Scand. J. of Economics 124(1), 278–300, 2022
DOI: 10.1111/sjoe.12451

Nationalistic bias among international experts: evidence from professional ski jumping*

Alex Krumer

Molde University College, NO-6402 Molde, Norway
alex.krumer@himolde.no

Felix Otto

University of Tübingen, DE-72074 Tübingen, Germany
felix.otto@uni-tuebingen.de

Tim Pawlowski

University of Tübingen, DE-72074 Tübingen, Germany
tim.pawlowski@uni-tuebingen.de

Abstract

Ski jumping competitions involve subjective evaluations by judges from different countries. This might lead to nationalistic bias, according to which judges assign higher scores to their compatriots. To test this claim empirically, we exploit within-performance variation of scores from all World Cup, World Championship, and Olympic Games competitions between the 2010/11 and 2016/17 seasons. Our findings confirm that judges assign significantly higher scores to their compatriots. The magnitude of this nationalistic bias is significantly higher in more corrupt countries. We do not find that judges assign significantly different scores to jumpers whose compatriots are present on the judging panel.

Keywords: In-group favoritism; judging panel; nationalistic bias; replication study; subjective performance evaluation

JEL classification: *D*71; *D*91; *Z*20

1. Introduction

Can well-trained and professional experts resist inherent preferences toward in-group members in their subjective evaluations? Do these experts use strategic motives when they evaluate in-group members of their counterparts? We try to answer these questions by studying the subjective

*T. Pawlowski is also affiliated with the LEAD Graduate School & Research Network. We thank international ski jumping judge, Ole Walseth, the Olympic medalist and former world record holder, Johan Remen Evensen, as well as Kjetil K. Haugen and Geir Oterhals, for providing valuable background knowledge on professional ski jumping.

evaluations of a panel of international experts who evaluate the performance of highly skilled professionals in real-life tournament settings with high monetary rewards.

In general, in-group favoritism based on the division of people into groups, according to some predefined rule, is a very well-established phenomenon. For example, Efferson et al. (2008) showed that even different signs on shirts were enough to divide people into groups and create in-group favoritism, according to which members of one group favor in-group members over out-group members. Thus, it is likely that in-group favoritism is one of the more primitive human instincts that developed during the evolutionary process (Sumner, 1907; Yuki, 2003), whose effects can even be observed in neurological processes in our brain.[1]

In-group favoritism has also been documented in various non-experimental settings. For example, Shayo and Zussman (2011) found that legal claims are more likely to be accepted if the judge and the plaintiff have the same ethnicity. Spierdijk and Vellekoop (2009) showed in-group favoritism based on geographical proximity in Eurovision Song Contests. Several other studies have shown evidence of favoritism in professional sport. For example, Price and Wolfers (2010) found that NBA players have fewer fouls called against them when their race matches that of the refereeing crew. Similarly, Pope and Pope (2015) demonstrated that referees favor their compatriot players by assigning them more beneficial foul calls in UEFA Champions League games. Very recently, Faltings et al. (2021) investigated Swiss soccer matches and showed that referees from one linguistic group assign significantly more yellow and red cards to teams from a different linguistic area.

In this paper, we build on the efforts of Zitzewitz (2006), who studied the subjective evaluation by judges in professional ski jumping based on data from 25 competitions in 2002. In these competitions, jumpers maximize their aggregate point score, which is determined by jumping distance (an objectively measured performance) and style points (a subjectively measured performance). His findings were striking: using within-performance (jump) variation of scores, he showed that judges assigned a significantly larger number of style points to their compatriot jumpers than the other judges who observed the same performance. In addition, Zitzewitz (2006) found a similar pattern of a nationalistic bias in figure skating competitions. Using a similar estimation strategy,

[1]As evidence, Andrews et al. (2019) tested the brain activities of fans from two rival teams who watched the same soccer game. They found a correlation between supporters of the same team in brain activities in areas that are known to be active in reward, self-identity, and control of movement. However, these brain activities were significantly different between the two groups of fans.

Sandberg (2018) showed an analogous result in dressage competitions. More recently, Scholten et al. (2020) used data from 41 World Cup ski jumping competitions, and provided suggestive evidence that nationalistic bias is still present. However, the number of events analyzed in their study is limited and the estimation strategy employed misses several key issues, making any comparison with the early findings by Zitzewitz (2006) impossible.[2]

We replicate and extend the analyses by Zitzewitz (2006) using data on 76,775 different evaluations of ski jumping judges from 203 competitions, comprising all the World Cups, Nordic World Ski Championships, and the Olympic Games between 2010/11 and 2016/17. Such an exercise seems highly relevant for two reasons. First, there is a growing consensus about the importance of replication studies in science.[3] Second, it seems highly relevant from a policy perspective to see whether problems that had been identified before have been solved over time.[4]

Comparing the score of a compatriot judge to scores of the other members of the panel within each jump, we find that compatriot judges assign close to 0.09 points more compared to their counterparts. This is equivalent to 29 percent of the within-jump standard deviation of scores, a non-negligible difference. As such, the nationalistic bias in professional ski jumping is remarkably persistent and still exists more than a decade after the initial findings by Zitzewitz (2006), which were also featured in the media.[5]

Further analysis suggests that the nationalistic bias is higher in more corrupt countries. Out of the 12 most observed countries in our data, only Norway and Finland had negligible estimates of a nationalistic bias,

---

[2]Scholten et al. (2020) neither exploited the within-performance (jump) variation of scores nor controlled for competition fixed effects. Moreover, they did not investigate the possibility of a compensating bias, according to which judges assign different scores to jumpers whose compatriots are present on the judging panel.

[3]For example, Open Science Collaboration (2015) replicated the results of only 36 out of 100 experimental and correlational studies that were published in top academic journals in psychology. In the same vein, Silberzahn et al. (2018) showed a high variance in the results of 29 scientific teams that investigated the same dataset, highlighting the importance of crowdsourced research that "can balance discussions, validate scientific findings and better inform policymakers" (Silberzahn and Uhlmann, 2015, p. 190). Finally, Ioannidis and Doucouliagos (2013) discussed the empirical evidence on the lack of a robust reproducibility culture in economics and business research. Therefore, replication of original findings is an important scientific task.

[4]For instance, Pope et al. (2018) performed a follow-up study to Price and Wolfers (2010) and showed that the racial bias among NBA referees disappeared after widespread media coverage.

