Citation for published version:
Franke, J & Papadopoulos, A 2024 'Strategic Reciprocity in a Contest with Large Stakes' Bath Economics Research Papers , no. 104/24, Department of Economics, University of Bath, Bath, UK.

Publication date:
2024

Link to publication

University of Bath

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 23. Jun. 2024

Strategic Reciprocity in a Contest with Large Stakes

Jörg Franke and Alexandros Papadopoulos

No.104/24

Bath Economics Research Papers 104/24

Department of Economics

Strategic Reciprocity in a Contest with Large Stakes*

Jörg Franke[1], Alexandros Papadopoulos[2]

[1] University of Bath (TU)
Department of Economics
Claverton Down
BA2 2AY Bath
United Kingdom

e-mail: j.franke@bath.ac.uk

[2] Department of Justice
HR Operations and Data Reporting
51 St. Stephen's Green
Dublin 2, D02 HK52
Ireland

e-mail: alexandrospapadopoulos00@gmail.com

April, 2024

Abstract

Using the unique properties of a German TV game show, we analyze the extent and implications of strategic reciprocity in sequential performance evaluations in a contest with large stakes. The sequential order of performances implies that the scope for strategic reciprocity differs systematically between participants: Contestants that perform later in the sequence evaluate their rivals before they are evaluated themselves, which creates incentives for strategic reciprocity. We find that earlier contestants benefit from this effect, resulting in a substantial negative sequence order bias. We provide estimates for the change in winning probabilities and for the financial implications of this bias.

Key Words: Reciprocity, contest, game show, shopping.

JEL classification: C57, D72, D9

*We would like to thank Debopam Bhattacharya, Santiago Oliveros, Javier Rivas, Andy Zapechelnyuk, Robertas Zubrickas and audiences at the University of Bath and TU Dortmund for helpful discussions.

1 Introduction

In the last decades experimental economists demonstrated that behavior of human subjects in laboratory experiments like ultimatum, trust and gift exchange games is systematically more cooperative and efficient than predicted by classical non-cooperative game theory based on self-centered payoff maximizing behavior, see Camerer (2003). In response to these observations, behavioral economists developed reciprocity-based modifications of non-cooperative game theory, where subjects prefer to treat others in a 'kind/unkind' way if they have been treated 'kindly/unkindly' by the respective individuals, compare Dufwenberg and Kirchsteiger (2004) or Falk and Fischbacher (2006) for original contributions and Sobel (2005) for a survey. By now, these modifications have been thoroughly analyzed and tested successfully in numerous experimental studies to demonstrate the improved predictive power of these behaviorally motivated modifications, see Hoffman et al. (1998), Fehr et al. (2002) and Malmendier et al. (2014) for surveys. However, finding direct empirical evidence for reciprocally-driven behavior in real-life contexts is comparatively scarce and typically determined in an indirect manner, for instance, using a combination of survey and experimental methods in a field context as in Finan and Schechter (2012) or Barr and Serneels (2009).

In this project we attempt to directly test for the existence and the extent of strategic reciprocity[2] in a real-life contest with large stakes that goes beyond the somewhat abstract and artificial environment of a laboratory setting. Using the unique feature of a weekly TV game show ('Shopping Queen'), where contestants compete against each other based on a sequence of consecutive performances and all contestants have to evaluate each others' performances, we find evidence for strategic reciprocity in the evaluation patterns among the contestants. In our empirical strategy we specifically exploit the sequential structure of individual performances and the fact that each contestant is evaluated immediately after her respective performance by all her rivals and, additionally, by a neutral expert. These properties imply that the scope for strategic reciprocity differs systematically between contestants depending on the position in the performance sequence. More specifically, contestants that perform later in the sequence must be aware that they themselves are more often evaluated after they evaluate their rivals performing before them which could induce reciprocal behavior by their rivals. In contrast, contestants that perform early in the sequence evaluate their rivals after their own performance, which implies that reciprocal reactions of their rivals are not relevant for them at the time they evaluate later

[2]By strategic reciprocity we refer to a situation where an individual might behave in a strategic (or instrumental, comp. Sobel (2005) for the relevant terminology) way in order to either induce future positive or avoid negative reciprocal behavior by their peers.

performances.

Hence, in this specific sequential contest strategic reciprocity would suggest that a contestant who performs later in the sequence is more likely to evaluate her rivals (who perform before her) in a more generous way either to induce positive or to avoid negative reciprocity, while a contestant that performs earlier in the sequence is not affected (when she evaluates later rivals, her own performance has already been evaluated previously). Overall, the implications of this hypothesized strategic reciprocity motive should benefit contestants that perform earlier in the sequence and be disadvantageous for contestants performing later in the sequence. Based on a theoretical analysis of the potential motivations of a contestant who decides which score to allocate to the performance of a specific rival, we derive three hypotheses regarding the respective evaluation patterns. The first two hypotheses are formulated as between-subject differences for contestants that perform on different days and specify that, firstly, contestants that perform later in the sequence should allocate more points in total to their rivals, while secondly, contestants that perform early in the sequence should obtain more points in total from their rivals. The third hypothesis is formulated as with-in subject prediction: Each contestant should allocate more points (on average) to those rivals who perform before them than after them.

Our empirical analysis comprises of 2240 evaluations from 560 contestants and confirms all three hypotheses, where we control for heterogeneity in individual performance quality by incorporating the objective performance measure of the neutral expert when appropriate. The results are robust for various econometric specifications, including non-parametric tests and different regression specifications. For the most extreme case (i.e. comparing a contestant performing on the first versus the last day), strategic reciprocity leads to an estimated total score difference of ca. 1.2 points based on an average score of 30.78 points received in total (roughly one third of a standard deviation), which translates to a ca. 5%-point difference in winning probability and a difference of ca. EU 50 in expected prize earnings. These differences are substantial especially when considering that performing contestants are not informed about the specific evaluation decisions of their rivals and are instead required to deduce imprecise signals regarding the generosity of specific rivals from the public discussion of their performance.

Our results are not only important in this specific instance of a highly popular TV game show, but in any situation where group decisions have to made in a sequential way. Consider for example a recruiting situation, where a committee has to determine the pool of candidates that should be shortlisted for interviews by sequential approval majority voting. In this situation committee members who will have a preference for candidates who are voted upon later in the sequence might vote for earlier candidates in the hope of inducing positive reciprocity from those

committee members that favor earlier candidates. The closely related concept of 'log-rolling' in sequential majority voting and elections, as well as other related contributions from the literature are discussed in the remainder of this section.

Related Literature

Our empirical approach utilizes data on observed behavior from participants in a TV game show. Similar approaches have been frequently applied in the past to test the predictive power of various economic theories and concepts; compare, for instance, Gertner (1993), Berk et al. (1996), Levitt (2004), List (2006), Thaler et al. (2012) and Camerer et al. (2015). As far as we are aware, the specific game show that we focus on has not been analyzed so far and combines several distinctive features in a unique way that facilitates the identification of strategic reciprocity in this specific context. These features are also related to different areas of the existing literature. Participants in this game show, for instance, are not only direct competitors for the prize but act, at the same time, as evaluators (or jurors) of each other because each participant has to allocate a score for the performance of each of her rivals. Hence, similarly to the classical sequential jury decision problem, there is an element of objective performance measurement involved and the question regarding the existence of a potential sequential order bias is of relevance.

Sequential order biases in jury decisions have been extensively analyzed in the recent past using data from various contests and competitions, e.g. song contests, sport tournaments, as well as research funding competitions. However, as juror members are typically not contestants in the competition, strategic considerations are mostly absent in this type of situations. This is also reflected in the empirical results from this literature. In contrast to the decreasing sequential order bias that we find in our contest (earlier contestants have a higher probability to win than later contestants), the literature on sequential jury decisions typically either finds an increasing sequential order bias where later (more recent) contestants are favored, see Glejser and Heyndels (2001), de Bruin (2005), Antipov and Pokryshevskaya (2017), Bian et al. (2022), or a so called J-shaped order bias, where the first and later contestants are favored, see Haan et al. (2005), Page and Page (2010) and Collins et al. (2019). While these empirical results seem to be robust with respect to the specific timing of evaluations (directly after each performance or at the end of the tournament) and the type of jury (amateur or professionals), the empirical approach in this literature is typically not able to control for individual performance quality due to the lack of an objective performance measure. Hence, the mentioned contributions are not able to differentiate between the two main explanations for the sequence bias; that is, endogenous objective quality differences in performances due to, for instance, learning, information acquisition and differences in motivations of the contestants during the course of the competition versus psycho-

logical biases of the jurors like primacy and recency effects or availability bias. In contrast, the unique features and properties of the game show that we consider allow us to control for individual performance quality (due to the fact that the neutral expert also evaluates each performance but not automatically in the same order) and to exclude primacy and recency effects because each contestant is evaluated by all her rivals immediately after each performance.

