games

MDPI

Article

Cooperation through Image Scoring: A Replication

Yvan I. Russell *, Yana Stoilova and Aura-Adriana Dosoftei

Department of Psychology, Middlesex University, London NW4 4BT, UK; y.t.stoilova@gmail.com (Y.S.); a.dosoftei@ucl.ac.uk (A.-A.D.)

* Correspondence: yvanrussell@gmail.com

Received: 11 August 2020; Accepted: 25 November 2020; Published: 30 November 2020

Abstract: "Image scoring" is a type of social evaluation, originally used in agent-based models, where the reputation of another is numerically assessed. This phenomenon has been studied in both theoretical models and real-life psychology experiments (using human participants). The latter are aimed to create conditions in the laboratory where image scoring can be elicited. One influential paper is that of Wedekind and Milinski (2000), WM. Our paper is a replication of that study, deliberately employing very similar methodology to the original. Accordingly, we had six groups of ten participants play an economic game. In each round, each player was randomly paired with another player whose identity was unknown. The participant was given a binary choice of either (1) donating money to that person, or (2) not donating money. In each round, the player was passively exposed to information about the past generosity of the other player. In our study, we successfully replicated the central result of WM. Participants in our replication gave significantly more money to partners with higher image scores (more generous reputations) than those with lower image scores (less generous reputations). This paper also provides a critical review of the methodology of WM and the study of image scoring.

Keywords: cooperation; reputation; psychology; replication

1. Introduction

"Reputation" refers to knowledge about the typical behavior of an observed individual [1–3]. A numerical index of reputation, called an "image score", was created in theoretical studies (agent-based modelling) that aimed to investigate the evolution of cooperation [4]. The image score scale ranged from –5 to +5 (including a neutral zero), allowing agents to accumulate a negative score for selfishness and a positive score for generosity. The motivation behind research in this area was the search for solutions to the "tragedy of the commons", a common phenomenon whereby a useful or necessary public good is diminished (perhaps fatally) because of overconsumption by self-interested parties [5]. The evolution of cooperation (that which would solve the tragedy) requires a complex and multi-faceted explanation [6]. Indirect reciprocity [7,8] is just one proposed solution to the tragedy of the commons, expressed in positive form as "A observed B help C, therefore A helps B" [2]: generosity becomes self-serving because the generosity itself might be rewarded [1,7]. A substantial number of agent-based models have been published [8] that explore the conditions that allow indirect reciprocity, notably that of Nowak and Sigmund [4], who found that a discriminating "image scoring" strategy (give only to those who have given to others) was an evolutionarily stable strategy in contrast to undiscriminating/selfish strategies. Following the Nowak and Sigmund model [4], Wedekind and Milinski [9] (hereafter, WM) published a short paper in *Science* titled "cooperation by image scoring in humans". This study converted the computer simulation into a psychology experiment, using actual human participants (instead of "agents" in a model). The results of this human study appeared to provide support for the conclusions of the Nowak and Sigmund model [4] because the human participants appeared to

act in accordance to the predictions of the model. In the twenty years since WM [9] was published, the general methodology of the human studies have become known as the indirect reciprocity game (IRG). In our new study, we adopt the basic design of the IRG. In doing so, we model our study design most directly on that of WM. Below, we describe the IRG as we implement it in our current study.

The IRG entails that a group of participants are brought into a group and given a sum of money. Then, the experimenter proceeds to engineer pairwise encounters between participants (a different pairing in each round). Upon pairing, the donor and receiver proceed to participate in a version of the dictator game [10], where the donor makes a unilateral decision about whether to donate money to the receiver. Participants never choose their own role (donor or receiver). The role of a player is never fixed, however, because it can change from round to round (often, the role is randomly scheduled by the computer program). Crucial to the design of the IRG is that the possibility of direct reciprocity is removed. Players cannot take revenge, nor reward a good deed. There are two ways that direct reciprocity is prevented. One way is to tell players that they will never be identifiably paired with the same player in reversed roles. Another way is that the identities of the other players are hidden. Even though participants are often sitting in the same room together, they do not know with whom they are being paired in a given round. Because direct reciprocity is not possible, players are consequently guided toward the perception that they are playing a string of one-shot games. Donors know nothing about the receiver except for a numerical measure of image score. The image score is an index of a donor's generosity in previous rounds. Players who choose to donate to another player are awarded one point. Players who decline to donate lose one point. Just like in theoretical models (e.g., [4]), a high score connotes generosity and a low score connotes selfishness. At the end of the game, participants are paid according to how much money is in their account. Participants had been allocated a sum at the beginning of the study. During the game, money was very likely added to the account (due to being generous) or subtracted from the account (due to being selfish). IRG is also called a "donor game" by some authors. In the following, some figures from WM are specifically mentioned. To avoid confusion with figures in our current paper, the figures from WM are prefixed by "WM-" (e.g., WM-Figure 2 refers to Figure 2 in their paper).

In the WM [9] version of the IRG, they tested seventy-nine human participants. Their main result was that reputation played a significant role in the participants' decisions about whom to reward. Participants who received more money had higher image scores (for more details see Table 1 and discussion below). This result was interpreted as a demonstration of the importance of reputation as an explanatory factor in why we cooperate. The WM paper [9] can justifiably be described as influential, because, to date, it has been cited more than 900 times. We carefully combed over these citations and found that the vast majority of the citations were brief and uncritical, used as theoretical background in literature reviews (e.g., in Ref. [11]). Among the 900+ citations, we found only nine papers [12–20] that we considered to be replications. All of them successfully replicated WM. Table 1 is a summary of these replications, compared against the original WM paper (Ref. [9]) and compared against our own replication (bottom row). Five of the replications were co-authored by the same people who did the original study [12–16] and four of the replications were from other groups [17–21]. There were two criteria that we used for including papers in Table 1. The first was that the replication directly cited WM [9]. The second criterion was that the replication adopted the basic design of the IRG (as described in the previous paragraph). This excludes empirical studies of indirect reciprocity which did not use the IRG (e.g., studies which used the Prisoner's Dilemma game, PDG, instead). As shown in Table 1, the replications of WM should probably be called "partial replications", or "conceptual replications". None of those studies used exactly the same methodology, nor the same analysis. This was true even for those with the same authors as the original [12–16]. Looking at the aims, methods, and results columns, it is clear that all replications were using the basic IRG design as a core procedure, but adding new conditions, additional new games (e.g., alternating IRGs with non-IRGs), and new hypotheses (e.g., Ref. [19] used the IRG paradigm to study strategic reputation building, something that WM had not investigated). In Table 1 (third and fourth columns), one can also see great variation in the