[5]For example, the findings of Zitzewitz (2006) were summarized and discussed in the article "How ski jumping gets Olympic judging right (and figure skating gets it wrong)" by E. Zitzewitz, in the *Washington Post* on 12 February 2014 (https://www.washingtonpost.com/news/monkey-cage/wp/2014/02/12/how-ski-jumping-gets-olympic-judging-right-and-figure-skating-gets-it-wrong/).

both statistically and economically. In contrast, Russian judges assigned, on average, 0.22 points more to Russian jumpers than the other judges on the panel.[6]

Finally, we test whether there is evidence of strategic voting, according to which judges assign different scores to jumpers whose compatriots are present on the judging panel. The evidence of such a strategic voting is mixed. On the one hand, Zitzewitz (2006), who coined the term "compensating bias" for that phenomenon, found that, for some specifications, the ski jumping judges assign significantly lower scores to jumpers if the other judge on the panel is a jumper's compatriot. On the other hand, Sandberg (2018), who used the term "indirect bias", and Zitzewitz (2006) found an opposite result for dressage and figure skating, respectively. We do not find evidence for compensating bias. Among other factors, this discrepancy might be explained by differences between our study and the previous studies in terms of dealing with home advantage in the analyses. In fact, when controlling for the home variable, we find that the compensating bias loses most of its magnitude and becomes insignificant, both statistically and economically.[7]

The remainder of the paper is organized as follows. In Section 2, we describe the institutional settings of ski jumping competitions. In Sections 3 and 4, we present the data and descriptive statistics, and the empirical strategy, respectively. In Section 5, we present the baseline results, while we explore effect heterogeneity in Section 6. In Section 7, we compare our results with the results in other studies. We offer concluding remarks in Section 8.

2. Ski jumping rules

Ski jumping is a sport in which athletes ski down a track to generate speed and then jump from a ramp, with the aim of maximizing the length of the jump and the style points awarded by a judging panel. Three different hill sizes (HS) are used in professional ski jumping events: normal hills (HS 85–109 m), large hills (HS 110–184 m), and flying hills (HS 185 m and larger). Usually, 50 competitors jump in the first round. In flying hills, this

[6]This result adds to the previous finding of Elaad et al. (2018), who showed that the more corrupt the country, the higher the probability that a team will achieve the desired result in order to avoid relegation in the last soccer game of a season. It also relates to Fisman and Miguel (2007), who found that United Nations diplomats living in New York who represent governments from very corrupt countries accumulated significantly more unpaid parking violations than their counterparts from less corrupt countries.

[7]See Section 7 for a detailed discussion on differences between our findings and findings in ski jumping, figure skating (Zitzewitz, 2006), and dressage competitions (Sandberg, 2018).

number is reduced to 40. Based on the results of the first round, the top 30 jumpers advance to the second round. The winner of the competition is the jumper with the highest number of aggregate points achieved in both rounds.[8]

The aggregate point score is determined by the jumping distance and the style points. The jumping distance is an objective performance measure and quantified in intervals of 0.5 m. This distance is converted to a point value that contributes to the aggregate score. In addition, there is a subjective performance evaluation by a judging panel. The panel consists of five judges from five different countries, one of which is always the host nation. These judges award style points for the execution of the jump, landing, and outrun, based on predefined judging criteria for each part of the jump. Each judge awards a score of between 0 and 20 points, with intervals of 0.5. The lowest and highest scores are truncated to exclude extreme votes. The three remaining scores are summed up to the total style points. The athletes also receive compensation points for the starting gate and wind conditions to make the competition safer and fairer.

The judges of the panel are considered highly skilled and professional experts in this task. They are selected by the international governing body for winter sports, the Fédération Internationale de Ski (FIS). Judges must have a minimum of three years of experience at the national level, followed by a qualification period of at least two additional years. After the successful completion of the practical examination, the candidates receive their license to judge international ski jumping competitions. Moreover, ongoing training and an annual certification program is required to keep the status as an officially licensed judge (FIS, 2017b).

The judging process is designed to ensure the independent and discrete decision-making of the panel. According to the rules of the FIS (2017b), the athletes' performances must be evaluated objectively and without any prejudice. No communication with others is allowed and the decision must be entered into the scoring system without any assistance. Moreover, the judging tower where the judges are located is constructed in a way that provides optimal conditions for executing the judging task and ensuring compliance with the rules. More specifically, the tower is located at the side of the jumping hill such that each judge can clearly observe all parts of the jump. In addition, each judge has their own compartment in the judging tower so that they cannot view the scores of the other judges or be distracted by others.

[8]At World Cup competitions, the top 30 athletes receive World Cup points and prize money. For each World Cup point, the jumpers receive 100 Swiss francs (CHF), which amounts to 10,000 CHF for the winner of the competition. Extra prizes are awarded for special competitions such as the Four Hills Tournament (see FIS, 2017a, for additional information).

Table 1. Sample size

Number of ski jumpers	268
Number of ski jumper countries	24
Number of judges	172
Number of judge countries	19
Number of total competitions	203
Number of World Cups	165
Number of Four Hills	28
Number of Nordic World Championships	8
Number of Olympic Games	2
Number of jumps (performances)	15,355
Average number of jumps per athlete	57.29 (81.70)
Average number of jumps per athlete and season	17.73 (17.28)
Number of style point scores	76,775
Average number of scores per judge	446.37 (291.87)
Average number of scores per judge and season	143.24 (66.41)

Note: Standard deviations are presented in parentheses.

3. Data and descriptive statistics

We collected data from the official website of the FIS on all men's World Cups, Nordic World Ski Championships, and Olympic Games (in Sochi 2014) for the seasons between 2010/11 and 2016/17. These are the most prestigious tournaments in professional ski jumping. This period was selected because of the introduction of wind and gate compensation points in the 2010/11 season.

For each jump, we have full information on athletes' names and nationality, competition date, and hill characteristics. We also have information on the judges' names and nationalities, as well as the individual judges' style point scores for each jump.

As summarized in Table 1, the data include performances of 268 jumpers from 24 countries, evaluated by 172 different judges from 19 countries, covering 203 competitions. Overall, we have information on 15,355 jumps, each of which was evaluated by five different judges, totaling up to 76,775 different jump evaluations. As described in Table A1 in the Appendix, the competitions took place in 14 countries. Events were most frequently held in Norway and Germany, with 38 competitions in each country. German judges were part of the panels in 71 percent of competitions, followed by Norwegian judges (59 percent) as the second most frequent country.

Table 2 provides the summary statistics for subsamples considering whether a judge and jumper are from the same country or not. On average, judges assign a higher score to their compatriots. In 9 percent of cases (6,941 evaluations overall), a judge was a compatriot of the evaluated

Table 2. Descriptive statistics

	Ski jumpers	
	Compatriot jumpers	Non-compatriot jumpers
Style points		
Mean	17.61	17.48
(overall SD)	1.05	1.07
(within-jump SD)		0.31
Min–max	5.0–20.0	4.0–20.0
Compatriot on panel		
Mean	0	0.40
Home event		
Mean	0.30	0.12
Jumping distance		
Mean	133.59	131.42
(overall SD)	31.43	30.24
Min–max	55.0–251.5	51.0–251.5
Wind points		
Mean	−0.92	−0.87
(overall SD)	8.30	8.44
Min–max	−34.9–43.4	−34.9–45.7
Gate points		
Mean	0.18	0.18
(overall SD)	4.56	4.53
Min–max	−29.4–45.2	−29.4–52.7
Country CPI score (2012–2017)		
Mean	72.90	70.82
(overall SD)	13.20	16.83
Min–max	28.33–88.67	28.33–88.67
Number of observations	6,941	69,834

Notes: Standard deviations (SDs) are presented only for metrical variables. CPI denotes the corruption perceptions index published by Transparency International. Starting in 2012, the CPI uses a standardized scale from zero (very corrupt) to 100 (very uncorrupt), and it includes information from several sources of the respective and previous years. For additional details on the CPI, see https://www.transparency.org/en/cpi. Given a very small within-country CPI variation, we use the average CPI score for each country between the years 2012 and 2017.

jumper. This means that, in 36 percent of cases, four judges of the panel evaluated a compatriot of the remaining fifth judge of this panel. In addition, compatriot jumpers compete more frequently in their home countries and also perform better jumps in terms of jumping distance.