The fact that contestants in Shopping Queen also act as evaluators with antagonistic preferences regarding the contest outcome implies that there is some relation to the literature on sequential majority voting, specifically log-rolling. Originating from Tullock (1959), the situation of interest is a sequence of consecutive approval decisions about separated issues that are voted upon in a successive way through simple majority voting, where decision makers have idiosyncratic preferences about issues. As each decision-maker (elector) has one yes/no-vote for each issue, there is an incentive for vote trading ('log-rolling') in the sense that elector B might vote for issue A (even though she does not care about this specific issue) in exchange for elector A (who has a strong preference for issue A) voting later for issue B (for which elector A might not care about but elector B does).[3] Note, that this mechanism requires binding commitment power regarding vote exchanges, which is typically assumed for simplicity in theoretical contributions but does frequently not apply in real-world situations. Fischbacher and Schudy (2013) and Fischbacher and Schudy (2020) designed a stylized laboratory experiment on sequential voting decisions (involving three issues with three electors) without commitment power to analyze whether reciprocity can substitute binding commitments and therefore facilitate successful vote trading. The authors demonstrate that successful reciprocity-based vote exchange requires publicity regarding the individual voting decisions and that electors whose issues are voted on first are more successful (in the sense that their issues are implemented with a higher probability). Our empirical results are in line and therefore complementary to their experimental results as we also find that contestants who perform early in the sequence are favored which suggests (an intent to establish) reciprocal score exchange. Moreover, our approach goes beyond the simplified experimental setup in the sense that firstly, our analysis is based on a real-world contest with large stakes, and secondly, we find indication for strategic reciprocity even in a situation where contestants are not directly informed about the score (vote) that they receive from their rivals. Hence, for situations where voting behavior is actually less secretive, our empirical results could constitute a lower bound on the extent of strategic reciprocity.

[3]While Tullock (1959) as well as Buchanan and Tullock (1962) conjectured that log-rolling will lead to Pareto improvements in the voting outcomes (because it allows electors to express the intensity of their preferences), this hypothesis has been rejected theoretically by Riker and Brams (1973), among others; see also Casella and Palfrey (2019) for a recent theoretical contribution that synthesizes these different approaches.

The two contributions that are most closely related to our framework are Schüller et al. (2014) and Haigner et al. (2010), both based on data from the TV game show 'Come dine with me'. This game show has similar features as 'Shopping Queen' in the sense that five participants compete against each other on five consecutive days (on each day one participant has to prepare a dinner for their rivals) and that all participants have to evaluate the performance of each other. However, one important difference between the two game shows relates to the nature and the evaluation of the respective performance: While evaluating the quality of a prepared dinner in 'Come dine with me' is more likely to be influenced by personal and idiosyncratic preferences (tastes) of the participants/jurors, the evaluation of a purchased outfit in 'Shopping Queen' seems to be more objective and less taste-based. A straight-forward implication is that there is more scope for learning and information acquisition regarding the tastes of specific participants in 'Come dine with me', which benefits contestants who perform on later days.[4] This learning and information acquisition channel might therefore confound the countervailing effect of strategic reciprocity, which is in fact reflected in the empirical analysis: While Haigner et al. (2010) suggest that the first contestant is disadvantaged, Schüller et al. (2014) observe that later contestants are advantaged. Both results are in marked contrast to the decreasing sequential order bias that we observe in our study. These results suggest that in 'Come dine with me' learning and information acquisition is more pertinent and dominates the impact of the strategic reciprocity channel. In 'Shopping Queen' instead, learning and information acquisition is of less relevance, which (together with the fact that we can control for individual performance quality) allows us to associate the observed decrease in performance evaluations over the course of the competition with the strategic reciprocity motive.

The paper proceeds as follows. Section 2 describes and discusses the unique features of the TV game show 'Shopping Queen' in more detail. Section 3 introduces a theoretical framework to clarify the motivational channels that are relevant for a decision-maker in this context and derives three testable hypotheses on evaluation patterns. Section 4 reports results from the empirical analysis. Section 5 discusses the implications of these results and provides some estimates regarding the economic and financial impact. Section 6 concludes by suggesting some policy implications that should be relevant for any organization where sequential group decisions occur.

[4]For instance, participants who have to prepare their dinner on later days might become aware during the course of the competition that some of their rivals do not like fish-dishes and will therefore prepare a meat dish. This type of information advantage is more substantial for participants performing later in the sequence.

2 The Framework: Shopping Queen

'Shopping Queen' is a highly successful German TV game show, which is broadcasted on a daily basis since 2012 with more than 2500 episodes and a comparatively high market share ($3-10\%$) and audience numbers ($100,000-500,000$ spectators daily). The format of the show is a non-scripted reality contest, where each week five female candidates from a specific city compete against each other for a prize of 1000 EU. During the competition, each candidate receives 500 EU with the task to shop/create an outfit that fits to a pre-specified motto (e.g. 'Dress to impress! Find the perfect outfit for your first day at work!') which is publicly announced at the beginning of the week. On each day, one of the candidates is allowed to spend the amount in various shops of her choice to purchase a creative outfit in line with the respective motto. At the end of the day, the candidate has to perform her outfit on a catwalk in front of the other candidates. The outfit is then discussed in front of the performing candidate by all her rivals in a public discussion. After that, each rival privately evaluates the outfit of the respective candidate by allocating a score between 0 and 10 points. This score is initially concealed (i.e. neither of the other rivals, nor the performing candidate, nor the expert is informed about the specific score) but later publicly broadcasted in the corresponding episode. Hence, from the perspective of an evaluating candidate there is room for strategic behavior (because individual scores are private) but at the same time reputational concerns might be of relevance (because individual behavior might be scrutinized by the large public TV audience after the episode is broadcasted). At the end of the week, a neutral fashion expert evaluates the (recorded) performance of each contestant by also allocating a score between 0 and 10 points to each candidate. The winner of the contest is the candidate with the highest number of points (sum of scores by rivals and expert) who receives the title of 'Shopping Queen' and the prize of 1000 EU.

In our analysis we focus specifically on the evaluation stage, where a candidate has to decide which score to allocate for the performance of the respective rival on her performance day. Three potential channels might be of relevance in this situation: Firstly, there is a direct strategic effect from scoring because allocating a higher score to a rival directly affects the win probability of the evaluating contestant in a negative way. Hence, the strategic effect induces downward pressure on scores and in its extreme form would suggest allocating the lowest possible score of zero to each rival irrespective of the quality of her performance.[5] Secondly, there exists a countervailing reputation or neutrality effect in the sense that candidates might be either externally pressured

[5]Based purely on strategic considerations, the Nash equilibrium of this game would imply that all contestants allocate zero points to each other, which would presumably imply that the show would cease to exist in the future.

or intrinsically motivated to evaluate performances of their rivals in a neutral or objective way.[6] These two countervailing effects might be shaped by individual preferences and therefore differ between individual candidates; however, these preference differences should not be related in a systematic way to the specific sequence or order in which candidates perform.[7]