parameters. The mean sample size (refs. [12–20]; $n = 10$ studies) was 128.8 participants (std. dev. = 55.09), median 114, range 79–228. The mean group size (students playing together; $n = 134$ groups) was 8.45 participants (std. dev. = 3.18), median 7.00, range 4–16. The mean number of rounds of the IRG ($n = 10$ studies) was approximately 32 rounds (std. dev. ≈ 31.2) (it was not possible to calculate this precisely due to randomized game-endpoints in Ref. [18]). Also shown in Table 1 (fourth column) is the length of the history of interactions that allows the donor to evaluate the image score of the receiver. The history length increases as the game proceeds. At the beginning of the game, the history length is zero. As shown in Table 1, the studies took two main approaches. Some studies [14,16–20] censored the history length to a maximum number of rounds. In the studies that did this, the mean maximum history length was approximately five rounds. Other studies [12,13,15] appeared to maintain a history length through the duration of the game. Therefore, a study like Ref. [12], for example, had 16 rounds, allowing a history length of 0–15 rounds (caveat: some of the papers did not mention history length, and therefore the inference was made that the history consisted of all rounds minus one).

Research using the specific WM design lasted only a few years. Most of the studies in Table 1 were conducted around the early 2000s (this includes refs. [17–19], which were also conducted around the early 2000s despite being published years later). Only Ref. [20] was conducted much later than the rest. There have been more recent (post-2015) developments on the study of reputation and indirect reciprocity, but these have branched out into new and different methods (we review more recent papers in our Discussion below). In our laboratory, we chose to replicate the original WM [9] paper. We did this because all of the prior replications [12–20] lack the same analysis as WM. The main claim of WM is based on a somewhat unusual means of transforming the data. The raw image score of each player was converted into a "deviation" score. This measured how much higher or lower that player's image score was compared to the group average on the given round. WM justified this approach as useful "to correct for group and round effects" (Ref. [9], p. 851). No further explanation is given, but presumably this refers to group-specific and round-specific confounds that influence the rate of giving. Following this transformation, there was another unusual aspect, concerning the analysis which supports the key result. The unusual part is *not* the statistical procedure (a standard repeated measures ANOVA was used), but the way that the data were set up. The focus was on the donor's perspective. During the course of the game, the donor had a number of opportunities to donate money. In each round, the question for the donor was always a YES or NO. In a given round, the donor is paired with a receiver, and the receiver's image score is shown. The donor's choice is whether or not to donate money to that receiver. The theoretical expectation in WM [9] was that positive image scores will be rewarded (YES) and negative image scores punished (NO). However, in the game, the donor was free to violate this expectation. It was completely possible to unjustly decide YES for a negative image score and NO for a positive one. Whatever the case, WM splits the donor's data into two columns: mean image score for (1) recipients of all YES decisions, and (2) all NO decisions. We believe that they did it this way because it allowed the binary YES/NO choice to be applied directly to the image score as encountered by the donor in a given round. WM did not report the descriptive statistics for the YES column and the NO columns, but the repeated measures ANOVA showed that those who benefited from YES decisions had significantly higher image scores than those who suffered NO decisions. WM's approach, which focused on "type of donors' decision" (Ref. [9], p. 851) was quite different from that in the replications that followed. None of the replications [12–20] did it that way. Yet, all of them were inspired by WM and copied their basic IRG paradigm. In our replication of WM below, we did our best to adhere to the specifics of the original analysis. We felt it was important to confirm that the original is replicable.

Table 1. List of replications, plus original and current study. Abbreviations: WM = Wedekind and Milinski; IRG = indirect reciprocity game; PGG = public goods game; PDG = prisoner's dilemma game; CAG = competitive altruism game; UNICEF = United Nations Children's Fund.

Ref.	Author(s) (Year)	Overall n (Number of Groups × Group Sizes)	Rounds (Shown History)	Aims and Methods	Result
[9]	WM (2000) (*original study*)	79 (7 × 10, 1 × 9)	6 (0–5)	Human experiment to investigate effects shown in Ref. [4]. Each round gave the participant one opportunity to donate and two opportunities to receive. Independent variable was a corrected version of receiver image score, measured not as raw value but as deviations from the group mean in a given round.	" . . . the image score of the receivers who were given money... was on average higher than the score of those who got nothing" (Ref. [9], p. 851). Compared to more generous players, less generous players were more discriminating, giving more to recipients with higher image scores.
[12]	Milinski et al. (2001)	161 (23 × 7)	16 0–15)	IRG was used as control group in study about the usefulness of additional, "standing strategy" information (e.g., when a player had declined to donate in an instance where recipient was not deserving). Statistical unit was group, not participant.	Image scoring was successful (analogous results to Ref. [9]), but the standing strategy was not successful. Analysis based on donations to "NO" player (confederate who never donated).
[13]	Milinski et al. (2002)	72 (12 × 7)	16 (0–15)	The donor, after making the yes/no decision to donate, was given an additional question of whether to donate money to UNICEF. Reputational information was shown for both decisions. Each group had confederates: "always yes" and "always no" players.	Both types of generosity (giving to other players/giving to UNICEF) tended to be rewarded, including in results of a mock-election at the end of the game to vote for other players for student council.
[14]	Milinski et al. (2002)	114 (19 × 6)	20 (0–7)	Information about the generosity of other players was derived from a PGG which alternated with an IRG in first sixteen rounds. In the first treatment, the games alternated. In the second treatment, eight PG games were followed by eight IRG. Last four rounds consisted of PGG only.	Players who were more generous in the PGG received more money in IR game. Final PGG showed very high cooperation if players uncertain about whether future IRGs would occur. Showed that concern for constant reputation monitoring increased cooperation.
[15]	Wedekind and Brathwaite (2002)	114 (6 × 9, 6 × 10)	24 [1] (0–23)	Investigated the relation between direct and indirect reciprocity. Each group played three games: PDG, IRG, then PDG again.	Players who were more generous in the IRG received more money (as in Ref. [9]) but also in the subsequent PDG.