4. Empirical strategy

In order to explore a possible nationalistic bias in professional ski jumping, we use *style points* awarded by each judge for a given jump as the unit

of observation. In general, it is quite challenging to study the effect of a nationalistic bias on performance evaluation. Obviously, a naïve approach of correlating a dummy variable evaluating a compatriot jumper with the style points would yield biased and inconsistent estimates because a jumper's unobserved ability is likely to affect their performance, and therefore the decision-making of the judges. For example, it is possible that jumpers whose compatriot is on the panel have, on average, a higher quality, given that both the jumper and the judge come from nations where ski jumping is more popular. However, our data allow us to compare the style points of a compatriot judge with those of non-compatriot judges within the same jump. In other words, we compare the scores from different judges who observed the same performance, estimating the following model,

$$style\ points_{jip} = \alpha_1\ compatriot\ jumper_{jip} + \theta_p + \lambda_{js} + \varepsilon_{jip}, \qquad (1)$$

where the dependent variable $style\ points_{jip}$ denotes the style points that judge j assigns to jumper i for jump p. The variable $compatriot\ jumper_{jip}$ is a dummy variable, equal to one if judge j and jumper i are from the same country; θ_p represents jump fixed effects. To control for idiosyncratic tendencies across judges (such as leniency or strictness), which might differ between judges, but also within a judge over the years, we use judge-per-season fixed effects, which is denoted by λ_{js}. A positive sign of α_1 implies a bias in favor of a compatriot jumper (in-group bias), whereas a negative sign of α_1 implies a bias against a compatriot jumper (out-group bias).

Beyond the issue of a nationalistic bias, another concern is that non-compatriot judges will assign lower (or higher) scores to jumpers if they have a compatriot judge on the panel (Zitzewitz, 2006; Sandberg, 2018). Obviously, any type of compensation (or reciprocation) is not allowed and might reinforce bias in evaluations by judging panels. To test the existence of a compensating bias, according to which judges consider whether one of the other judges is a compatriot of the evaluated jumper, we cannot use jump fixed effects because the composition of the judges is fixed within each jump. As noted earlier, a naïve approach of correlating a dummy variable of having a compatriot judge on the panel with the style points would yield biased and inconsistent estimates, because jumpers' unobserved ability is likely to affect their performance. However, ability can vary over time, differing over the years due to different preparations between seasons, injuries, or a natural decrease in physical strength that can appear at some point in a career. Hence, we need to take the different sources of unobserved heterogeneity into account. For example, Harb-Wu and Krumer (2019) investigated shooting accuracy in professional biathlon by using

biathlete-per-season fixed effects.[9] As our panel data follow the same jumpers over many years, we follow the same approach as in Harb-Wu and Krumer (2019) and use jumper-per-season fixed effects as well as competition fixed effects, along with other observed characteristics of the jump, estimating the following model:

$$\begin{aligned} \textit{style points}_{jipr} &= \alpha_1\ \textit{compatriot jumper}_{jipr} + \alpha_2\ \textit{compatriot on panel}_{jipr} \\ &\quad + \lambda_{js} + \delta_{is} + \mu_r + X_{ipr} + \epsilon_{jipr}. \end{aligned} \tag{2}$$

Here, $\textit{compatriot on panel}_{jipr}$ is a dummy variable that receives the value of one if judge j has a colleague on the judging panel of competition round r who is a compatriot of jumper i. This specification includes fixed effects for judges-per-season (λ_{js}), jumpers-per-season (δ_{is}), and each competition round (μ_r). X_{ipr} is our set of controls that includes a dummy variable for whether jumpers compete in their home country. It also includes an objective performance measure (i.e., the length of the jump), which is fully observed, and its squared term, as well as the wind and gate compensation points to observe the different conditions between jumps. These wind and gate compensation points, which were absent before 2010, enable us to better control for the objective quality of the jump. For this identification approach, we need to assume that there is no correlation between the composition of nationalities on the judging panel and the quality of jumps beyond what is already captured by the observables. A positive sign of α_2 implies bias in favor of jumpers who have a compatriot judge on the panel (positive reciprocation bias), while a negative sign of α_2 implies bias against such jumpers (negative compensating bias).

5. Baseline results

In Column 1 of Table 3, we present the results from model (1), controlling for jump fixed effects. Standard errors, which are three-way clustered at the judge, jumper, and jump level, appear in parentheses. We find that judges assign 0.09 style points more to their compatriot jumpers, corresponding to 29 percent of the within-jump standard deviation (as reported in Table 2). This result is statistically significant at the 1 percent level.[10]

To test the existence of a compensating bias, according to which judges take into account whether a certain jumper has a compatriot

[9]In addition, see Genakos and Pagliero (2012) and Genakos et al. (2015) for a discussion about fixed effects estimations in multi-stage sports competitions.

[10]A concern might be the possible risk of bias from censoring as there are observations with the maximal possible score of 20. However, we only observe 104 such observations (0.14 percent). Therefore, there is no serious risk of bias from censoring.

judge on the panel, we estimate model (2) because we cannot use jump fixed effects. First, we follow the approach of Zitzewitz (2006) and Sandberg (2018), who did not use the dummy variable for whether jumpers compete in their home country (Column 2).[11] We find that the *compatriot on panel* variable is positive, but not statistically significant at conventional levels ($p = 0.16$), whereas the *compatriot jumper* coefficient increases slightly. However, because 30 percent of all jumps from a compatriot jumper in our sample were performed at a home event, we consider a potential home effect as highly relevant when analyzing performance evaluations. When additionally controlling for the home event (Column 3), the *compatriot on panel* variable loses most of its magnitude and becomes almost zero and highly insignificant ($p = 0.86$). In other words, having a counterpart on the judging panel who is from the same country as the jumper has no statistically significant effect on judges' evaluation.

Theoretically, at least two explanations for such a home effect seem plausible. First, judges might be affected by the home crowd and thus bias their decision in favor of local jumpers (e.g., Garicano et al., 2005; Page and Page, 2010; Price et al., 2012; Waguespack and Salomon, 2015). Second, jumpers might simply perform better when competing in their home country, resulting in higher style points. In order to test the latter, we explore whether jumpers make longer jumps when competing in their home country, estimating the following model:

$$\begin{aligned} \textit{length of jump}_{ipr} &= b_1\ \textit{home event}_{ipr} + \alpha_2\ \textit{compatriot on panel}_{ipr} + \delta_{is} \\ &\quad + \mu_r + X_{ipr} + \epsilon_{ipr}. \end{aligned} \tag{3}$$

Here, the dependent variable is the length of jump p in meters of jumper i in competition round r, $\textit{home event}_{ipt}$ is a dummy variable that receives the value of one if a jumper competes in his home country. This specification includes fixed effects for jumper-per-season (δ_{is}), and for each competition round (μ_r), as well as a dummy of whether a jumper has a compatriot judge on the panel. X_{ipr} is our set of controls that includes the wind and gate compensation points.