This order independence property, however, does not hold for the third effect, strategic reciprocity, which varies systematically depending on the respective position in the sequence of the performance. Strategic reciprocity becomes relevant when a candidate i, who evaluates a candidate j before being evaluated by j on a later day, anticipates that candidate j could behave in a reciprocal way. Specifically, candidate i might expect that j would allocate a higher score to i after having previously received a comparatively high score by i (positive reciprocity), while j is expected to allocate a low score to i after having received an unexpected low score by i (negative reciprocity), compare Dufwenberg and Kirchsteiger (2004) for a theoretical model along these lines. Hence, anticipating this behaviour by candidate j, candidate i might try to evoke positive reciprocity by choosing a comparatively generous score for rival j or at least avoid negative reciprocity by not choosing a low score for j. This implies that the strategic reciprocity effect induces upward pressure on the scores that candidate i allocates, restricted to those rivals j that will evaluate her afterwards. This effect therefore countervails the direct strategic effect further (in addition to the neutrality effect) but crucially depends on the order of performance. The effect becomes especially pertinent, for instance, for the candidate who performs at the last day because all her rivals have been evaluated previously by her and therefore are able to potentially reciprocate when evaluating her performance on the last day. In contrast, the effect is irrelevant for the candidate who performs on the first day as her performance has been evaluated always before she is called upon evaluating her rivals, making it impossible for her rivals to react to her scoring decision. For a candidate who performs on day 2, 3, or 4, strategic reciprocity implies that she should allocate comparatively more generous scores to those rivals who perform earlier than herself in contrast to rivals who perform later. Based on this discussion we can formulate three empirically testable hypotheses, that are backed up by a formal theoretical framework in

[6] External pressure can be exerted, for instance, during ex-ante interactions with the show producers but also through the fact that each candidate's scoring behavior is screened in all details to a very large TV audience (allocating unjustified low scores could be perceived as self-centered or greedy behavior that endangers the future of the show and potentially leads to a severe public backlash and reputation loss). The empirical analysis will reveal that scores at the lower bound are hardly ever observed which suggests that the strategic effect is effectively attenuated to a large extent by the reputation effect.

[7] While we were not able to find public information on how the order is determined by the producers, basic fairness consideration would require a random order. There is no indication that either candidates or the audience suspects systematic violations of fair randomization. In any case, the fact that there is an objective performance measure for each contestants (the score of the neutral expert) allows us to check and control for non-random assignment of order positions (see section 4.2).

the next section.

3 Theoretical Framework and Hypotheses

A basic theoretical framework of the scoring decision process allows us to distinguish the different motivations of a candidate when deciding a score for the performance of a rival and helps us to formulate the corresponding hypotheses. We use the following notation. Let $x_{ij} \in \{0, 1, \ldots, 10\}$ be the score or points that candidate i allocates to candidate j who performs on day j with $x_i = \left\{x_{ij}\right\}_{j \neq i}$ denoting the corresponding vector of scores. The total points given by candidate i are denoted by $X_i = \sum_{j \neq i} x_{ij}$, while the total points that candidate i receives by her rivals for her performance on day i are denoted by $X^i = \sum_{j \neq i} x_{ji}$. Similarly, we denote the score that the expert allocates to the performance of candidate i by $q_i \in \{0, 10\}$, which is the benchmark for an objective performance measure for each candidate. Note, that the winner of the contest is the candidate with the highest number $X^i + q_i$ of total points which includes the score of the expert.

The expected utility of candidate i is captured by the following function

$$u_i(x_i, x_{-i}) = u_i\left(P(x_i, x_{-i}), T\left(x_i, \left\{q_j\right\}_{j \neq i}\right)\right),$$

which depends on two terms: The first term $P(x_i, x_{-i})$ is defined as candidate i's probability to win the contest; that is, the probability that she obtains the highest number of total point:

$$P(x_i, x_{-i}) = Pr\left(X^i + q_i \geq \max\left\{X^j + q_j\right\}_{j \neq i}\right).$$

Obviously, this probability is decreasing in x_{ij} (i.e., $\frac{\partial P}{\partial x_{ij}} < 0$), but increasing in x_{ji} (i.e., $\frac{\partial P}{\partial x_{ji}} > 0$) for each $j \neq i$.

The second term is a measure of (aggregated) deviations of candidate i's allocated scores from the respective benchmark scores:

$$T(x_i, q) = T\left(\left\{\left|x_{ij} - q_j\right|\right\}_{j \neq i}\right),$$

where this measure is increasing in each individual entry (i.e., $\frac{\partial T}{\partial \left|x_{ij} - q_j\right|} > 0$).

Utility is assumed to be increasing in the win probability (i.e., $\frac{\partial u_i}{\partial P} > 0$), and decreasing in the deviation measure (i.e., $\frac{\partial u_i}{\partial T} < 0$) to capture the direct strategic and the reputation effect. We refrain from making more specific assumptions on functional forms to maintain the setup as general as possible.

Strategic reciprocity is captured indirectly in the following way. If candidate i anticipates that her rival j reacts in a reciprocal ways to x_{ij} (whenever j evaluates i after i evaluates j, i.e., whenever $i > j$), then x_{ji} becomes dependent on x_{ij} as follows:[8]

$$\frac{\partial x_{ji}}{\partial x_{ij}} \begin{cases} > 0 \text{ if } i > j, \\ = 0 \text{ if } i < j. \end{cases}$$

Consider now the derivative of the utility function of candidate i allocating x_{ij} points to candidate j on day j:

$$\frac{\partial u_i(x_i, x_{-i})}{\partial x_{ij}} = \underbrace{\frac{\partial u_i}{\partial P}\frac{\partial P}{\partial x_{ij}}}_{<0} + \underbrace{\frac{\partial u_i}{\partial T}\frac{\partial T}{\partial \left|x_{ij} - q_j\right|}\frac{\partial \left|x_{ij} - q_j\right|}{\partial x_{ij}}}_{\begin{cases} > 0 \text{ if } x_{ij} < q_j \\ < 0 \text{ if } x_{ij} > q_j \end{cases}} + \underbrace{\frac{\partial u_i}{\partial P}\frac{\partial P}{\partial x_{ji}}\frac{\partial x_{ji}}{\partial x_{ij}}}_{\begin{cases} > 0 \text{ if } i > j \\ = 0 \text{ if } i < j \end{cases}}$$

The first term on the right-hand-side captures the direct strategic effect. It is decreasing because allocating points to rival candidates decreases own winning probability. The second term captures the attenuating reputational concerns in the sense that the candidate prefers to be closely aligned (avoid deviations) with the objective performance measure of the expert, comp. Cialdini and Goldstein (2004), Bernheim (1994). The third term captures the countervailing effect of strategic reciprocity: Candidate i hopes to induce positive (or avoid negative) reciprocity from candidate j by allocating a higher score to j in case i is evaluated subsequently by j.[9]

This specification suggests that candidate i evaluates performances of rival candidates that happen before day i systematically different than those happening after day i due to the fact that

[8]Note that for the case $i > j$ this specification captures positive as well as negative reciprocity in the following sense. If x_{ij} falls short of a j's subjective expectation then further decreases in x_{ij} would lead to even more drastic decreases in x_{ji}, implying that x_{ji} is increasing in x_{ij}. If x_{ij} instead exceeds j's subjective expectation then further increases in x_{ij} would also lead to further increases in x_{ji}, implying again that x_{ji} is increasing in x_{ij}.

[9]Note that direct reciprocity (in contrast to strategic reciprocity) is not explicitly modelled in this setup. Experimental results from the strategically related trust game, where the amount sent from the sender to the receiver is typically multiplied by a factor of 2 or 3 implying a rate of return larger than 1 (which is not the case in our scoring game where the rate of return is equal to 1), suggest that the amount returned to the sender is typically lower than the amount sent to the receiver and decreasing in the rate of return. Moreover, the amount returned to the sender is also lower if subjects play both roles (which is the case in our scoring game), see Johnson and Mislin (2011). Both observations suggest, that in our scoring game direct reciprocity is either negligible or dominated by strategic reciprocity. In the latter case, our hypotheses could be modified accordingly which would make empirical verification more demanding. However, given that our empirical analysis yields significant results, the effect of strategic reciprocity might actually be stronger than our empirical results suggest.

strategic reciprocity is only relevant for a candidate that evaluates a rival before she is evaluated by the respective rival. Using these insights we can formulate the following three hypotheses:

Hypothesis H1: Candidates that perform later allocate more points in total than candidates that perform earlier: $X_i > X_j$ for $i > j$. Hence, X_i is increasing in day i of performance.

