Supplementary Information for "Viewing men's faces does not lead to accurate predictions of trustworthiness"

Charles Efferson[1,2] & Sonja Vogt[1,2]

[1]*Department of Economics, University of Zurich, Zurich, Switzerland*

[2]*Laboratory for Social and Neural Systems Research, University of Zurich, Zurich, Switzerland*

1 Experimental methods, the trust game

Munich, the sequential social dilemma. The trust game took place at the Munich Experimental Laboratory for Economic and Social Sciences at the University of Munich. 134 males participated in a total of six sessions. Sessions had between 18 and 24 participants and took about one hour. The vast majority of subjects were students at the University of Munich. Students studying psychology or economics were not invited. All participants received a show-up fee of 4 Euros, which along with any money earned during the experiment was paid privately in cash at the end of the session. The experiment itself consisted of two parts. In the first part subjects played an anonymous one-shot trust game on a local computer network using z-Tree[1]. In the second part of the experiment each subject had his photograph taken in private.

The precise sequence of events was as follows. After everyone had read a set of instructions and correctly answered several control questions to demonstrate understanding, participants were paired at random, with one participant in the role of first mover and the other in the role of second mover. Each player received an endowment of 9 Euros. The first mover began the game by deciding whether to transfer this endowment. If the first mover transferred, the money was doubled by the experimenter, and the second mover thus received 18 Euros in addition to his initial endowment. If the first mover did not transfer, both players proceeded to the second stage of the game with their original endowments. In the second stage, the second mover specified an amount between 0 and 9 Euros, in 1-Euro increments, to back transfer to the first mover. This back transfer, whatever it was, was also doubled by the experimenter.

For the duration of the experiment, subjects could refer to printed instructions, which included a matrix specifying the exact amounts received for each of the two players under every possible combination of choices.

After the game was completed, we took photographs of each participant in private. Participants were informed about the photos when initially invited to participate. Moreover, before beginning an experimental session, we informed participants in Munich that subjects in a study in another city in Germany would later see these photographs, but we would not reveal the names or behaviours of specific participants in Munich. Participants in Munich signed a consent form at the beginning of the experimental session, and they were free to decline or opt out of the study at any point in time[1]. Our methods were approved by the Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich and by the Office of Data Protection and Privacy at the University of Munich (i.e. Datenschutzbeauftragte der Universität München). All photos were taken privately, against the same backdrop, and according to a standardised procedure such that the camera lens was approximately one meter from the participant's face and approximately four centimetres below the top of the participant's head. Participants were asked to adopt a neutral facial expression. We took one picture with the participant's normal clothing, although we asked them to remove eyeglasses, coats, and any conspicuous necklaces they may have been wearing. We also took a second photograph in which participants wore surgery caps in a standardised way that obscured their hair. Because hair styles carry all sorts of social meaning, we presented the photographs taken with surgery caps to raters in Konstanz.

As mentioned in the main paper, we wanted to isolate the ability of raters to predict second mover behaviour without age or ethnicity as a confound. For example, our subjects in Munich were generally undergraduates at the University of Munich. We had one subject, however, who at 61 was substantially older than the other players. To avoid having to control for the age of second movers with most of the variation due to a single outlying data point, we simply excluded him from the sample of second movers

[1]Only two participants opted out before the end of the session. One of them was a first mover whose trust was exploited by his partner. He also happened to be near the end of the line for having his picture taken, and he argued that he did not want to wait around to collect only the show-up fee of 4 Euros. The other person who left had not carefully read our initial invitation and thus did not realise we would take photographs. He left when we explained this again at the beginning of the session.

before analysing any of our behavioural data or conducting any sessions with raters in Konstanz. For analogous reasons, we also excluded 12 additional pairs in a strictly ex ante fashion because each of these second movers had a father or mother who did not speak German as a native language. Altogether, this procedure yielded 54 second movers as a sample for the sessions with raters in Konstanz.

Munich, measuring faces. We calculated the width-to-height ratios for second movers using the photographs taken while they wore surgery caps. Specifically, upper facial height was measured as the distance from the top edge of an individual's upper lip to the upper edge of the eyelids. After establishing this distance, a rectangle was formed from one edge of the individual's face to the other. The dimensions of this rectangle yielded the bizygomatic width-to-height ratio [2–4], and this ratio is the unit-free measure of facial structure we used in all the relevant analyses discussed in the main paper. Each author independently measured the facial width-to-height ratios without consulting each other in any way and without any information whatsoever regarding the behaviours of the men in the photographs. This procedure resulted in an inter-rater correlation of 0.928 ($P < 0.001$), and analyses discussed in the main paper are based on the measurements of C.E. We both used the Preview program standard on Apple computers to measure width-to-height ratios, but we also tested a few individuals to confirm that our Preview measurements were similar to those achieved with the software ImageJ. We did this simply because ImageJ seems to be frequently used for measuring facial width-to-height ratios. We found, however, that the graphical depiction of rectangles in Preview allowed for more precise measurements. In the end, because the width-to-height ratio is a unit-free measurement, the choice of software is largely irrelevant. Any software that allows one to look at photographs at the correct aspect ratio and draw rectangles should be perfectly suitable.