Table 1. *Cont.*

Ref.	Author(s) (Year)	Overall n (Number of Groups × Group Sizes)	Rounds (Shown History)	Aims and Methods	Result
[16]	Semmann, Krambeck, and Milinski (2005)	228 (19 × 12)	16 [2] (0–5)	In a similar design to Ref. [14], information about the generosity of other players was derived from a PGG which alternated with an IRG. Statistical unit was group, not participant, in most analyses.	Players who were more generous in the PGG tended to receive more money in the IRG, showing that reputational information transfers between games and groups.
[17]	Bolton, Katok, and Ockenfels (2006)	192 (16 × 12)	14 (0–1)	Had similar aims to Refs. [12,18]. Manipulated cost (high/low) and type of information. For the latter, they distinguished between first-order (image score) and second-order information (cf. standing strategy).	Contrary to Ref. [12], giving was highest in response to second-order information (compared to first-order and a no-information control condition). Giving is higher in the low-cost condition.
[18]	Seinen and Schram (2006)	168 (12 × 14)	90+ (0–6)	Had similar aims to Refs. [12,17]. Manipulated cost of giving (high/low) and information (information about past generosity in high/low cost condition, no-information about past generosity in high cost condition only. Number of rounds were designed to be unpredictable (min. 90, avg. 99).	Dependent variable was the fraction of helping behavior (from 0–1). More donations occurred in the information condition and when the cost was low. Players with best image score tended to receive more money. Individual strategies were partitioned into six categories.
[19]	Engelmann and Fischbacher (2009)	80 (5 × 16)	80 (0–5)	A study of strategic reputation building, comparing "public" (image score seen by all) and "private" (image score not seen) conditions. Half the participants had public scores in the first 40 rounds and private scores in the last 40 rounds (and vice versa for the other half).	Image scoring was successful. Contributions were higher when image scores were public compared to private, suggesting that participants altered their behavior in response to being observed. Analysis of individual results showed a mix of apparent strategies among players.
[20]	Sylwester and Roberts (2013) [3]	80 (20 × 4)	30 (6+)	A study of reputation building in both direct and indirect contexts. PGG was alternated with a IRG (for half the sample) and alternated with a CAG (for the other half) (there were two types of alternating scheme: "one-shot" and "iterated"). In the CAG, players could choose each other in advance for a directly reciprocal game. Analysis based on groups rather than individuals.	Image score was derived from the PGG (rather than in IRG play). The IRG was successful, but was not the main focus of the paper. Generosity was higher overall (including in PGG) for CAG than IRG. Participant showed an immediate reaction to the introduction of reputational incentives (causing them to contribute more).

Table 1. *Cont.*

Ref.	Author(s) (Year)	Overall n (Number of Groups × Group Sizes)	Rounds (Shown History)	Aims and Methods	Result
-	Russell, Stoilova, and Dosoftei (*current paper*)	60 (6 × 10)	12 (0–5)	Replication of WM [9], using the same horizon (0–5 rounds) of reputational information and the same group sizes, but smaller overall sample. Unlike in Ref. [9], but like Ref. [19], not every player played in every round. However, on average, the randomization allowed six recipient rounds per player, as in Ref. [9]. Image scores were displayed only to donor in a given pairing (like private condition in Ref. [19], but unlike the displays of most other refs).	The main result of WM [9] was replicated, using the analysis where image score was measured as deviations from the group mean in a given round, contrasting those who received a high amount versus a low amount of money. See main article for details of methodology and further analyses.

[1.] This excludes PDG rounds played before and after the IRG rounds. [2.] This includes six PGG rounds alternating with IRG rounds (included because PGG contributed to reputation). [3.] Earlier versions of this study are reported in Ref. [21].

We know that some readers may question the value of replicating such an old paper. Our response is that, in the context of the "replication crisis" in psychology [22], replications can be considered as opportunities to look forward as well as backward [23]: replications can lead to the development of new methods for paradigms that perhaps need to be rethought. Furthermore, each new replication adds a data point to quantitative analyses of the success/failure rates of replications [22]. Finally, we can also point to equivalent replication programmes in other areas of psychology. The 1963 Milgram [24] study of obedience is a good example. That classic paper has been replicated umpteen times as a means of probing the nature of obedience as thoroughly as possible [25]. Replications of Milgram [24] continue to the present day, utilizing the most modern methodologies and theoretical perspectives [26]. We think that the WM study [9]—like the Milgram study [24]—deserves to be explored further, allowing us, in future, to probe the effects of reputation as thoroughly as possible.

2. Results

In our replication, the main result of WM [9] was replicated. Data files and other material are viewable in the Supplementary Material (Documents S1–S12). In summary, we found that players provided YES decisions preferentially to receivers who had higher image scores in contrast to the receivers who received NO decisions. Our methods differed in several small ways from WM (see Methods for details), but produced directly comparable results. Figure 1 is modelled directly on WM-Figure 2. As shown, there were six groups of players. To conduct an equivalent analysis to WM (as described in our introduction), we measured image score as individual deviations from the mean image score for the group and round that a player is in. Thus, if the group mean for in a given round were +0.2, but an individual's raw image score were −1 in that round, then that individual's adjusted image score is −1.2. The overall mean for adjusted image score was 0.00 (std. dev. = 2.90). We then calculated two other variables: image scores of individuals who (1) received and (2) did not receive a donation in a given round (i.e., YES-donate, NO-donate). The first step of this calculation was to identify the recipient in every round and determine their adjusted image score at the beginning of the round (i.e., the image score perceived by the donor before making a choice). In the data file (Document S9) where the rows were participants, the image score of the recipient would be inputted into the row of the donor. For example, if participant 47 was the donor in round 10 and participant 46 was the receiver, then participant 46's image score was inputted into participant 47's row as the image score of the receiver from the point of view of participant 47 in that round. Thereafter, for every player, two variables were created that were the mean adjusted image scores of all recipients to which a player had (1) donated, or (2) not donated. The overall mean adjusted image score for recipients (from the point of view of the donor) was 0.92 (std. dev. = 0.97; range –2.00 to +2.90; n = 58) for those who received donations, and −0.46 (std. dev. = 1.47; range -4.20 to +2.10; n = 42) for those who were declined donations. Because this was a comparison of two separate variables, the analysis below included only those players who were seen to make both decisions (YES-donate; NO-decline donation) during their play. Players who were all-yes (n = 8) and all-no (n = 3) were therefore excluded. A repeated measures ANOVA (the same analysis as WM) was then conducted with donors as replicates using the two aforementioned variables (if the donor gives or not) as the within-subjects variables and with groups of individuals as between-subjects factors. The effect of image score in giving/not giving was significant, $F(1, 34) = 6.563$, $p = 0.015$, $\eta^2_p = 0.1618$, but there was no significant effect of group, $F(5, 34) = 0.346$, $p = 0.881$, $\eta^2_p = 0.0484$, and no significant interaction, $F(5, 34) = 1.093$, $p = 0.382$, $\eta^2_p = 0.1384$.