Hypothesis H2: Candidates that perform earlier receive more points in total than candidates that perform later: $X^i > X^j$ for $i < j$. Hence, X^i is decreasing in day i of performance.

Hypothesis H3: Candidates that perform on day $i = 2, 3, 4$ allocate more points to those candidates that perform before them than after them: $x_{ij} < x_{ik}$ for $j < i < k$.

The first two hypotheses are formulated as between-candidates comparisons because the above equation implies that the reciprocity concern is more or less pertinent depending on the respective performance day of the candidate. The last hypothesis focuses directly on within-candidate differences in performance evaluations depending on the relative order of the specific rival candidates in comparison to the performance day of the evaluating candidate (before-after comparison). The empirical analysis presented in the next section reveals that these hypotheses are mostly confirmed.

4 Empirical Analysis

Our data set consists of 2240 evaluations from 560 contestants in 112 contests as well as the corresponding 560 evaluations from the expert, extracted from public broadcasts of the game show. Moreover, we collected data on contestants' personal characteristics like age, height, weight, size and shoe size, which allows us to construct additional controls for the potential effects of individual characteristics on evaluations.[10] The summary statistics of all variables in our data set are provided in Table 1.

[10]During the contest each contestant is required to provide this personal information. Four contestants did not provide their personal weight information and one contestant did not provide any personal details.

Table 1: Summary Statistics

	Obs	Mean	SD	Min	Max
Points	2,240	7.70	1.24	0	10
Expert	560	7.41	1.22	0	10
Age	559	36.67	11.79	18	72
Height	559	168.73	6.52	148	189
Weight	556	63.56	13.03	40	150
Size	559	38.00	3.58	30	58
Shoesize	559	38.73	1.63	35	47

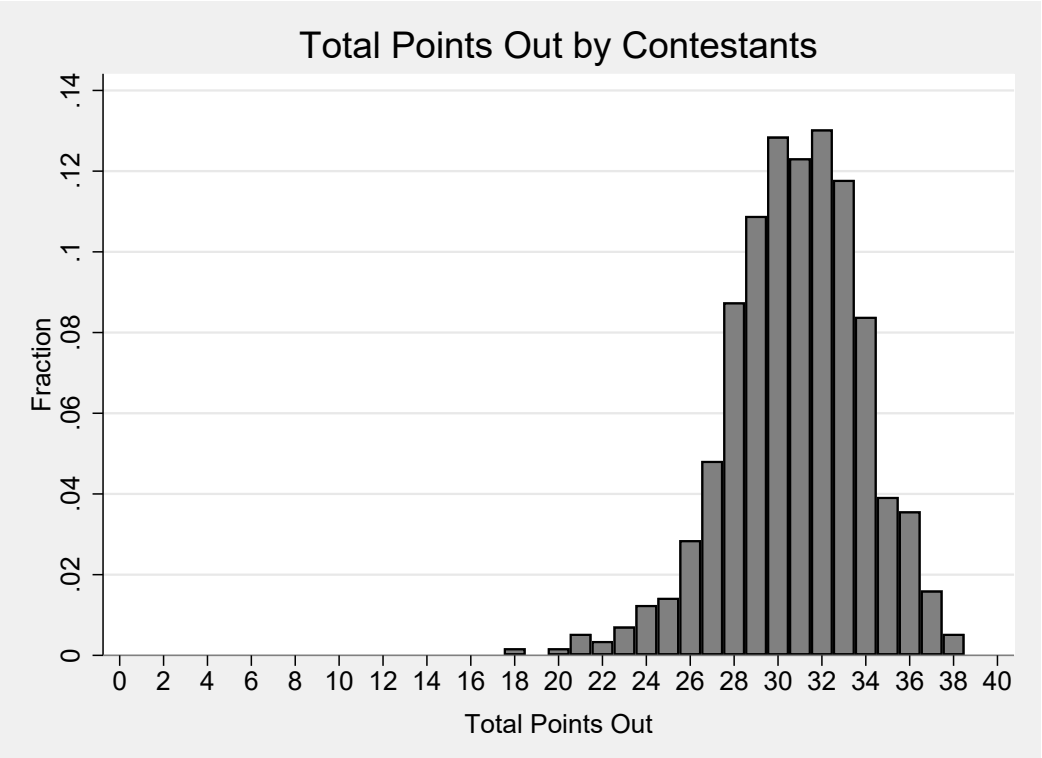

Figure 1: Histogram of Total Points Out by Contestant

Comparing the average score allocated by contestants ('Points') and the expert ('Expert') suggests that contestants evaluate each other on average more generously than the expert (the difference of 0.29 points is statistically significant at $p < 0.01$ significance level using a two-sided t-test as well as a non-parametric Wilcoxon signed-rank test), which could be a first indication for the existence of positive strategic reciprocity among contestants.

Figure 1 shows a histogram of the total number of points that each contestant allocated to all their rivals ('Total Points Out'). It should be noted that all observations are in the interior of the strategy space with none of the contestants allocating zero points to all of her rivals. This observation already suggests that the neutrality/reputation effect is effective in counterbalancing the strategic channel (which would imply allocating zero points to all rivals in order to maximize winning probabilities). We are now going to address the validity of each of the three hypothesis in the following subsections.

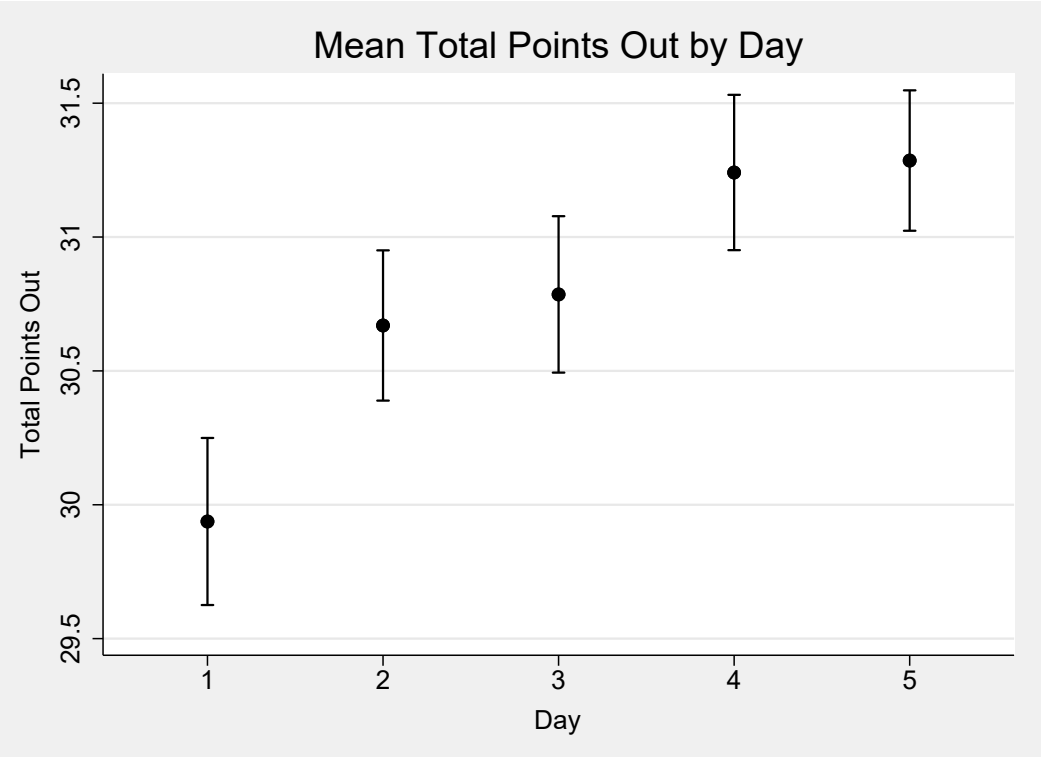

Figure 2: Total Points Out by Performance Day of Evaluating Players

4.1 Hypothesis H1

The statement in hypothesis H1 is formulated with respect to the total number of points allocated by a specific contestant to all their rivals. In this context strategic reciprocity implies that those candidates that perform on a later day in the performance sequence are more inclined to take reciprocal motives into account when evaluating others (because they are aware that their rivals will evaluate them later), resulting in a more generous score for their rivals. For candidates that perform earlier in the sequence, this effect is less relevant because when they are asked to allocate

a score to their rivals they have been already evaluated before. Hence, hypothesis H1 predicts an increase in average total points allocated ('Total Points Out') with respect to the performance day of the evaluating candidate. Figure 2 suggests that this is in fact the case when comparing the means of total points allocated by candidates that perform on different days.