In Column 4 of Table 3, we demonstrate that jumpers who compete in their home country jump, on average, 1.86 m longer compared with their jumps in competitions abroad. Similar to the case of subjective evaluation, having a compatriot judge on the panel has no statistically significant relationship with the length of jump, which is an objective

[11] Although neither of the studies controlled for the home variable in that specification, they report in footnotes 11 (Zitzewitz, 2006) and 24 (Sandberg, 2018) that their findings on the existence of compensating bias are robust to exclusion of home participants.

Table 3. Fixed effect estimates for the judges' style point scores and the length of jump

Dependent variable	Style points			Length of jump
	(1)	(2)	(3)	(4)
Compatriot jumper	0.091***	0.109***	0.094***	
	(0.008)	(0.013)	(0.014)	
Compatriot on panel		0.018	0.002	0.072
		(0.013)	(0.014)	(0.231)
Home event			0.056**	1.860***
			(0.022)	(0.487)
Jump fixed effects	Yes	No	No	No
Judge-per-season fixed effects	Yes	Yes	Yes	No
Jumper-per-season fixed effects	No	Yes	Yes	Yes
Competition-round fixed effects	No	Yes	Yes	Yes
Number of observations	76,775	76,775	76,775	15,355

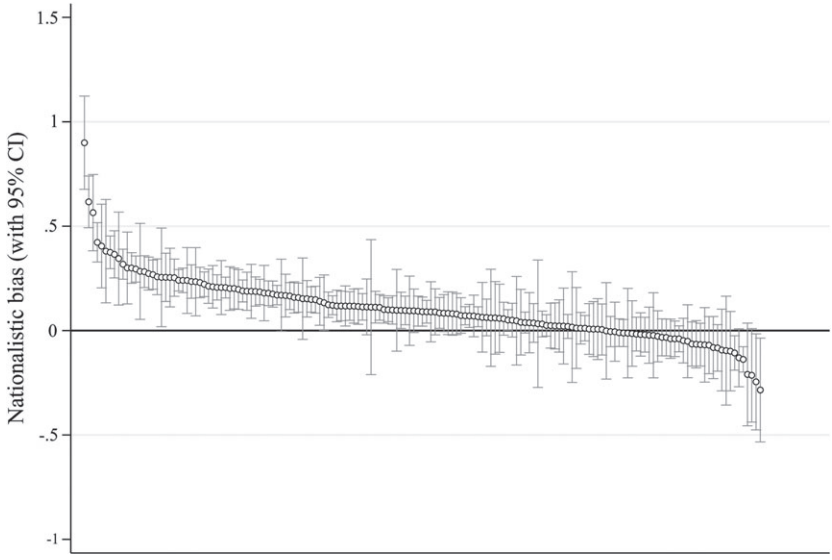
Notes: In Columns 1–3, the dependent variable is the style points of each individual judge for a given jump. If no jump fixed effects are used, we control for performance indicators, which include the jumping distance and its squared term, as well as the wind and gate points. Standard errors are three-way clustered at the judge, jumper, and jump level, and presented in parentheses. In Column 4, the dependent variable is the length of a jump in meters. Note that, in this case, the *compatriot on panel* variable defines whether one of the five judges is a compatriot. Here, we control for performance indicators, which include the wind and gate points. Standard errors are two-way clustered at the jumper and competition-round level and presented in parentheses. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

measure of performance. Thus, we conclude that jumpers perform better when competing in their home country, which might also explain their higher style point scores. A possible explanation for such a home advantage is familiarity with the facilities (e.g., Barnett and Hilditch, 1993; Koning, 2011), a crucial factor in this technical discipline, which involves complex aerodynamic elements.

To test whether our findings on nationalistic bias are driven by extreme judges (outliers), we follow the approach of Sandberg (2018) by replacing the *compatriot jumper*$_{jip}$ variable in model (1) with an interaction term between a dummy for a specific judge and *compatriot jumper*$_{jip}$, and including judge fixed effects instead of judge-per-season fixed effects. We run this estimation separately for each of the 172 judges to obtain coefficients indicating how much, on average, judge j deviates from the other judges on the panel when jumper i is a compatriot minus how much, on average, judge j deviates from the other judges on the panel when jumper i is of another nationality.

In Figure 1, we present the results of this judge-specific degree of nationalistic bias. The figure shows that 76.7 percent of judges show a positive nationalistic bias and 49.1 percent are positive and statistically significant ($p < 0.05$), while only 2.5 percent of judges show a negative and

Figure 1. Judge-specific nationalistic bias



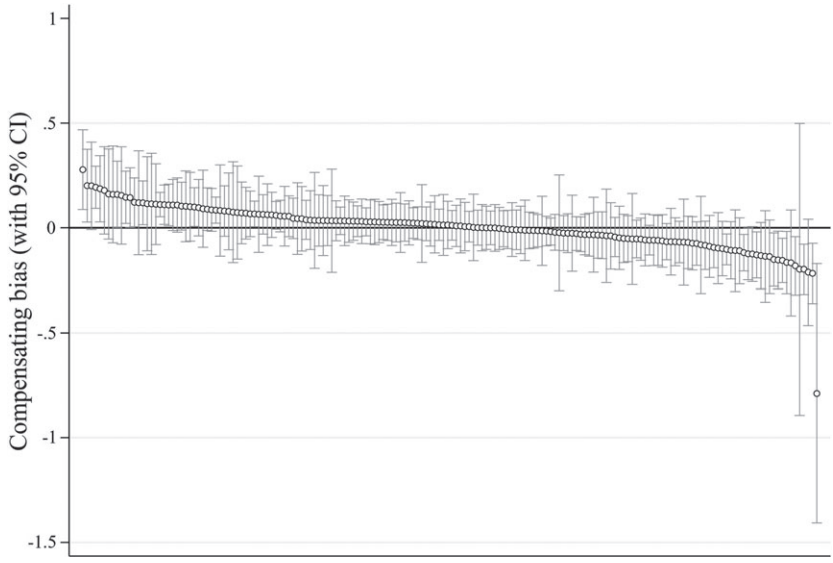
Notes: Each point represents a judge-specific estimate of the degree of nationalistic bias with 95 percent confidence intervals based on model (1) with judge fixed effects instead of judge-per-season fixed effects.

statistically significant nationalistic bias.[12] Therefore, we conclude that the finding on positive nationalistic bias is not driven by only a few extremely biased judges.