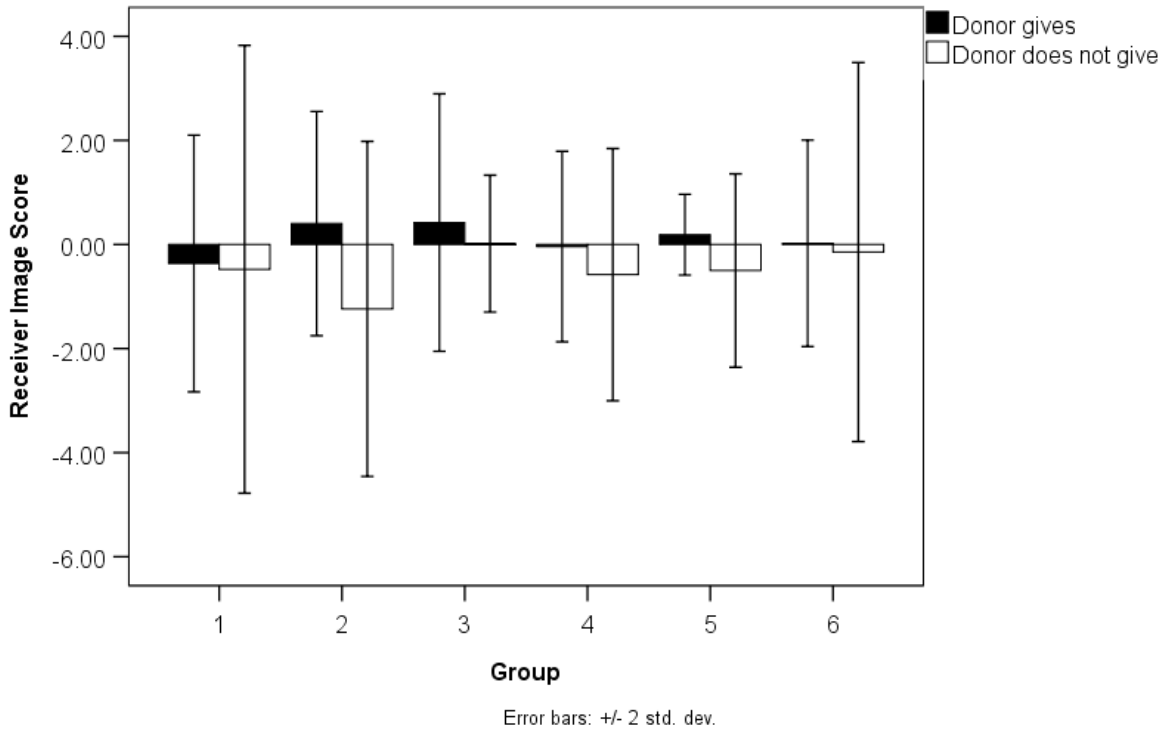

Figure 1. Image score of receivers when money given (black bars) or not (white bars).

We decided to perform an alternative analysis, to confirm that the WM result was not an artefact of their choice of analysis. WM had partitioned receiver's image score into two separate variables, according to whether they were (1) YES versus (2) NO decisions. In contrast, we decided to merge receiver's image score into a single variable. We constructed a logistic GLMM (generalized linear mixed model) using donor trial as repeated measures and a binomial distribution with a logit link. The donor trial referred to every instance that a player was assigned the donor role by the computer program (these varied across participants; see Methods). The opportunity to donate was randomly assigned, not fixed to a particular round. For example, looking at the data file, we see that participant one had five opportunities to donate and these occurred in rounds 2, 5, 7, and 8. Player two, in contrast, had four opportunities to donate and these occurred in rounds 3, 4, 10, and 12. All players had a different pattern in this regard. That is why the donor trial is differentiated from the game round. There was a significant positive correlation between receiver image score (original) and the donor trial, rho = 0.214, $p < 0.001$, which shows the accumulative nature of the image score (the same correlation was not significant using the adjusted image score). The binary target (dependent variable) was the yes-no decision about whether or not to donate money to the recipient in a given round (hereafter called "decision"). This new analysis required the construction of a new data file where each row in an SPSS file recorded a player's decision in a given round. This created a file with 330 decisions made by sixty players. A player's group membership and ID number was recorded in columns, allowing a data structure where participants were nested within their testing groups. The random effect was therefore the testing group ($n = 6$). In the datafile, we made two exclusions. One is that we excluded all decisions from the first round. We did this because it was too early to accumulate an image score in the first round. Second, we excluded all-YES and all-NO players (leaving $n = 49$). The number of exclusions was therefore the same as in the repeated measures ANOVA reported above. The results below were not significant unless these exclusions were applied. The fixed effect (independent variable) was the image score of the recipient in a given round. We ran the test using both versions of the receiver's image score (original and adjusted). The result was significant for the adjusted image score only, BIC (Bayesian Information Criterion) = 895.783, fixed effects $F = 4.767$, $df = 192$, $p = 0.030$, fixed co-efficient = 0.178 ± 0.082, $t = 2.183$, $p = 0.030$. The co-variance parameters

for each donor trial (from second to ninth opportunity) were: second (n = 41), β ± std. err. = 0.765, Z = 4.216, $p < 0.001$, third (n = 40) β ± std. err. = 0.923, Z = 4.431, $p < 0.001$, fourth (n = 38) β ± std. err. = 1.057, Z = 4.272, $p < 0.001$; fifth (n = 33) β ± std. err. = 1.183, Z = 4.013, $p < 0.001$; sixth (n = 25) β ± std. err. = 1.351, Z = 3.433, $p = 0.001$; seventh (n = 12) β ± std. err. = 1.679; eighth (n = 4) β ± std. err. = 1.024, Z = 1.395, $p = 0.163$; ninth (n = 1) β ± std. err. = 1.475, Z = 0.686, $p = 0.493$.