To test directly for a linear increasing trend in the performance day we resort to a regression approach. Table 2 presents the corresponding results from an OLS specification (Models 1&2) as well as fixed-effects estimations to control for differences in evaluation patterns that are common within each weekly contest (Models 3&4).[11] Individual controls based on the personal characteristics are included in Models 2 and 4 but are mostly non-significant.[12] The variable of interest 'Day' is positive, statistically significant, and of similar size in all specifications. The results imply that a candidate who performs one day later allocates on average $0.27 - 0.33$ points more to their rivals, which confirms hypothesis H1. These results are also robust with respect to alternative specifications not reported here, for instance, including the score of the expert as additional explanatory variable or using additionally squared transformations of the control variables to allow for non-linear effects.

Table 2: Total Points Out

	1		2		3		4	
Day	0.3268	***	0.2761	***	0.3268	***	0.2879	***
	(0.0666)		(0.0733)		(0.0666)		(0.0717)	
N	560		556		560		556	
R^2	0.0226		0.0357		0.0473		0.0567	
FE	N		N		Y		Y	
Controls	N		Y		N		Y	

*** p<.01, ** p<.05, * p<.1

Note: Model 1 OLS, Model 2 OLS with Controls, Model 3 Fixed Effects, Model 4 Fixed Effects with Controls. All models are based on Total Points Out as dependent variable applying robust standard errors that are clustered on week/contest.

[11] Note that OLS and FE specifications in Models 1 and 3 lead to the same estimation for the variable of interest because 'Day' is the only explanatory variable in these regressions. We keep this specification in the table for completeness.

[12] The only exception is a positive coefficient for Age in Model 2 at $p = 0.080$, which becomes non-significant in Model 4. The non-significance of the individual control variables comes not unexpected and implies that personal characteristics of the evaluator like weight size, etc. do not impact her generosity of scoring others.

4.2 Hypothesis H2

The statement in hypothesis H2 is based on the total number of points received by a specific candidate. In this case, strategic reciprocity would imply that contestants performing earlier receive a higher number of points in comparison to contestants performing later, which leads to a decrease in total points received over performance days. Figure 3 (left) shows the average number of points received ('Average Points in') per day, which allows for direct comparisons with the corresponding evaluation by the expert ('Points by Expert') in Figure 3 (right). In contrast to the hypothesized pattern, the trend in average points received over performance days in Figure 3 (left) is rather non-monotonic which is also mirrored to some extent in the scoring pattern of the expert in Figure 3 (right). This similarity in the evaluation patterns of candidates and expert suggests that the non-monotonic trend is driven by differences in the objective quality of performances between different performance days.

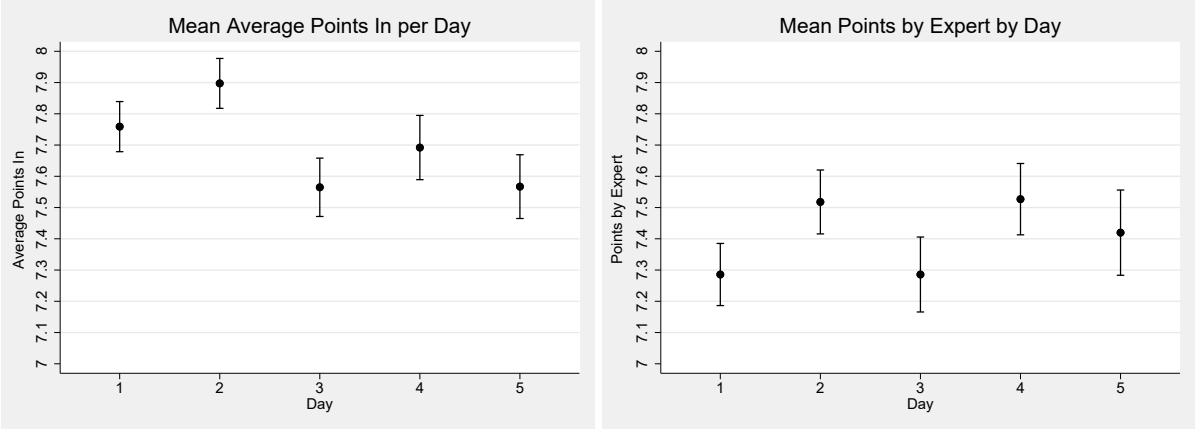

Figure 3: Average Points In by Contestants (left) versus Expert (right) by Day

We will therefore control for differences in the objective quality of individual performances by incorporating the corresponding score of the expert as a quality benchmark. In order to use the score of the expert as a neutral and objective quality benchmark, two properties must hold. Firstly, there should be some correlation between the score of the expert and the corresponding average score received by the respective candidate (which indicates that each performance has an objective quality dimension which is perceivable by experts and candidates), which is in fact the case in the data: The Pearson correlation coefficient of 0.505 is statistically significant at $p < 0.01$, which also holds for the non-parametric Spearman rank correlation. Secondly, the score of the expert should not be affected by the performance day or other individual charac-

teristics of the performing candidate (which indicates that the expert is in fact neutral and is able to evaluate performances in an unbiased way). Table 6 in the appendix presents the corresponding regression results from several specifications and demonstrates that the expert's score does neither depend on the performance day[13], nor on most of the personal characteristics of the performing candidates.[14]

Hence, we correct the average points received for each candidate by subtracting the respective benchmark score allocated by the expert. The resulting variable therefore describes for each candidate the difference between average points received by her rivals and by the expert. The corresponding graph which plots averages of these differences separately for each performance day is presented in Figure 4 and shows clearly a monotonic decreasing trend in line with H2.

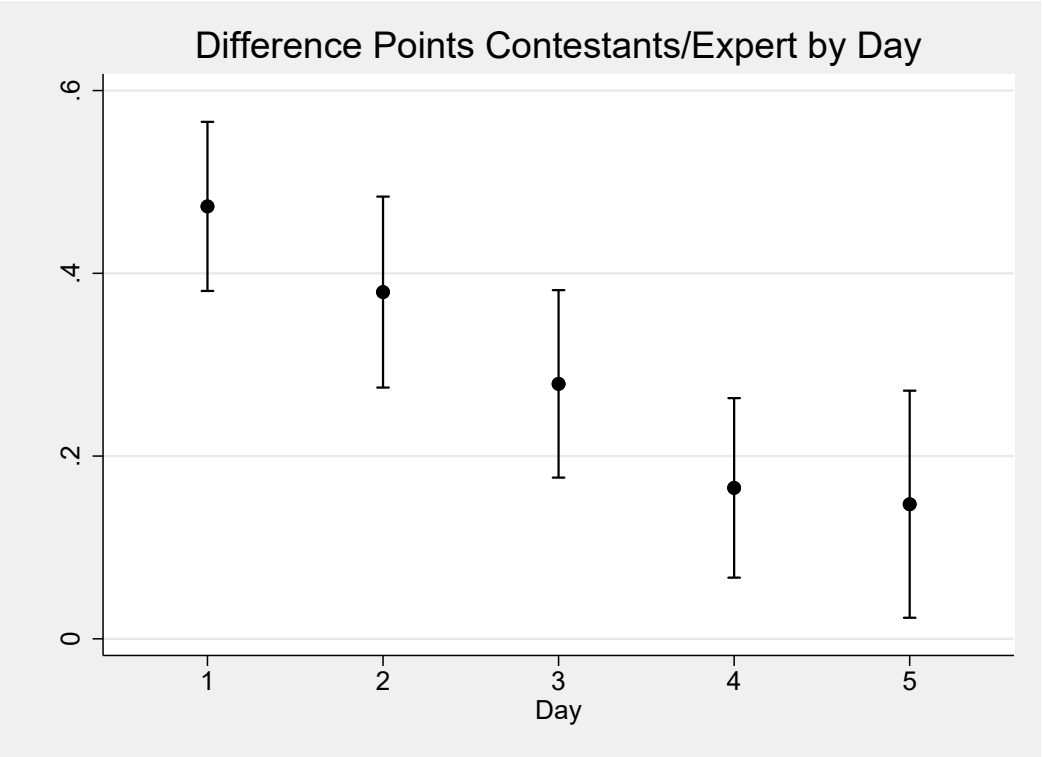

Figure 4: Difference between Average Points In by Contestants/Expert by Day

[13] Alternative specifications where variable 'Day' is substituted by indicator variables for the respective performance days yield some weakly significant day dummies in the respective version of Model 1, which however become non-significant in the corresponding versions based on Models 2-5.

