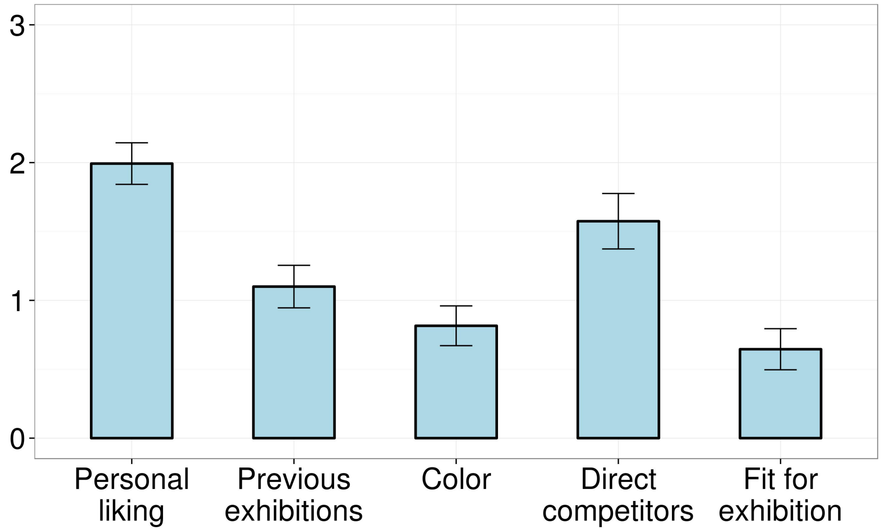


Figure 24. Average level of agreement with statements about the criteria used for reviewing images. From left to right the complete statements are: *(i) My judgment was based only on my personal liking; (ii) My judgment was influenced by what had been displayed in the previous exhibitions and their scores; (iii) My judgment was influenced by the color of the submission; (iv) I gave lower scores to players submitting in the same exhibition to which I had submitted; (v) I tried to judge if the image was good for a specific exhibition.* Error bars are 95% confidence intervals.

with the three proposed statements. However, 19 participants filled in the optional open-ended comments section and explained that they used the copy functionality mainly to copy their *"own old paintings"*, or when they *"wanted to change [...] drastically, to save some time"*. About one third said they did not copy any image (see Fig. 25).

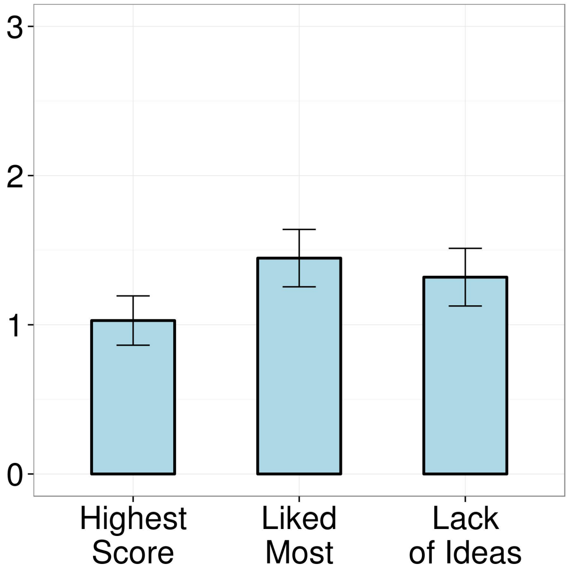


Figure 25. Average level of agreement with statements about the usability of the criteria used for reviewing images. From left to right the complete statements are: *(i) I copied the images that had the highest score; (ii) I copied the images that I liked the most; (iii) I copied the images when I didn't have a good idea myself.* Error bars are 95% confidence intervals.

References

1. Herman Chernoff. The use of faces to represent points in K-Dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368, 1973.

2. R.A. Finke, T.B. Ward, and Steven M. Smith. *Creative Cognition: Theory, Research, and Applications.* MIT press Cambridge, MA, 1992.

3. L. Fogg and D.W. Fiske. Foretelling the judgments of reviewers and editors. *American Psychologist*, 48(3):293–294, 1993.

4. S.M. Garcia, A. Tor, and T.M. Schiff. The psychology of competition a social comparison perspective. *Perspectives on Psychological Science*, 8(6):634–650, 2013.

5. D.T. Kenrick, N.P. Li, and J. Butner. Dynamical evolutionary psychology: Individual decision rules and emergent social norms. *Psychological Review*, 110(1):3–28, 2003.

6. B. Latane and M.J. Bourgeois. *Handbook of Social Psychology: Vol, 4 Group Processes*, chapter Dynamic Social Impact and the Consolidation, Clustering, Correlation, and Continuing Diversity of Culture, pages 235–258. Blackwell, 2001.

7. J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences (PNAS)*, 108(22):9020–9025, 2011.

8. W.A. Mason, F.R. Conrey, and E.R. Smith. Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*, 11(3):279–300, 2007.

9. H. Tajfel and J.C. Turner. The social identity theory of intergroup behavior. In S. Worchel and L.W. Austin, editors, *Psychology of Intergroup Relations*, pages 7–24. Nelson-Hall, 1986.