Next, we addressed another analysis from WM. In WM-Figure 3, they showed an analysis that suggested that players with lower image scores appeared to be more discriminating than players with higher image scores. Their y-axis was "image score of receivers who get something" and their x-axis was "number of donations (donor image score at the end)". As with our Figure 1 above, the image score was the adjusted one that measured the deviations per group and round. Using a Kruskal–Wallis test, they found that players who accumulated higher image scores tended to donate money to players with image scores significantly lower than those players who had accumulated lower image scores. We did the same analysis, using our YES corrected image score variable as the dependent variable, and a simple count of the amount of donor's donations as the independent variable. In contrast to WM, our result was not significant (Kruskal–Wallis H = 8.145, $df = 8$, $p = 0.419$).

The extent of payoffs vary widely across participants for a number of reasons. Because of randomization in the way that the computer program paired players, players had unequal chances to donate and receive (i.e., some players had more turns than others). The mean amount of donations was 4.20 (std. dev. = 2.02). The mean income at the end of the game was £10.28 GBP (std. dev. = 3.55), range £1.50–19.00 (for context, a typical 500 mL bottle of cola in nearby shops costed £1.00 at that time). Income was inversely correlated with the number of chances to donate, Pearson $r = -0.722$, $p < 0.001$, and positively correlated with the number of chances to receive, $r = 0.722$, $p < 0.001$. Also, income was inversely correlated with the number of donations actually made, Pearson $r = -0.521$, $p < 0.001$. Image score was calculated the same way as in Nowak and Sigmund [4]: starting at zero at round one; the allocation thereafter was +1 for every donation and −1 for every non-donation (e.g., a donation followed by two non-donations would yield a score of −1). Overall image score of all players at the end of the game was 1.98 (std. dev. = 2.66), range −5 to +5. The correlation between end income and (uncorrected) end image score (i.e., score at the beginning of round 12) was not significant, Pearson $r = -0.144$, $p = 0.273$.

3. Discussion

We successfully replicated the key result of WM, showing that, from the donor's perspective, the partner's image score played a role in the decision-making about whether or not to donate money to a given recipient. We took care to conduct the same analysis as WM did (compare WM-Figure 2 to our Figure 1). In an alternate analysis of the same data, we found that their main result held up. This shows that WM's results were not simply an artefact of their unusual analysis. However, we did not successfully replicate WM's results that appear to show that players with lower image scores are more discriminating in whom they choose to reward (WM-Figure 3). One possible explanation for this latter result is that players with higher image scores were more attentive to the image-scoring stimuli (or, conversely, that those with lower image scores were particularly inattentive). This is only conjecture on our part, and would need further study to ascertain why this particular result contradicts that of WM.

We acknowledge the limitations of our study. Despite our attempts to hew closely to WM's methodology and analysis, it was unavoidable that we deviated in a number of ways (see Methods for details). One obvious (but easily forgiven) deviation was with respect to the actual machinery used in WM (see WM-Figure 1). Their slightly antiquated physical set-up consisted of metal boxes, tangled wires, and light bulbs. Instead of attempting to reconstruct their non-computerized testing room, we used PCs, using the z-Tree software [27]. Another deviation (perhaps, less easily forgivable) is that we chose to differentiate our procedures from WM in terms of the number of trials, number of rounds, group size, and history length (see Table 1). That said, our study has several characteristics

that provided an analogous experience to WM. For example, our range of history length was the same as theirs. Players in WM had six opportunities to donate and this allowed the viewing of up to five previous interactions (image scoring points gained in the final round was not viewed by other players because the game was done). In our study, players had a variable amount of opportunities to donate, but the history length was capped at five. Another example of how we provided a roughly equivalent experience is that the player in our had roughly six opportunities to be a donor. In WM, there were six rounds where participants interacted with other players thrice (once as donor, twice as receiver). This is a different definition of "round" from follow-up studies (Table 1) where each round consisted of a single decision and the players stayed in the same role within a round (they did not switch roles until there was a new round). Therefore, WM's game had *de facto* eighteen rounds of play (if you adopt the definition of "round" used in our study and other studies in Table 1). In our study, all role designations were randomized, which caused a range of two to nine opportunities to donate, but with a population mean of six times (approximating the random six times of WM). Another parallel to WM is that our group sizes were the same. We had ten participants in every group (WM had one group with nine players, which we conjecture was not planned, but due to a no-show participant). Despite the similarities mentioned above, our study did differ in important ways. Notably, we had ~25% fewer participants then WM did. In Table 1, our study has the lowest number of participants. Obviously, a larger sample size would have provided better statistical power. Our sample size was not larger because we had significant budgetary and time constraints at the time that the study was run. However, the replication was successful, despite our smaller sample size.

Another thing to mention is that the computer program was not programmed to exactly match the parameters of WM. In creating our program, we obtained the file that was used to program Ref. [19] (generously donated to us by the creator of Z-tree). During programming, we decided to retain a number of the features in that program, rather than altering the program to the furthest extent to more closely resemble the original procedure in WM (see Methods for a summary of all differences between WM and our study). For example, there was no obvious reason to create the same system of rounds as WM did, when the round system used in Ref. [19] accomplished generally the same thing. However, one difference between WM and our study that we should particularly mention is that we did not display the image scores on a large screen for all players to see (as WM did). This made the image scores relatively private (only visible to the donor who is paired with that receiver on the given round). Another issue to mention is that we did not use names at all (a few studies, e.g., [12], used fake names, such as the names of planetary moons, to "identify" other players). These two issues (no public screen/no names) made the image score considerably more private than in WM and replications. Thus, our study resembled the "private" condition in Ref. [19]. Having said all of the above, we should note that exactness is not always the best way to judge a replication. As Stroebe and Strack [28] wrote: "A finding may be eminently reproducible and yet constitute a poor test of a theory" (p. 60). Accordingly, there is some validity in the argument that a conceptual (as opposed to a methodologically identical) replication is desirable, because it shows that the previous empirical support for an underlying theory was not merely a quirk that arose out of a particular operationalization [28]. We argue that the methodology in the present study is functionally equivalent to that of WM's, without attempting to recreate every aspect of their design.