[14] A positive and statistically significant ($p < 0.01$) 'Age' variable in Models 2 and 4 is the exception. In fact, there is anecdotal evidence that the expert is biased in favor of contestants of highly advanced age. This is also in line with the results in Model 3, where 'Age' and 'AgeSquared' are positive and significant, suggesting that the age bias of the expert might be driven by a low number of candidates with highly advanced age.

This negative linear trend is also confirmed in the regression results presented in Table 3: The respective variable of interest 'Day' is consistently negative and statistically significant in all specifications, ranging from -0.0614 to -0.866 (with a significance level of $p = 0.006$ in Models 1&3, $p = 0.052$ in Model 2 and $p = 0.054$ in Model 4). Hence, hypothesis H2 seems to be confirmed as well, although the significance is less robust with respect to the various specifications in comparison to hypothesis H1. A further observation of interest relates to the negative and statistically significant control variable 'Age' (at $p < 0.05$), which however can be explained by the specific preferences of the expert mentioned in footnote 14. All other control variables are non-significant, which implies that evaluations of contestants by their rivals are (as for the expert) not affected by individual characteristics of the respective performing candidates.

Table 3: Difference Points Contestant/Expert

	1		2		3		4	
Day	-0.0866	***	-0.0614	*	-0.0866	***	-0.0626	*
	(0.0311)		(0.0313)		(0.0311)		(0.0321)	
Age			-0.0099	**			-0.0091	**
			(0.0041)				(0.0043)	
Height			-0.0109				-0.0102	
			(0.0101)				(0.0104)	
Weight			-0.0021				0.0081	
			(0.0096)				(0.0102)	
Size			0.0381				0.0022	
			(0.0331)				(0.0322)	
Shoesize			0.0262				0.0009	
			(0.0453)				(0.0454)	
Intercept	0.5487	***	0.3397		0.5487	***	1.8899	
	(0.1064)		(1.6962)		(0.0933)		(1.6906)	
R^2	0.0121		0.0266		0.0180		0.0339	
N	560		556		560		556	
FE	N		N		Y		Y	
Controls	N		Y		N		Y	

**** p<.01, ** p<.05, * p<.1*

Note: Models 1&2 OLS, Models 3&4 Fixed Effects.

All regressions use robust standard errors that are clustered on week/contest.

While sections 4.1 and 4.2 focused on between-subject differences based on total points (allocated and received), we are now going to address with-in subject differences in allocated points to be able to test the validity of hypothesis H3.

4.3 Hypothesis H3

Hypothesis H3 is based on individual scoring differences for the subset of candidates that perform on day 2, 3, or 4. Note that these candidates evaluate some rivals before and some rivals after they have performed themselves. For these candidates the strategic reciprocity motive suggests that rivals should be evaluated differently depending on whether they have performed already or not. More specifically, candidates should evaluate those rivals more generously that perform before them because only those can reciprocate afterwards when it is their turn to evaluate the respective contestant. Figure 5 aggregates all scores allocated by this subset before (left data point) and after their own performance (right data point). It is obvious that there is a sharp and statistically significant decline of 0.27 allocated points ($p < 0.01$ using a two-sided t-test as well as a non-parametric Mann-Whitney test), which is in line with hypothesis H3.

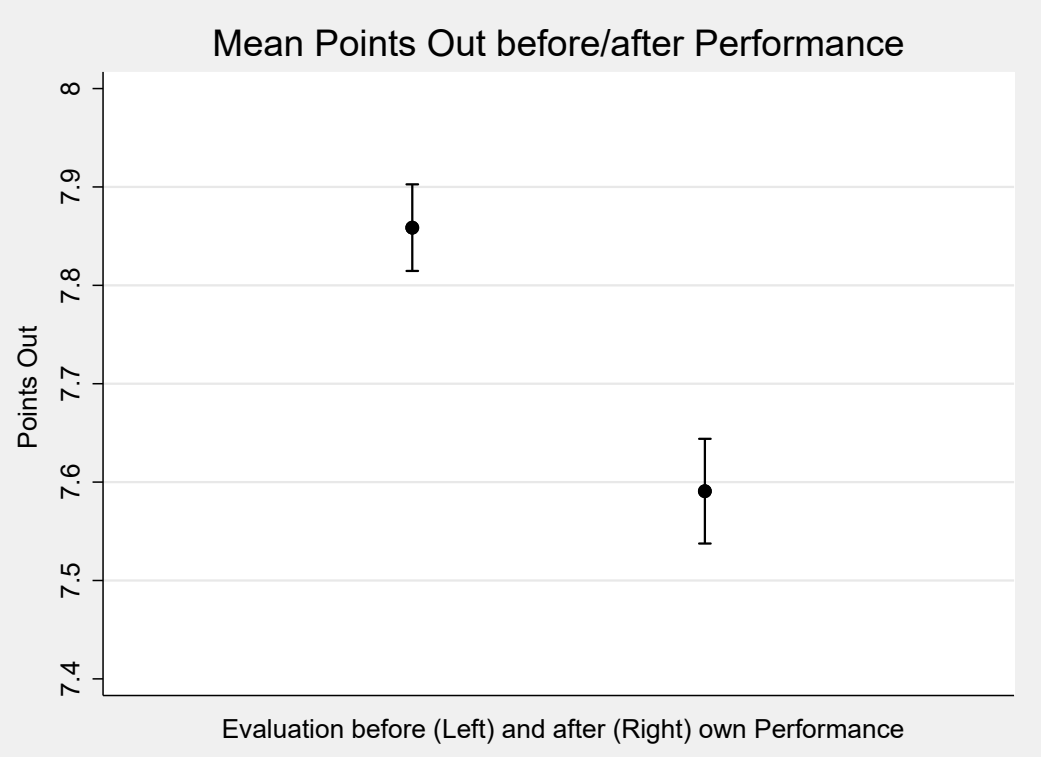

Figure 5: Points Out before/after own Performance

Table 4 presents regression results from various specifications to corroborate the insights from the graphical analysis. Models 2, 4 and 6 are based on the restricted subset of candidates that evaluate rivals before and after their own performance (i.e. that perform on day 2, 3 or 4).

For robustness checks the candidate set is extended to the entire pool in Models 1, 3 and 5. The variable of interest is 'Performed' which indicates whether the respective candidate evaluates a rival before ('Performed'=0) or after ('Performed'=1) their own performance. Table 4 shows that this variable is negative and statistically significant in all specifications ($p < 0.01$ in all specifications except Model 6, where $p = 0.014$), which confirms hypothesis H3. Table 4 also includes results from regressions using fixed effects on the evaluating contestant in Models 3 and 4 as well as on the evaluated contestant in Models 5 and 6. Using fixed effects controls for any heterogeneity in evaluations that are constant for the evaluating or the evaluated agent. This covers, for instance, quality differences in performances (which are constant when considering an evaluated contestant) or candidate-specific generosity (captured by the fixed effect on the evaluating candidate). Hence, score differences before and after own performances in Models 3-6 are purely identified through variations in evaluations based on the differences in the day of the performances between evaluators or evaluated candidates.