Likewise, to test whether our findings on the absence of compensating bias are driven by some abnormal patterns of individual judges, we present the judge-specific degree of compensating bias. For this analysis, we follow Sandberg (2018) to estimate a modified model (2), by replacing the $compatriot\ on\ panel_{jipr}$ variable with interaction terms between a dummy for each judge and $compatriot\ on\ panel_{jipr}$, and including judge fixed effects instead of judge-per-season fixed effects. In Figure 2, we present the results of this judge-specific degree of compensating bias. While 54.7 percent of judges show a positive compensating bias, only 8.1 percent are positive and statistically significant ($p < 0.05$). The remainder (45.3 percent) show a negative compensating bias,[13] with only 7.0 percent of judges showing a negative and statistically significant compensating bias. Taken together,

[12]We also estimated the judge-per-season specific degree of nationalistic bias. These results are available upon request and show a very similar pattern to that in Figure 1.
[13]We also estimated the judge-per-season specific degree of compensating bias. These results are available upon request and show a very similar pattern to that in Figure 2.

Figure 2. Judge-specific compensating bias



Notes: Each point represents a judge-specific estimate of the degree of compensating bias with 95 percent confidence intervals based on model (2) with judge fixed effects instead of judge-per-season fixed effects.

the results suggest that compensating bias is not likely to play a significant role in performance evaluations by ski jumping judges.

6. Effect heterogeneity

We further explore potential sources and variation of the nationalistic voting behavior of judges. First, we analyze event-specific variation of nationalistic bias. As ski jumping competitions consist of two rounds and only the top 30 jumpers qualify for the second round, their performances are decisive for determining the final ranking, including the winner and the distribution of prize money. Thus, stakes are higher in the second round and judges might have incentives to increase their nationalistic bias. Competitions also have different hill size categories (normal, large, flying) and the importance of the style point score varies across these categories because of a different calculation of the final score. For example, in our data, the shares of style points from the final score are 45 percent, 44 percent, and only 30 percent for normal, large, and flying hills, respectively. Thus, the judges' contributions to the final outcome are less important at flying hill competitions, which reduces incentives for biased behavior. The nationalistic

Table 4. Event-specific variation of nationalistic bias

Subsample estimations	No. of obs.	Coefficient	Standard error	p-value
Round 1	49,020	0.095***	0.009	0.000
Round 1 Top 30	27,755	0.087***	0.009	0.000
Round 2	27,755	0.087***	0.010	0.000
Normal hills	6,215	0.089***	0.026	0.001
Large hills	58,435	0.093***	0.008	0.000
Flying hills	12,125	0.086***	0.016	0.000
World Cups	62,205	0.092***	0.008	0.000
Four Hills	10,610	0.088***	0.014	0.000
World Championships	3,185	0.082**	0.031	0.016
Olympic Games	775	0.217*	0.104	0.093

Note: The dependent variable is the style points of each individual judge for a given jump. All estimates are based on subsample estimations of model (1) with judge-per-season and jump fixed effects. Standard errors are three-way clustered at the judge, jumper, and jump level. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

bias might also be stronger for events with a national character, such as the Olympic Games or World Championships, because national identity becomes more salient (e.g., Sandberg, 2018).[14]

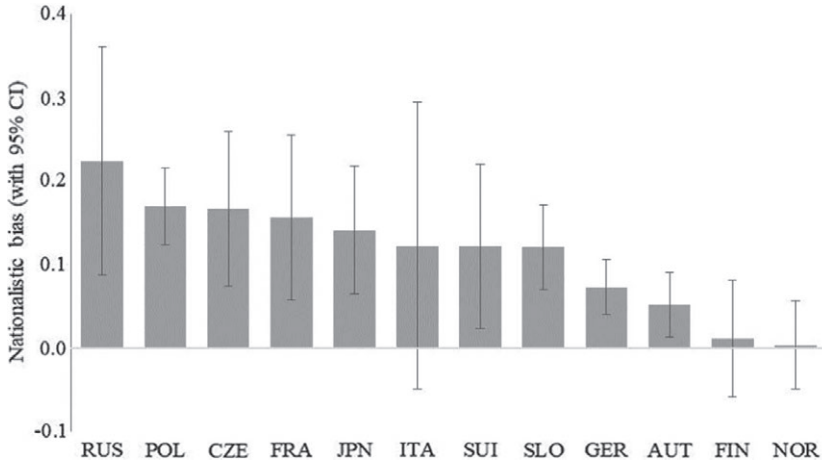
In Table 4, we present the results of model (1) for the different subsamples of the data. Overall, the degree of nationalistic bias is similar for event rounds, hill sizes, and event types.[15] An exception is the Olympic Games, where the nationalistic bias is more than twice as large. However, when estimating model (1), including an interaction between *compatriot jumper* and Olympic Games, we find no statistically significant difference (coefficient = 0.13, $p = 0.17$). Still, the nationalistic bias in the Olympic Games does not seem to be economically negligible, even if it is not statistically significant at conventional levels.

We further test whether nationalistic bias might vary by country. In Figure 3, we present results for the nationalistic bias estimates of the 12 most observed countries in our dataset, based on subsample estimations of model (1) without judge-per-season fixed effects. We see that Russia has the highest nationalistic bias (0.22). Out of the 12 countries, Norway (0.00) and Finland (0.01) are the only two countries whose coefficients are negligibly

[14]We also consider the Four Hills tournament as a separate event type because it includes the most prestigious World Cups in the calendar. The event has taken place in Germany and Austria each year since 1953. Winning all four events in one Four Hills Tournament edition is known as a grand slam. For additional information, see https://en.wikipedia.org/wiki/Four_Hills_Tournament.

[15]We also estimated model (1) with all data and interaction terms between *compatriot jumper* and event round, hill size, and event type, finding no statistically significant differences, except for a slightly larger nationalistic bias in the first event round compared with the second round (coefficient = −0.02, $p = 0.09$).

Figure 3. Country-specific variation of nationalistic bias



Notes: The figure shows the average nationalistic bias with 95 percent confidence intervals of judges when they evaluate performances of their compatriot jumpers. The estimates are based on subsample estimations of model (1) without judge-per-season fixed effects for the performances of all ski jumpers from the respective countries. The 12 countries are those with the most performance observations. The order of countries is based on the size of nationalistic bias. Please see Table A1 for the country abbreviations.

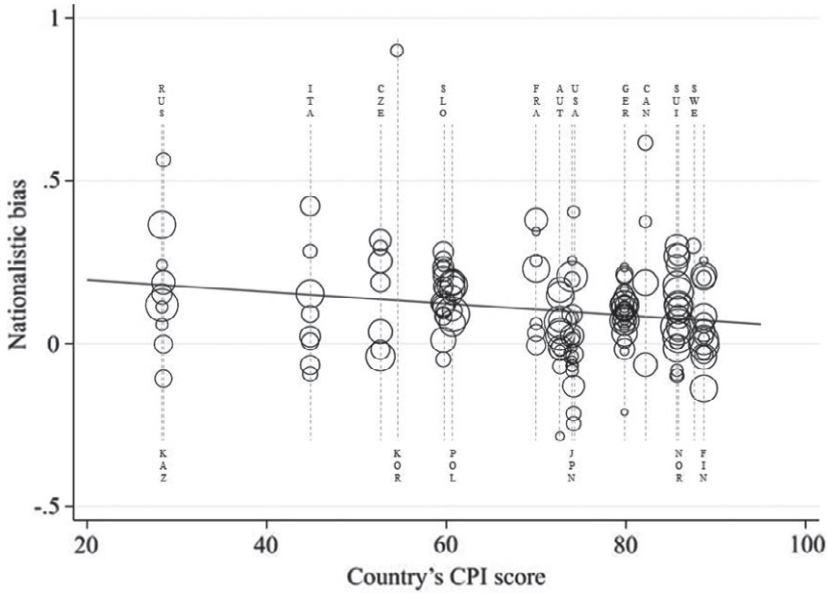
small, both economically and statistically.[16] Such country-specific variation of nationalistic favoritism also seems a plausible explanation for the large but statistically insignificant effect for the Olympic Games. By looking at the composition of the judging panel at the 2014 Sochi Olympic Games, we find large differences in the judges' nationalistic bias. Again, Russia has the largest estimated bias (coefficient = 0.64, $p = 0.01$), which is 70 percent larger than the second highest estimated bias of Switzerland (coefficient = 0.38, $p = 0.08$).[17]