2 Experimental methods, drawing inferences

Konstanz, guessing the behaviour of second movers in Munich. The second task took place at the Lakelab at the University of Konstanz, Germany. 13 females and 15 males participated in a single session that took about an hour. The instructions explained to subjects that they would see photographs of people who

had already participated in another study in another city in Germany. In addition to their own instructions, subjects were also provided with the original instructions from the study in Munich. Before beginning, in order to ensure that everyone understood both their own task and the task faced by participants in the Munich study, all subjects in Konstanz had to correctly answer several questions associated with both sets of instructions[2].

Each subject saw the photos of second movers from the Munich study. We cropped photos to a standard height of 1180 pixels. Photos were presented in z-Tree[1] in a random order that was independently determined for each subject. Each rater viewed the photos on a 19-inch Belinea monitor with a resolution of 1280 $\times$ 1024. We used the multimedia box in z-Tree to actually present the photos, and we programmed photo presentation so that photos constituted 67.2% of the monitor's height while maintaining the aspect ratio of the original photos. When viewing the photo of a specific second mover, each subject was also informed whether this individual's partner (i.e. the first mover) transferred 0 or 9 Euros to the second mover in the photo. Each subject then guessed how many Euros (i.e. an integer between 0 and 9) the second mover back transferred to the first mover. Subjects could take as long as they wanted to view a photo and specify a guess, and they were encouraged to look at the payoff matrix from the instructions used in Munich as often as necessary. Subjects received a show-up fee of 4 Euros. In addition, if a subject guessed the back transfer of a second mover correctly, she received 50 cents with a probability of 0.9 for this guess. If a subject's guess deviated 1 Euro from the actual back transfer of the second mover, the participant received 50 cents with a probability of 0.5 for the guess. If a subject's guess deviated 2 Euros or more from the actual choice of the second mover, the subject received 50 cents with a probability of 0.1. Subjects received no feedback about their accuracy during the guessing task; they only received information about their total earnings at the end of the session. Our payoff procedures preserved anonymity by ensuring that subjects in Konstanz could never infer the actual behaviour of any specific second mover from Munich. They also, however, ensured that subjects had a clear incentive to guess accurately because on average subjects earned more money for correct inferences.

Konstanz, estimating the attractiveness of second movers in Munich. We also conducted a second session in Konstanz with 15 women and 13 men. These subjects viewed the photographs of the second

[2]Both instructions in German are available upon request.

movers from Munich and rated each one in terms of attractiveness on a five-point scale ranging from "very attractive" to "very unattractive". These subjects received 4 Euros as a show-up fee, and they received 20 cents for every face they viewed and rated. In the relevant analyses, we used the mean attractiveness of each second mover as an independent variable. We calculated these mean attractiveness values for each second mover by averaging over the 28 separate ratings of attractiveness obtained for each second mover. As discussed in the main paper, this independent measure of attractiveness enabled us to evaluate if attractive second movers were especially pro-social or anti-social and if they were perceived as such by raters.

1. Fischbacher, U. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* **10**, 171–178 (2007).

2. Weston, E. M., Friday, A. E. & Lio, P. Biometric evidence that sexual selection has shaped the hominin face. *PLoS ONE* **2**, e710 (2007). URL `http://dx.plos.org/10.1371/journal.pone.0000710`.

3. Carré, J. M. & McCormick, C. M. In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society B* **275**, 2651–2656 (2008).

4. Stirrat, M. & Perrett, D. I. Valid Facial Cues to Cooperation and Trust. *Psychological Science* **21**, 349–354 (2010).

5. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).

Table S1. Binary accuracy of rater guesses about the choices of the 41 second movers who were trusted. We coded back transfers and guesses as either zero or positive. A guess was accurate if it matched the actual back transfer in either direction (i.e. zero or positive), and the table presents the number of accurate guesses by rater. The table also shows P values based on two-sided binomial tests and no correction for multiple hypothesis tests. Under random guessing, the expected number of correct guesses is 20.5. After correcting for multiple tests with an overall false discovery rate[5] of $\delta = 0.05$, only rater 9 was significantly above chance.

Rater	# correct	P-value
1	21	1.000
2	23	0.533
3	29	0.012
4	24	0.349
5	23	0.533
6	13	0.028
7	15	0.117
8	30	0.004
9	32	<0.001
10	20	1.000
11	20	1.000
12	18	0.533
13	26	0.117
14	27	0.060
15	18	0.533
16	20	1.000
17	29	0.012
18	29	0.012
19	23	0.533
20	21	1.000
21	21	1.000
22	19	0.755
23	23	0.533
24	23	0.533
25	27	0.060
26	28	0.028
27	28	0.028
28	13	0.028

Table S2. Binary accuracy of rater guesses about the choices of all 54 second movers. We coded back transfers and guesses as either zero or positive. A guess was accurate if it matched the actual back transfer in either direction (i.e. zero or positive), and the table presents the number of accurate guesses by rater. The table also shows P values based on two-sided binomial tests and no correction for multiple hypothesis tests. Under random guessing (R), the expected number of correct guesses is 27. Raters 8, 9, 14, 17, 18, 25, 26, and 27 had accuracy rates significantly higher than this after correction[5] using an overall false discovery rate of $\delta = 0.05$. A different null is based on a simple reciprocity heuristic (H). The heuristic posits that, when trusted by their partners, second movers back transfer some positive amount. Otherwise, they do not. Because back transfers were positive for 28 of the 41 second movers who were trusted and two of the 13 who were not trusted, this heuristic produces 39 correct guesses in our sample. With correction, raters 1, 6, 7, 12, 15, 16, 20, 21, 22, and 28 had accuracy rates significantly lower than this. No one was significantly above.