Further on the topic of limitations, we should mention that, in our study, we could not guarantee that the players were strangers to each other, or to the experimenters. In fact, in many cases, we knew the participants personally and we knew that some players were friends with each other. These issues may have introduced a confound in that it introduces the possibility of a player behaving pro-socially due to a desire to achieve a positive self-presentation in front of friends or the experimenters [29]. However, we should mention three counterarguments against the idea that having non-strangers in the room would have greatly influenced our results. The first is that *most* people in the room were definitely strangers to each other, many having been recruited from outside the psychology department. Another counterargument is that all players' actions were anonymous. Personal identities

were never linked to the image scores in the dyad. All players knew that, on a particular pairing, they were playing with someone else in the room, but they did not know the identity of that person. A third counterargument is that all players knew it was a game. They had received some rather opaque instructions on how to earn money. It is possible that players did not construe not-giving as a "selfish" thing to do. The act of not-giving may have been considered a gaming strategy rather than a personality assessment. Thus, players who were sitting in a room with non-strangers may not have felt inhibited in adopting a competitive ("selfish") strategy. On the "game-playing" issue itself, we acknowledge the issue of external validity that accompanies laboratory experimentation. Do players in the game-world behave differently from players in real-world? If so, then do results in our study, even if internally valid, apply to real world image scoring? It has long been recognized [30] that behavioral laboratory experiments are subject to criticisms of their artificiality (when comparing laboratory versus field studies, the trade-off involves control versus real-life applicability). One way to make progress on the problem of external validity is to design future experiments that incorporate some new element of realism. For example, later in our discussion, we summarize new studies of image scoring that take into account the real-life imperfections of social perception (e.g., Ref. [31]). On the topic of "opaque instructions" (see Document S1 in Supplementary information) we had a situation where we had to facilitate image scoring without overtly directing the player to reward the observed generosity of others. We told them that they could earn money without telling them *how* (hence the opaqueness). Players were not instructed to pay attention to an "image score" (the term and its theoretical background was not mentioned at all, until the debriefing). If one scrutinizes our data file (e.g., Document S9 in Supplementary information), then it is quickly apparent that the image score was a hit-or-miss determinant of whether participants decided YES or NO. The data files show many instances of NO decisions on high image scores and YES decisions on low image scores. However, as our results show, the image score had a significant effect on YES/NO decisions *on average*. This result would suggest that we found the right balance in our instructions, nudging participants toward utilizing image score, but obscuring the importance of the image score in an opaque set of instructions.

There are further theoretical limitations to mention. For example, despite the success of our replication, we should note that we have demonstrated image scoring, but not the whole cycle of indirect reciprocity. In our study (as in WM), generosity did not pay off. In fact, those who earned the most were those who contributed the least. At first, this may seem to undermine the idea of indirect reciprocity (where the principle is that generous individuals prosper and selfish individuals suffer). However, we think that the reason that selfish players benefited is due to the short-term nature of the game. The procedure consisted of only twelve rounds, giving insufficient time for generosity to pay off. A longer-term game may have been different. There is also the depersonalized nature of the game. As mentioned above, players never knew exactly with whom they were playing. In this paradigm, reputation is stripped down to only one characteristic: their donation history. This was enough to show image scoring in the abstract, but it might be useful to conduct new versions of this experimental paradigm, where more personal relationships are established and using longer time frames. Indirect reciprocity has been shown to work in numerous theoretical studies (e.g., Ref. [4]), but it would also be worthwhile to explore future paradigms that provide naturalistic settings for their participants. Additionally, it would be useful to learn the participant's point of view in the game. Looking at our results, should we conclude that participants engaged in image scoring because of a genuine desire to reward goodness? Or, should we conclude that they engaged in image scoring as a deliberate strategy of profit-maximization? These questions relate to the longstanding debates in behavioral economics [32] about whether or not game players behave pro-socially because of some inherent "other-regarding" niceness (the alternative explanation being that apparent niceness is a by-product of selfish motivations). It is well-known that players spontaneously adopt a diversity of strategies when playing economic games [17,33–35] and it would be fair to state that the broad results of studies, as typically reported, are shrouding the heterogeneity of player strategies. It should be more commonplace to investigate such strategies, knowing that the micro patterns upwardly determine the

macro patterns. Another, important, means of assessing individual experience is to gauge sensitivity to parameter difference. Take the example of history length (cf. Table 1). In our study and others, history lengths varied widely, often in the 5–6 round range, but often higher. Some studies (e.g., Ref. [17]) go as low as one round of history length. In empirical research, some studies have shown that differences in history length can impose significant effects on game play, but other studies have not [36,37].

It should be noted that the original agent-based model by Nowak and Sigmund [4]—which inspired WM [9]—has been strongly criticized. Leimar and Hammerstein [38], for example, ran their own agent-based model that exposed weaknesses in Nowak and Sigmund's study. Their "island model" simulation produced conditions that favored a more sophisticated strategy called the "good standing" strategy, which allowed acts of non-cooperation to have no detrimental effect on one's image score if the prospective recipient was known to be undeserving of a donation (i.e., "A observed B refuse to give to C, therefore A refuses to give to B"). Subsequently, there have been real-life psychology experiments that have strove to demonstrate "standing strategy". Looking at two early studies (see Table 1), one found evidence for the standing strategy [17], but another one did not [12]. In a more recent and sophisticated study, Okada et al. [31] argued that, when thinking about the reputations of others, real-life decision-making does not consider all of the available information. In fact, they argue, researchers should focus on the reality of *selective inattention*: reputational information is multi-faceted and consists of different types of information (such as "what" data, the actions of player/donors, and "whom" data, the actions of recipients). Using an IRG-style paradigm, Okada et al. [31] presented participants with a number of different patterns of information prior to the decision about whether or not to donate. They found that participants were selective in the information that they utilized for making a decisions to donate to another recipient, attending to some information, but ignoring other information. On a similar topic, Hilbe et al. [39] published an agent-based model (a descendent of the Nowak and Sigmund [4] study) that explored the effects of "private, noisy, and incomplete information" (Ref. [39], p. 12241) on the processes of indirect reciprocity. In their results, they found that the introduction of errors caused previously known models of indirect reciprocity to become unstable. This model is an example of how tricky it is to increase the realism of agent-based modelling (i.e., a better approximation of the complexities of real life cooperation). Similarly, the psychology experiment of Duca and Nax [40] was an attempt to introduce more realism by exploring cooperative situations that go beyond pairwise interactions. They presented their participants with different variations on image scoring: besides the standard image scoring mechanism (as in WM), they had other versions such as group scoring (all members of the group get the same score) and self-scoring (where players manually assign scores to others). Their analysis found that the group methods were not an improvement over WM-style imaging score in terms of preventing the decay of cooperation. Judging from these recent studies, there are exciting possibilities opening up in research in regards to our understanding of how image scoring operates in the natural world.