Table 4: Points Out

	1		2		3		4		5		6	
Performed	-0.2955	***	-0.2679	***	-0.2643	***	-0.2643	***	-0.3554	***	-0.2946	**
	(0.0522)		(0.0691)		(0.0688)		(0.0688)		(0.0483)		(0.1190)	
N	2,240		1,344		2,240		1,344		2,240		1,344	
R^2	0.0141		0.0111		0.0091		0.0140		0.0268		0.0077	
FE	N		N		Y		Y		Y		Y	
Day=2,3,4	N		Y		N		Y		N		Y	

*** p<.01, ** p<.05, * p<.1*

Note: Models 1 and 2 OLS, Models 3 and 4 Fixed Effects on Evaluator, Models 5 and 6 Fixed Effects on Evaluated Contestant. Models 2, 4, 6 restricted to contestants that perform on day 2, 3, or 4.

All regressions use robust standard errors that are, for Models 3 - 6 clustered on the respective FE variable.

5 Outcome Implications and Impact Analysis

While our empirical analysis confirms the predicted evaluations pattern as stated in the three hypotheses, the question of the overall magnitude and financial implications of this effect remains. This issue is addressed in Table 5 which presents results from two empirical specifications: Models 1&2 are based on ordered probit regressions with the rank[15] of the candidates as regression

[15]This ranking variable is constructed by ranking candidates with respect to the total number of points received by their rivals (i.e. without the points by the expert), which allows us to explicitly analyze the correlation between

variable to directly test for the effect of the performance day on the ranking position. Models 3&4 are based on probit regressions with a binary winning indicator dummy as regression variable to directly test for changes in winning probability. All specifications include the score of the expert as explanatory variable (which is positive, as expected, and statistically significant at $p < 0.01$ in all specification). Additional controls for personal characteristics of the evaluated candidate are included in Models 2&4.

Table 5: Ranking and Winning Probability

	1		2		3		4	
Day	0.0877	***	0.0839	**	-0.0728		-0.0682	
	(0.0319)		(0.0347)		(0.0502)		(0.0521)	
Expert	-0.3634	***	-0.3723	***	0.7328	***	0.7434	***
	(0.0409)		(0.0423)		(0.0782)		(0.0805)	
Age			0.0016				-0.0040	
			(0.0049)				(0.0065)	
Height			0.0017				0.0095	
			(0.0104)				(0.0147)	
Weight			0.0076				-0.0268	*
			(0.0082)				(0.0150)	
Size			-0.0449				0.0953	*
			(0.0319)				(0.0546)	
Shoesize			0.0254				-0.0315	
			(0.0497)				(0.0662)	
adj. R^2	0.0497		0.0528		0.2106		0.2205	
N	560		556		560		556	

*** p<.01, ** p<.05, * p<.1*

Note: Models 1 and 2 Ordered Probit with ranking position ($1, 2, \ldots, 5$) as dependent variable, Models 3 and 4 Probit with binary winning indicator as dependent variable. All regressions use robust standard errors that are clustered on week/contest.

The results from Models 1&2 reveal a positive and statistically significant effect of the performance day on the final ranking position of a candidate ($p = 0.006$ in Model 1 and $p = 0.016$ in Model 2). This is in line with the empirical results obtained so far because contestants that perform later in the sequence obtain fewer points (hypothesis H2) and grant more points to their rivals (hypothesis H1). Both effects imply that later performing candidates receive a higher (worse) rank with higher probability. In Models 3&4 the variable of interest, performance day,

the rank and the score of the expert. Modification of the ranking variable by including the score of the expert leads to qualitatively similar results.

has the expected positive sign but remains non-significant. We attribute this non-significance to the comparatively larger standard errors due to the loss in information induced by the use of this binary variable (regression variables in all other specification were less coarse and therefore contained more variability which allowed for more precisely estimated coefficients).

In order to provide a rough estimate for the overall impact of strategic reciprocity based on statistically significant estimations, the following indirect approach is taken: Firstly, we regress the binary winning variable on the total points received by rivals (plus the expert's score as additional explanatory variable as well as individual control variables). Results for these specifications are reported in Table 7 in the appendix. Secondly, based on these results the corresponding marginal effects for total points received can be calculated: Each additional point received increases the winning probability for an average candidate by roughly 4.16%-points (the mean of the three estimated coefficients: 4.32 in Model 1, 4.11 in Model 2 and 4.05 in Model 3, all marginal effects are statistically significant at $p < 0.01$). Thirdly, our results from section 4.2 imply that each additional performance day lowers the average points received by between $0.061 - 0.087$ points (see Table 3). Using the mean (0.074) of the four estimated coefficients from Table 3 as rough estimate for the decrease in points, the difference in average received points between a contestants that performs on the first and the last day results in roughly $0.074 \cdot 4 \approx 0.3$ points per rival or $0.3 \cdot 4 = 1.2$ points in total. This translates to a difference in winning probabilities of roughly $4.16 \cdot 1.2 \approx 5\%$-points or, in financial terms, a difference in expected payoffs of roughly 50 EU.

Alternatively, we can provide a rough estimate of the overall financial impact of strategic reciprocity by aggregating the differences between average points received by rivals and by the expert for the first four candidates who benefit from strategic reciprocity, weighted by the impact on probabilities derived before. Formally, the impact of strategic reciprocity (measured as hypothetical loss in aggregated expected payoff) can be calculated as follows:

$$\sum_{Day=1}^{4} 4 \cdot (0.074 \cdot Day) \cdot 0.0416 \cdot 1000 \text{ EU} = 123.14 \text{ EU}$$

Hence, the overall impact of strategic reciprocity corresponds to roughly 12% of the entire prize sum, which seems to be an economically relevant magnitude.

6 Concluding Remarks

Using the unique features of a TV game show with large stakes, our empirical analysis of observed scoring behavior identifies a pattern that is in line with strategic reciprocity of contestants with respect to each other. More specifically, contestants that perform later in the performance sequence are more generous when evaluating their rivals than candidates that perform earlier in the sequence. We attribute this difference in scoring behavior to the fact that later performing candidates are aware that they themselves are evaluated afterwards by their rivals which is not the case for earlier performing candidates. Hence, the strategic reciprocity motive is systematically more pertinent for later performing contestants in the sense that they are more likely to either induce positive or avoid negative reciprocity when evaluating their rivals.

We find additional evidence that is in line with the strategic reciprocity motive: Firstly, contestants that perform early in the contest sequence receive more points than their later performing rivals, and secondly, within-subject evaluations are more generous if the respective rival is evaluated before versus after the own performance of the evaluating candidate. Our empirical analysis validates these hypothesized scoring patterns in a statistically robust way and demonstrates that the overall effect of strategic reciprocity leads to a substantial bias in favor of contestants that perform early in the sequence. For a candidate who performs on the first versus the last day, this bias translates to an increased winning probability of roughly 5%-points, or alternatively, a difference in expected payoff from prize money of 50 EU. Aggregated over all five candidates in a given contest, the evaluation bias from strategic reciprocity therefore amounts to a substantial 12% of the entire prize sum.

It is conceivable that the producers of this game show spent substantial effort and thought in designing the evaluation procedure in such a way that strategic manipulations in scoring behavior are minimized. Concealed voting, no self-selection of contestants into the performance sequence, one-shot interactions and benchmarking through an external expert without personal stake in the contest (facing a difference performance sequence) are design features that should impede any type of collusive behavior among the participants. Our empirical analysis demonstrates that these measures have been at least partially successful: Candidates' evaluations are closely correlated with those of the neutral expert and are on average more generous which suggests that the direct strategic effect of allocating less points to rivals in order to further own winning prospects is under control. However, although concealed voting is implemented and contestants are only able to deduce or transmit reciprocal motives though a general public discussion of the respective performance quality of their rival, we still find evidence of a substantial bias in evaluation scores

due to strategic reciprocity.[16]

Our analysis also implies that the impact of strategic reciprocity in any type of organization where sequential decision-making (e.g. short-listing candidates or approval of projects in a committee) is implemented without these tight controls on performance evaluations is likely to be even more pertinent than our empirical results suggest. In this sense, the discussion and analysis of this game show and its specific design features might provide valuable lessons for any organization where some decisions are made by committees in a sequential order and which therefore faces similar challenges.