Rater	# correct	P-value (R)	P-value (H)
1	30	0.497	0.009
2	34	0.076	0.131
3	35	0.040	0.226
4	35	0.040	0.226
5	32	0.220	0.047
6	24	0.497	<0.001
7	25	0.683	<0.001
8	38	0.004	0.762
9	41	<0.001	0.649
10	31	0.341	0.021
11	31	0.341	0.021
12	28	0.892	0.002
13	35	0.040	0.226
14	38	0.004	0.762
15	29	0.683	0.004
16	30	0.497	0.009
17	40	0.001	0.880
18	39	0.001	1.00
19	34	0.076	0.131
20	29	0.683	0.004
21	28	0.892	0.002
22	30	0.497	0.009
23	34	0.076	0.131
24	34	0.076	0.131
25	38	0.004	0.762
26	39	0.001	1.00
27	39	0.001	1.00
28	24	0.497	<0.001

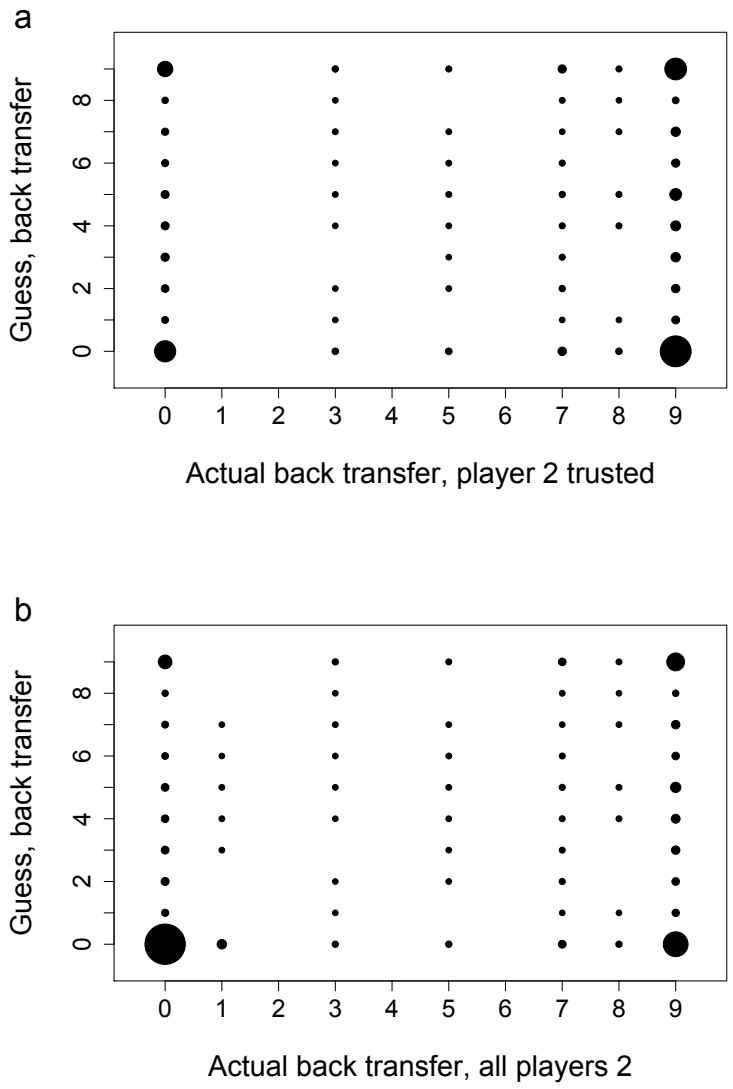

Figure S1. Guesses about back transfers. Panel **a** shows the distribution of guesses about back transfers and actual back transfers for the 41 second movers who were trusted. Each of the 28 raters in Konstanz made a guess, and so the total number of observations represented is 1148. The figure is a histogram in the sense that, for each combination of actual back transfer and guess about back transfer, the size of the circle is proportional to the percentage of observations associated with this combination. Panel **b** shows the same type of distribution for all 54 second movers, whether trusted or not, and the total number of observations represented is 1512. The large mass at (0,0) in this figure indicates that many raters understood that a transfer of 0 was likely to yield a back transfer of 0. This basic understanding of reciprocity was enough to generate significantly accurate inferences for some raters in Konstanz, but we have no evidence that raters did draw or could have drawn any inferences about pro-social tendencies from the photographs of second movers.