4. Materials and Methods

Sixty naïve participants (25 males, 35 females, mean age 22.58, std. dev. = 3.96) were recruited in 2017 from students at Middlesex University (London, UK) and from personal contacts of the investigators. Participants were paid for their participation in varying amounts, contingent on the outcome of the game (see details below). The game was designed using z-Tree [27], version 3.5.1 (University of Zurich, Zurich, Switzerland, and Thurgau Institute of Economics, Kreuzlingen, Switzerland). A template program for an indirect reciprocity game (the one used in Engelmann and Fischbacher, [19]) was obtained from the creator of z-Tree. The hardware used in WM was unavailable (e.g., metal boxes, wires, and lightbulbs). The template program in z-Tree was adapted to fit the WM paradigm, but we decided to retain six features of the Engelmann and Fischbacher [19] study: (1) Participants were randomly paired in each round; (2) they were randomly allocated as donor or receiver within a round; (3) image score was calculated for the most recent five donation rounds only; (4) image score was displayed on the individual's screen rather than on a public board; (5) there were no cues to the

partner's identity at all (in the original study, players were labelled by number); and (6) there were twelve rounds instead of six. The purpose of choosing twelve rounds was that it gave every participant the possibility (subject to random variation) of donating six times and receiving six times (in the original study, participants had six definite chances to donate and six definite chances to receive, across six multi-event rounds). In our study, the computer program gave players a mean number of six opportunities to donate (std. dev. = 1.54; range 2–9), and six opportunities to receive (std. dev. = 1.54; range 3–10). Original data files are shown in Supplementary Material (Documents S1–S9). The Z-tree programming code is shown in Supplementary Material (Document S10).

University computer rooms were used, which consisted of ten PCs for participants and an eleventh PC for investigators. Screen sizes were 34 × 27 cm for participants and the classroom screen was 185 × 105 cm. Opaque white plastic dividers were placed in between adjacent players. Participants also received an information sheet, consent form, debriefing sheet, and payment with receipt. During the game, participants viewed a series of screens that provided information: current balance (in GBP), indication of game round (1–12), description of player's status in current round (donor/receiver), and all/some of the following information about player history: (1) If receiver in previous round, whether that player received money or not; (2) if donor in current round, then past donating behavior (e.g., "in the last 5 rounds as donor, you gave money 2 times and 3 times you did not give money"), and (3) if donor in current round, then donating history of prospective recipient (e.g., "The other person's record: In the last 5 times as donor, this person gave out money 3 times and did not give out money 2 times"). Also, for the donor, a choice was presented: (1) give £0.50 GBP to partner, or (2), nothing. If receiver, the participant saw a message to wait. Screenshots of the game are shown in Supplementary Material (Document S11).

Upon entry to the testing room, participants were asked to sit at designated computers and then read/sign an information sheet and consent form. The investigator then presented a Powerpoint slide show (duration ~3 min.) on the main screen that further explained the game: (1) There are twelve rounds; in each round, (2) the computer randomly chooses your role (donor/receiver) (3) and the computer randomly pairs you with someone else in room, (4) all identities hidden, (5) all players present in room, (6) if donor, choice to give (£0.50/nothing), (7) if receiver, wait for round to end, (8) multiplying rule (every donation multiplied by four, hence donor gives £0.50, recipient gets £2), (9) participants have £4.00 at start of game, (10), you could gain or lose money, (11) you are paid the money "won" at end of game. The aim of the game was stated as maximizing income, but no information was provided on how to achieve it. The slideshow is shown in Supplementary Material (Document S12). There were six groups of ten players. In each round, participants were designated either as donor or receiver and partnered with a player in the opposite role. This role/pairing alternated between rounds. After twelve rounds, the players saw a final screen indicating how much money they had "won", and then another screen prompting them to indicate age and sex. Participants were debriefed and then paid. There was no separate show-up fee.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4336/11/4/58/s1. Document S1: Explanatory document for data files. Documents S2–S7: Z-tree output files. Document S8: Main SPSS datafile. Document S9: Secondary SPSS datafile. Document S10: Z-tree programming code. Document S11: Some screenshots of game. Document S12: Slideshow presentation for participants at beginning of study.

Author Contributions: Y.I.R. designed the methodology, programmed the Z-Tree game, wrote the final manuscript, and performed the data collection, participant recruitment, and data analysis. A.-A.D. and Y.S. each wrote an undergraduate dissertation using these data, helped to collect the data and recruit participants, and contributed to the literature review used in the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: Y.I.R. received internal funding from the Psychology Department at Middlesex University.