[16]Potential additional remedies that have the potential to ameliorate the evaluation bias from strategic reciprocity in this specific context could include the following: Increasing the weighting of the neutral expert in determining the winner of the contest, excluding the presence of the performing candidate in the discussion of her performance quality or abolishing the publish discussion in its entirety.

References

Antipov, E. A. and E. B. Pokryshevskaya (2017): "Order effects in the results of song contests: Evidence from the Eurovision and the New Wave," Judgment and Decision Making, 12, 415–419.

Barr, A. and P. Serneels (2009): "Reciprocity in the workplace," Experimental Economics, 12, 99–112.

Berk, J., E. N. Hughson, and K. Vandezande (1996): "The price is right, but are the bids? An investigation of rational decision theory," American Economic Review, 86, 954–70.

Bernheim, D. (1994): "A theory of conformity," Journal of Political Economy, 102, 841–877.

Bian, J., J. Greenberg, J. Li, and Y. Wang (2022): "Good to go first? Position effects in expert evaluation of early-stage ventures," Management Science, 68, 300–315.

Buchanan, J. and G. Tullock (1962): The Calculus of Consent, University of Michigan Press.

Camerer, C. F. (2003): Behavioral Game Theory, Princeton University Press.

Camerer, C. F., R. H. Thaler, D. van Dolder, and M. J. van den Assem (2015): "Standing united or falling divided? High stakes bargaining in a TV game show," American Economic Review, 105, 402–407.

Casella, A. and T. Palfrey (2019): "Trading votes for votes. A dynamic theory," Econometrica, 87, 631–652.

Cialdini, R. B. and N. J. Goldstein (2004): "Social influence: Compliance and conformity," Annual Review of Psychology, 55, 591–621.

Collins, A., J. McKenzie, and L. V. Williams (2019): "When is a talent contest not a talent contest? Sequential performance bias in expert evaluation," Economics Letters, 177, 94–98.

de Bruin, W. B. (2005): "Save the last dance for me: Unwanted serial position effects in jury evaluations," Acta Psychologica, 118, 245–260.

Dufwenberg, M. and G. Kirchsteiger (2004): "A theory of sequential reciprocity," Games and Economic Behavior, 47, 268–298.

Falk, A. and U. Fischbacher (2006): "A theory of reciprocity," Games and Economic Behavior, 54, 293–315.

Fehr, E., U. Fischbacher, and S. Gächter (2002): "Strong reciprocity, human cooperation, and the enforcement of social norms," Human Nature, 13, 1–25.

Finan, F. and L. Schechter (2012): "Vote-buying and reciprocity," Econometrica, 80, 863–881.

Fischbacher, U. and S. Schudy (2013): "Reciprocity and resistance to comprehensive reform," Public Choice, 160, 411–428.

——— (2020): "Agenda control and reciprocity in sequential voting decisions," Economic Inquiry, 58, 1813–1829.

Gertner, R. (1993): "Game shows and economic behavior: Risk-taking on "Card Sharks"," The Quarterly Journal of Economics, 108, 507–521.

Glejser, H. and B. Heyndels (2001): "Efficiency and inefficiency in the ranking in competitions: The case of the Queen Elisabeth music contest," Journal of Cultural Economics, 25, 109–129.

Haan, M. A., S. G. Dijkstra, and P. T. Dijkstra (2005): "Expert judgment versus public opinion: Evidence from the Eurovision song contest." Journal of Cultural Economics, 29, 59–78.

Haigner, S. D., S. Jenewein, H.-C. Müller, and F. Wakolbinger (2010): "The first shall be last: Serial position effects in the case contestants evaluate each other," Economics Bulletin, 30, 1–7.

Hoffman, E., K. Mccabe, and V. Smith (1998): "Behavioral foundations of reciprocity: Experimental economics and evolutionary psychology," Economic Inquiry, 36, 335–352.

Johnson, N. D. and A. A. Mislin (2011): "Trust games: A meta-analysis," Journal of Economic Psychology, 32, 865–889.

Levitt, S. D. (2004): "Testing theories of discrimination: Evidence from Weakest Link." The Journal of Law and Economics, 47, 431–452.

List, J. A. (2006): "Friend or foe? A natural experiment of the prisoner's dilemma," Review of Economics and Statistics, 88, 463–471.

Malmendier, U., V. L. te Velde, and R. A. Weber (2014): "Rethinking reciprocity," Annual Review of Economics, 6, 849–874.

Page, L. and K. Page (2010): "Last shall be first: A field study of biases in sequential performance evaluation on the Idol series," Journal of Economic Behavior and Organization, 71, 186–198.

Riker, W. H. and S. J. Brams (1973): "The paradox of vote trading," American Political Science Review, 67, 1235–1247.

Schüller, D., H. Tauchmann, T. Upmann, and D. Weimar (2014): "Pro-social behavior in the TV show 'Come Dine With Me': An empirical investigation," Journal of Economic Psychology, 45, 44–55.

Sobel, J. (2005): "Interdependent preferences and reciprocity," Journal of Economic Literature, 43, 392–436.

Thaler, R. H., M. J. van den Assem, and D. van Dolder (2012): "Split or steal? Cooperative behavior when the stakes are large," Management Science, 58, 2–20.

Tullock, G. (1959): "Problems of majority voting," Journal of Political Economy, 67, 571–579.

Appendix

Table 6: Points by Expert

	1	2	3	4	5
Day	0.0277	-0.0053	-0.0019	-0.0014	0.0012
	(0.0323)	(0.0348)	(0.0353)	(0.0356)	(0.0363)
Age		0.0120 ***	-0.0289	0.0100 **	-0.0237
		(0.0040)	(0.0250)	(0.0046)	(0.0259)
Height		0.0183	0.0232 *	0.0134	0.0065
		(0.0118)	(0.0121)	(0.0129)	(0.0134)
Weight		-0.0109	-0.0080	-0.0144	-0.0199
		(0.0126)	(0.0320)	(0.0112)	(0.0363)
Size		0.0106	-0.2659	0.0309	-0.2600
		(0.0420)	(0.2313)	(0.0388)	(0.2854)
Shoesize		-0.0516	-0.3542	-0.0540	0.4904
		(0.0489)	(0.9855)	(0.0547)	(1.0378)
AgeSquared			0.0005 *		0.0004
			(0.0003)		(0.0003)
HeightSquared			-0.0000 *		0.0000 ***
			(0.0000)		(0.0000)
WeightSquared			-0.0000		0.0000
			(0.0002)		(0.0002)
SizeSquared			0.0033		0.0037
			(0.0029)		(0.0035)
ShoesizeSquared			0.0039		-0.0070
			(0.0126)		(0.0133)
Intercept	7.3241 ***	6.1928 ***	17.6397	6.6282 ***	2.9137
	(0.1045)	(1.8695)	(17.3519)	(1.7531)	(18.0456)
N	560	556	556	556	556
R^2	0.0010	0.0214	0.0304	0.0222	0.0338

*** p<.01, ** p<.05, * p<.1

Models 1-3 OLS, Models 4-5 Fixed Effects. All regressions use robust standard errors that are clustered on week/contest.

Table 7: Winning Probability

	1		2		3	
Received	0.2357	***	0.2360	***	0.2376	***
	(0.0292)		(0.0298)		(0.0303)	
Expert	0.5206	***	0.5331	***	0.5455	***
	(0.0816)		(0.0838)		(0.0856)	
Day					-0.0683	
					(0.0522)	
Age			-0.0071		-0.0043	
			(0.0071)		(0.0074)	
Height			0.0122		0.0128	
			(0.0168)		(0.0169)	
Weight			-0.0227		-0.0218	
			(0.0170)		(0.0173)	
Size			0.0782		0.0749	
			(0.0612)		(0.0617)	
Shoesize			-0.0233		-0.0305	
			(0.0704)		(0.0713)	
Ps-R^2	0.3448		0.3535		0.3560	
N	560		556		556	

*** p<.01, ** p<.05, * p<.1

Note: Models 1, 2 and 3 Probit.

All regressions use robust standard errors that are clustered on week/contest.