In general, because the athletes' performances must be evaluated objectively and without any prejudice, such favoritism can be described as a corrupt type of behavior. Therefore, we explore whether nationalistic bias is related to the corruption perceptions index (CPI) of countries. In

[16]Please note that Italian judges participated in only 21 percent of competitions compared to 59 percent and 53 percent of Norwegian and Finnish judges, respectively. For additional details, see Table A1 in the Appendix.

[17]The estimates are based on subsample estimations of model (1) for the 2014 Sochi Olympic Games and only for jumpers from countries whose judges were part of the panel. Because of data constraints, we could not use judge or judge-per-season fixed effects. The full set of results is available upon request.

Figure 4. Nationalistic bias and the CPI



Notes: The circles show the judge-specific nationalistic bias. The size of the circles is relative to the number of observations for each judge in the data. The dashed vertical lines label the respective countries at the level of their CPI score. The regression line depicts the linear relationship between the judge-specific bias and the CPI score of the judges' countries.

Figure 4, we demonstrate a negative relationship between the judge-specific nationalistic bias and the CPI score for a country.

The coefficient of the corresponding regression is −0.002 and it is statistically significant ($p = 0.01$).[18] In other words, the higher the CPI (less corrupt country), the lower the nationalistic bias. To put this result into perspective, an increase in one standard deviation in CPI reduces the nationalistic bias by 0.03 style points, which is 10 percent of the within-jump standard deviation of the evaluation of style points.

Finally, because Russia had the highest estimated nationalistic bias in the 2014 Olympic Games, but was also the only country that hosted Olympic Games in our data, it is possible that our findings on the relationship between the CPI and nationalistic bias are driven by hosting the Olympic

[18]The regression is based on model (1) and includes an interaction term between *compatriot jumper* and the CPI score to estimate the relationship. We also run an alternative specification where we weigh by the number of observations per country. The results are very similar (coefficient = −0.002, $p = 0.02$).

Games and not by Russia per se. To obviate this concern, we remove the data of the Olympic Games and perform similar analyses to those in Figures 3 and 4. The results presented in Figures A1 and A2 in the Online Appendix show a very similar pattern. This finding is in line with previous cross-country evidence on positive relationships between unethical behavior and corruption levels in experimental settings (Barr and Serra, 2010; Gächter and Schulz, 2016) and non-experimental settings (Zitzewitz, 2006; Fisman and Miguel, 2007; Elaad et al., 2018).

7. A comparative view on nationalistic and compensating biases

Next, we compare the magnitude of nationalistic and compensating biases in our paper with the biases reported in Zitzewitz (2006) and Sandberg (2018).[19] Given the different scale of scores between the different sports, we present estimates standardized by the within-performance standard deviation. Figure 5 compares the standardized nationalistic bias estimates across studies. The nationalistic bias is the smallest in dressage (0.24), followed by our estimate for ski jumping (0.29). In comparison, the standardized coefficients for ski jumping (0.44) and figure skating (0.47), as reported in Zitzewitz (2006), are considerably larger.[20]

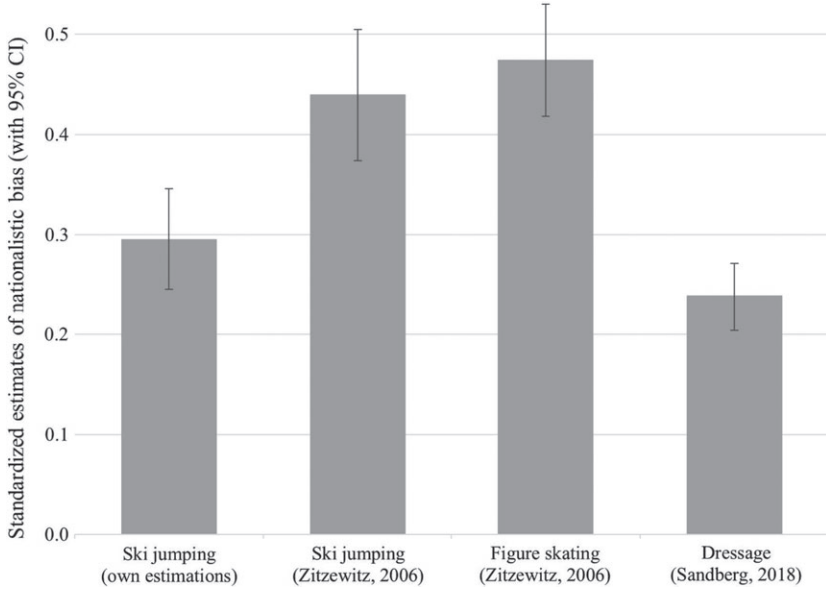
When comparing both findings for ski jumping, it should be noted that Zitzewitz (2006) only used data from 2002, an Olympic year. In both studies related to ski jumping, the nationalistic bias in the Olympic Games is the highest: 0.26 in Zitzewitz (2006)[21] and 0.22 in our study. As such, the inclusion of the Olympic Games generally increases the average estimate of nationalistic bias in ski jumping. However, in our case, the share of the Olympic Games is only 1 percent of the overall number of observations, while it is 13 percent in Zitzewitz (2006). When comparing our finding for ski jumping with the finding for figure skating, it should be noted that ice dancing accounts for one-third of the data on figure skating in Zitzewitz (2006). As noted by the author, "biases are larger where scoring is more subjective, as it is for ice dancing, where skaters do not have as many

[19]Because Zitzewitz (2014), using a figure-skating setting, was unable to differentiate between nationalistic and compensating biases and Scholten et al. (2020), using a ski-jumping setting, employed a different estimation approach, neglecting some key issues, as mentioned in the Introduction, their results are hardly comparable with ours and, as such, are not considered here.

[20]We also find a similar pattern when we standardize the point estimates by the overall-performance standard deviation, which yields a standardized coefficient of 0.09 for our ski jumping estimations, 0.13 for both ski jumping and figure skating (Zitzewitz, 2006), and 0.07 for dressage (Sandberg, 2018).

[21]See Panel A, Line 2 in Table 5 of Zitzewitz (2006).