Acknowledgments: We thank the editors of *Games* and three anonymous reviewers. Y.I.R. would like to thank his colleagues in the Psychology Department at Middlesex University (especially David Westley) and research assistant Dhana Letchmanan.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Milinski, M. Reputation, a universal currency for human social interactions. *Philos. Trans. R. Soc. B Biol. Sci.* **2016**, *371*. [CrossRef] [PubMed]
2. Russell, Y.I. Reciprocity and reputation: A review of direct and indirect social information gathering. *J. Mind Behav.* **2016**, *37*, 247–270.
3. Russell, Y.I. Reputation. In *Encyclopedia of Animal Cognition and Behavior*; Vonk, J., Shackelford, T.K., Eds.; Springer: London, UK, 2019; pp. 1–8.
4. Nowak, M.A.; Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **1998**, *393*, 573–577. [CrossRef] [PubMed]
5. Hardin, G. The tragedy of the commons. *Science* **1968**, *162*, 1243–1248.
6. Rand, D.G.; Nowak, M.A. Human cooperation. *Trends Cogn. Sci.* **2013**, *17*, 413–425. [CrossRef]
7. Alexander, R.D. *The Biology of Moral Systems*; Aldine de Gruyter: New York, NY, USA, 1987.
8. Okada, I. A Review of Theoretical Studies on Indirect Reciprocity. *Games* **2020**, *11*, 27. [CrossRef]
9. Wedekind, C.; Milinski, M. Cooperation through image scoring in humans. *Science* **2000**, *288*, 850–852. [CrossRef]
10. Leder, J.; Schütz, A. Dictator Game. In *Encyclopedia of Personality and Individual Differences*; Zeigler-Hill, V., Shackelford, T.K., Eds.; Springer: London, UK, 2018; pp. 1–4.
11. Russell, Y.I.; Call, J.; Dunbar, R.I.M. Image scoring in great apes. *Behav. Process.* **2008**, *78*, 108–111. [CrossRef]
12. Milinski, M.; Semmann, D.; Bakker, T.C.M.; Krambeck, H.-J. Cooperation through indirect reciprocity: Image scoring or standing strategy? *Proc. R. Soc. B Biol. Sci.* **2001**, *268*, 2495–2501. [CrossRef]
13. Milinski, M.; Semmann, D.; Krambeck, H.-J. Donors to charity gain in both indirect reciprocity and political reputation. *Proc. R. Soc. B Biol. Sci.* **2002**, *269*, 881–883. [CrossRef]
14. Milinski, M.; Semmann, D.; Krambeck, H.-J. Reputation helps solve the 'tragedy of the commons'. *Nature* **2002**, *415*, 424–426. [CrossRef]
15. Wedekind, C.; Brathwaite, V.A. The long-term benefits of human generosity in indirect reciprocity. *Curr. Biol.* **2002**, *12*, 1012–1015. [CrossRef]
16. Semmann, D.; Krambeck, H.-J.; Milinski, M. Reputation is valuable within and outside one's own social group. *Behav. Ecol. Sociobiol.* **2005**, *57*, 611–616. [CrossRef]
17. Bolton, G.E.; Katok, E.; Ockenfels, A. Cooperation among strangers with limited information about reputation. *J. Public Econ.* **2005**, *89*, 1457–1468. [CrossRef]
18. Seinen, I.; Schram, A. Social status and group norms: Indirect reciprocity in a repeated helping experiment. *Eur. Econ. Rev.* **2006**, *50*, 581–602. [CrossRef]
19. Engelmann, D.; Fischbacher, U. Indirect reciprocity and strategic reputation building. *Games Econ. Behav.* **2009**, *67*, 399–407. [CrossRef]
20. Sylwester, K.; Roberts, G. Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. *Evol. Hum. Behav.* **2013**, *34*, 201–206. [CrossRef]
21. Sylwester, K. The Role of Reputations in the Evolution of Human Cooperation. Ph.D. Thesis, University of Newcastle, Newcastle upon Tyne, UK, 2010.
22. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **2015**, *349*, aac4716. [CrossRef]
23. Rodgers, J.L.; Shrout, P.E. Psychology's replication crisis as scientific opportunity: A précis for policymakers. *Policy Insights Behav. Brain Sci.* **2018**, *5*, 134–141. [CrossRef]
24. Milgram, S. Behavioral study of obedience. *J. Abnorm. Soc. Psychol.* **1963**, *67*, 371–378. [CrossRef]
25. Burger, J.M. Replicating Milgram: Would people still obey today? *Am. Psychol.* **2009**, *64*, 1–11. [CrossRef] [PubMed]
26. Gonzalez-Franco, M.; Slater, M.; Birney, M.E.; Swapp, D.; Haslam, S.A.; Reicher, S.D. Participant concerns for the Learner in a Virtual Reality replication of the Milgram obedience study. *PLoS ONE* **2018**, *13*, e0209704. [CrossRef] [PubMed]
27. Fischbacher, U. z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* **2007**, *10*, 171–178. [CrossRef]
28. Stroebe, W.; Strack, F. The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* **2014**, *9*, 59–71. [CrossRef]

29. Christensen, L.B. *Experimental Methodology*, 5th ed.; Allyn & Bacon: Boston, MA, USA, 1991.
30. Schram, A. Artificiality: The tension between internal and external validity in economic experiments. *J. Econ. Methodol.* **2005**, *12*, 225–237. [CrossRef]
31. Okada, I.; Yamamoto, H.; Sato, Y.; Uchida, S.; Sasaki, T. Experimental evidence of selective inattention in reputation-based cooperation. *Sci. Rep.* **2018**, *8*, 14813. [CrossRef]
32. Binmore, K. Why do people cooperate? *Politics Philos. Econ.* **2006**, *5*, 81–96. [CrossRef]
33. Ledyard, J.O. Public goods: A survey of experimental research. In *The Handbook of Experimental Economics*; Kagel, J.H., Roth, A.E., Eds.; Princeton University Press: Princeton, NJ, USA, 1995; pp. 110–194.
34. Ule, A.; Schram, A.; Riedl, A.; Cason, T.N. Indirect punishment and generosity towards strangers. *Science* **2009**, *326*, 1701–1704. [CrossRef]
35. Swakman, V.; Molleman, L.; Ule, A.; Egas, M. Reputation-based cooperation: Empirical evidence for behavioural strategies. *Evol. Hum. Behav.* **2015**, *37*, 230–235. [CrossRef]
36. Camera, G.; Casari, M. Monitoring institutions in indefinitely repeated games. *Exp. Econ.* **2018**, *21*, 673–691. [CrossRef]
37. Kamei, K.; Nesterov, A. *Endogenous Monitoring through Gossiping in An Infinitely Repeated Prisoner's Dilemma Game: Experimental Evidence*; SSRN Working Paper; Durham University Business School: Durham, UK, 2020; pp. 1–53.
38. Leimar, O.; Hammerstein, P. Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. B Biol. Sci.* **2001**, *268*, 745–753. [CrossRef] [PubMed]
39. Hilbe, C.; Schmid, L.; Tkadlec, J.; Chatterjee, K.; Nowak, M.A. Indirect reciprocity with private, noisy, and incomplete information. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 12241–12246. [CrossRef] [PubMed]
40. Duca, S.; Nax, H.N. Groups and scores: The decline of cooperation. *J. R. Soc. Interface* **2018**, *15*. [CrossRef] [PubMed]