Figure 5. Standardized nationalistic bias estimates across studies



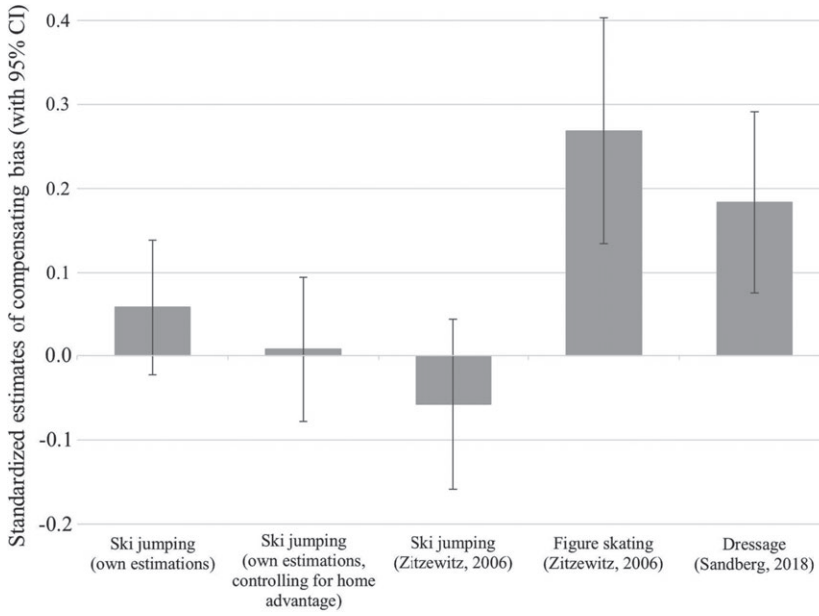
Notes: The figure shows the estimates of nationalistic bias with 95 percent confidence intervals, both standardized by the within-performance standard deviation. To calculate the values presented in this figure, we use the non-standardized point estimates from Column 1 in Table 3 for ski jumping (our own estimations); we also use values for ski jumping and figure skating from Zitzewitz (2006, Table 3, Lines 5 and 3, respectively), and values for dressage from Sandberg (2018, Table 3, Column 1).

mandatory deductions for falls, and for artistic impression as opposed to technical merit scores" (Zitzewitz, 2006, p. 79). This is in line with a recent paper by Joustra et al. (2021), who found a significant advantage for later performances in female gymnastics, which is likely to be driven by the existence of subjective evaluation only in female competitions considering artistry. In fact, the nationalistic bias found by Zitzewitz (2006) for ice dancing is 33 percent higher than that for non-ice dancing disciplines.[22]

Figure 6 compares the standardized compensating bias estimates across studies. While both of our standardized coefficients (i.e., with and without controlling for home advantage) are positive and the standardized coefficient for ski jumping in Zitzewitz (2006) is negative, none of these differs significantly from zero. In contrast to these findings, the standardized coefficients for figure skating (0.27) and dressage (0.18) are comparably

[22] See Panel B in Table 5 of Zitzewitz (2006).

Figure 6. Standardized compensating bias estimates across studies



Notes: The figure shows the estimates of compensating bias with 95 percent confidence intervals, both standardized by the within-performance standard deviation. To calculate the values presented in this figure, we use the non-standardized point estimates from Column 2 in Table 3 for ski jumping (our own estimations) and from Column 3 in Table 3 for ski jumping (our own estimations), where we also control for home advantage; we also use values for ski jumping and figure skating from Zitzewitz (2006, Table 4, Panel A/Line 5b and Panel B/Line 5, respectively), and for dressage from Sandberg (2018, Table 5, Column 2).

large and significant.[23] One possible reason for this might be that the estimating equations for figure skating and dressage do not include the home variable. However, the home variable is also excluded from the estimating equation for ski jumping in Zitzewitz (2006), and both Sandberg (2018) and Zitzewitz (2006) have reported that their results are robust to exclusion of home participants.

Another possible reason why we observe positive reciprocation biases in figure skating and dressage in contrast to ski jumping might be the differences in institutional settings that relate to truncation and exposure of

[23]We also find a similar pattern when we standardize the point estimates of compensating bias by the overall-performance standard deviation, which yields a standardized coefficient of 0.02 for our ski jumping estimations and 0.00 when controlling for home advantage, −0.03 for ski jumping and 0.07 for figure skating in Zitzewitz (2006), and 0.06 for dressage in Sandberg (2018).

scores in the sports. In the scoring system used in figure skating, judges' scores are transformed into votes about the skaters' relative performance. According to Zitzewitz (2006), such a system makes it easier to detect a defection from reciprocal arrangements than a system with continuous scores. Therefore, the transformation of scores into votes can make reciprocal arrangements easier to sustain. In dressage, a rule promotes consistency in scoring. According to this rule, the panel members must have an evaluation meeting after the competition if the scores for a performance differ by more than 5 percent among the judges. Thus, it seems possible that experienced dressage judges anticipate the nationalistic bias of their panel members and act accordingly – that is, they give better (and biased) scores to ensure consistency. This is different in ski jumping, where the truncation mechanism seems to lower incentives either for compensation (because extreme votes are excluded) or for consistency (because there is no such 5 percent rule).

8. Conclusion

In this paper, we have examined nationalistic bias in subjective evaluations by international experts, which has been shown to be a significant factor in previous studies (Zitzewitz, 2006; Sandberg, 2018). Our efforts in this regard are in line with the increasing importance of replication studies (Ioannidis and Doucouliagos, 2013; Open Science Collaboration, 2015) and crowdsourced research (Silberzahn and Uhlmann, 2015; Silberzahn et al., 2018).

Our findings confirm the existence of nationalistic voting of judges in professional ski jumping competitions more than a decade after this bias was first illustrated in similar settings. This nationalistic bias is found for a large share of judges and is positively associated with the level of corruption according to the CPI. Our results suggest that in-group favoritism is a strong feature of human behavior, especially in countries with a high prevalence of corruption in their institutional environment. In addition, unlike previous findings, our results show no evidence of strategic voting, according to which judges assign different scores to jumpers whose compatriots are present on the judging panel. This discrepancy can be partly explained by different approaches in dealing with home advantage and different institutional settings. Unlike figure skating and dressage, ski jumping uses a truncation mechanism, according to which the highest and lowest scores are excluded, which seems to lower incentives for strategic voting.

It is important to note that our results were obtained from fully observable sports competitions. Such in-group favoritism might even be stronger in less transparent settings that involve subjective decision-making,

such as policymaking processes, judging in legal proceedings, human resource management, etc.

Finally, we call for future research to investigate nationalistic favoritism in other settings to create higher awareness of this primitive human instinct that has not yet disappeared. This call is particularly important during times when the entire humanity faces difficulties, such as COVID-19, where the immediate and natural desire is to protect in-group members, which could lead to an increased nationalistic favoritism.

Appendix

Table A1. Frequencies of countries by groups of jumpers, judges, and competitions

Country name	Country code	Jumpers	Jumps	Judges	Competitions	Judges in competitions
Austria	AUT	27	2,077	12	19	91 (45%)
Bulgaria	BUL	1	110	0	0	0 (0%)
Canada	CAN	5	113	4	0	22 (11%)
Czech Republic	CZE	15	1171	7	6	50 (25%)
Estonia	EST	4	70	0	0	0 (0%)
Finland	FIN	19	717	16	23	107 (53%)
France	FRA	6	316	7	0	33 (16%)
Germany	GER	24	2,098	30	38	145 (71%)
Greece	GRE	1	3	0	0	0 (0%)
Italy	ITA	8	299	8	4	43 (21%)
Japan	JPN	27	1408	15	14	54 (27%)
Kazakhstan	KAZ	8	53	4	2	20 (10%)
South Korea	KOR	4	39	1	2	2 (1%)
Netherlands	NED	1	1	0	0	0 (0%)
Norway	NOR	27	2,037	12	38	120 (59%)
Poland	POL	20	1,666	8	16	79 (39%)
Romania	ROU	2	3	4	0	8 (4%)
Russia	RUS	18	688	6	8	38 (19%)
Slovenia	SLO	27	1,774	15	15	84 (41%)
Switzerland	SUI	10	564	10	15	60 (30%)
Slovakia	SVK	1	2	2	0	14 (7%)
Sweden	SWE	2	10	4	3	22 (11%)
Ukraine	UKR	2	2	0	0	0 (0%)
USA	USA	9	134	7	0	23 (11%)
Total	24	268	15,355	172	203	

Notes: The last column states the number of competitions in which the respective country has a judge on the panel. There are five judges in each competition. This is also presented as percentage share based on the total number of competitions in parentheses.

Supporting information

Additional supporting information may be found online in the supporting information section at the end of the article.

Online Appendix
Replication Files

References

Andrews, T. J., Smith, R. K., Hoggart, R. L., Ulrich, P. I., and Gouws, A. D. (2019), Neural correlates of group bias during natural viewing, *Cerebral Cortex 29*, 3380–3389.

Barnett, V. and Hilditch, S. (1993), The effect of an artificial pitch surface on home team performance in football (soccer), *Journal of the Royal Statistical Society: Series A (Statistics in Society) 156*, 39–50.

Barr, A. and Serra, D. (2010), Corruption and culture: an experimental analysis, *Journal of Public Economics 94*, 862–869.

Efferson, C., Lalive, R., and Fehr, E. (2008), The coevolution of cultural groups and ingroup favoritism, *Science 321*, 1844–1849.

Elaad, G., Krumer, A., and Kantor, J. (2018), Corruption and sensitive soccer games: cross-country evidence, *Journal of Law, Economics, and Organization 34*, 364–394.

Faltings, R., Krumer, A., and Lechner, M. (2021), Rot-jaune-verde. language and favoritism: evidence from swiss soccer, School of Economics and Political Science, Department of Economics, University of St Gallen, Working Paper.

Fédération Internationale de Ski (FIS) (2017a), *Rules for the FIS Ski Jumping World Cup (men)*, Edition 2017/2018, Oberhofen, Switzerland.

Fédération Internationale de Ski (FIS) (2017b), *The International Ski Competition Rules (ICR), Book III, Ski Jumping*, Oberhofen, Switzerland.

Fisman, R. and Miguel, E. (2007), Corruption, norms, and legal enforcement: evidence from diplomatic parking tickets, *Journal of Political Economy 115*, 1020–1048.

Garicano, L., Palacios-Huerta, I., and Prendergast, C. (2005), Favoritism under social pressure, *Review of Economics and Statistics 87*, 208–216.

Gächter, S. and Schulz, J. F. (2016), Intrinsic honesty and the prevalence of rule violations across societies,*Nature 531*, 496–499.

Genakos, C. and Pagliero, M. (2012), Interim rank, risk taking, and performance in dynamic tournaments, *Journal of Political Economy 120*, 782–813.

Genakos, C., Pagliero, M., and Garbi, E. (2015), When pressure sinks performance: evidence from diving competitions, *Economics Letters 132*, 5–8.

Harb-Wu, K. and Krumer, A. (2019), Choking under pressure in front of a supportive audience: evidence from professional biathlon, *Journal of Economic Behavior & Organization 166*, 246–262.

Ioannidis, J. and Doucouliagos, C. (2013), What's to know about the credibility of empirical economics?, *Journal of Economic Surveys 27*, 997–1004.

Joustra, S. J., Koning, R. H., and Krumer, A. (2021), Order effects in elite gymnastics, *De Economist 169*, 21–35.

Koning, R. H. (2011), Home advantage in professional tennis, *Journal of Sports Sciences 29*, 19–27.

Open Science Collaboration (2015), Estimating the reproducibility of psychological science, *Science 349*, aac4716.

Page, K. and Page, L. (2010), Alone against the crowd: individual differences in referees' ability to cope under pressure, *Journal of Economic Psychology 31*, 192–199.
Pope, B. R. and Pope, N. G. (2015), Own-nationality bias: evidence from UEFA Champions League football referees, *Economic Inquiry 53*, 1292–1304.
Pope, D. G., Price, J., and Wolfers, J. (2018), Awareness reduces racial bias, *Management Science 64*, 4988–4995.
Price, J., Remer, M., and Stone, D. F. (2012), Subperfect game: profitable biases of NBA referees, *Journal of Economics & Management Strategy 21*, 271–300.
Price, J. and Wolfers, J. (2010), Racial discrimination among NBA referees, *Quarterly Journal of Economics 125*, 1859–1887.
Sandberg, A. (2018), Competing identities: a field study of in-group bias among professional evaluators, *Economic Journal 128*, 2131–2159.
Scholten, H., Schneemann, S., and Deutscher, C. (2020), The impact of age on nationality bias and cultural proximity bias: evidence from ski jumping, *Journal of Institutional and Theoretical Economics 176*, 708–734.
Shayo, M. and Zussman, A. (2011), Judicial ingroup bias in the shadow of terrorism, *Quarterly Journal of Economics 126*, 1447–1484.
Silberzahn, R. and Uhlmann, E. L. (2015), Crowdsourced research: many hands make tight work, *Nature News 526*, 189–191.
Silberzahn, R. et al. (2018), Many analysts, one data set: making transparent how variations in analytic choices affect results, *Advances in Methods and Practices in Psychological Science 1*, 337–356.
Spierdijk, L. and Vellekoop, M. (2009), The structure of bias in peer voting systems: lessons from the Eurovision Song Contest, *Empirical Economics 36*, 403–425.
Sumner, W. G. (1907), *Folkways: A Study of the Sociological Importance of Usages, Manners, Customs, Mores, and Morals*, Ginn and Company, Boston, MA.
Waguespack, D. M. and Salomon, R. (2015), Quality, subjectivity, and sustained superior performance at the Olympic Games, *Management Science 62*, 286–300.
Yuki, M. (2003), Intergroup comparison versus intragroup relationships: a cross-cultural examination of social identity theory in North American and East Asian cultural contexts, *Social Psychology Quarterly 66*, 166–183.
Zitzewitz, E. (2006), Nationalism in winter sports judging and its lessons for organizational decision making, *Journal of Economics & Management Strategy 15*, 67–99.
Zitzewitz, E. (2014), Does transparency reduce favoritism and corruption? Evidence from the reform of figure skating judging, *Journal of Sports Economics 15*, 3–30.

First version submitted April 2020;
final version received May 2021.