Nationalistic bias among international experts: Evidence from professional ski jumping

Alex Krumer [a,], Felix Otto [b], Tim Pawlowski [b, c]*

[a] Faculty of Business Administration and Social Sciences, Molde University College, Britvegen 2, Molde, 6402, Norway

[b] University of Tübingen, Wilhelmstraße 124, DE 72074 Tübingen, Germany

[c] Tim Pawlowski is also affiliated with the LEAD Graduate School and Research Network as well as the Interfaculty Research Institute for Sports and Physical Activity in Tuebingen.

Email: alex.krumer@himolde.no*, felix.otto@uni-tuebingen.de, tim.pawlowski@uni-tuebingen.de

*Corresponding author

This version: October 2020

Date this version was printed:* *17 October 2020

Abstract: Ski jumping competitions involve subjective evaluations by judges from different countries. This may lead to nationalistic bias, according to which judges assign higher scores to their compatriots. In order to test this claim empirically, we exploit within-performance variation of scores from all World Cup, World Championship, and Olympic Games competitions between seasons 2010/11 and 2016/17. Our findings confirm that judges assign significantly higher scores to their compatriots. Further analysis reveals that the magnitude of the nationalistic bias is significantly higher in more corrupt countries, according to the Corruption Perceptions Index. Moreover, out of the 12 most observed countries in our data, only Norway and Finland had negligible estimates of the nationalistic bias both statistically and economically. However, in contrast to previous studies, we do not find that judges assign significantly different scores to jumpers whose compatriots are present on the judging panel. We discuss our findings in the context of the growing debates about the importance of replication studies and crowdsourced research.

Keywords: Subjective performance evaluation; nationalistic bias; in-group favoritism; judging panel; replication studies

JEL Classification: D71, D91, Z20

Declarations of interest: none

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

1 Introduction

Can well-trained and professional experts resist the inherent preferences toward in-group members in their subjective evaluation? 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 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 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, 1904; 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. There were also several studies that 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 the UEFA Champions League games. Very recently, Faltings et al. (2019) investigated Swiss

[1] As evidence, Andrews et al. (2019) tested brain activities of fans from two rival soccer teams who watched the same soccer game. The authors 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.

soccer games 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 the jumping distance and is an objectively measured performance, and the style points, which is 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. Using figure skating competitions, Zitzewitz (2006) found a similar pattern of a nationalistic bias. More recently, Sandberg (2018) showed an analogous result in dressage competitions.

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.[2] Second, it seems highly relevant from a policy perspective to see whether problems that had been identified before, have been solved over time.[3]

Comparing within each jump the score of a compatriot judge to scores of the other members of the panel, we find that compatriot judges assign close to 0.1 points more compared

[2] 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 spirit, 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) discuss 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.

[3] 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.

to their counterparts. This is equal to about 29% of the within-jump standard deviation of scores, which is a non-negligible result. 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.[4]

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, 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. Thus, we conclude that the prevalence of corruption in the institutional environment within the judges' countries seems to shape their behavior in international evaluation settings.[5]

Finally, we test whether there is evidence of strategic voting, according to which judges assign significantly different scores to jumpers whose compatriots are present on the judging panel. The evidence on such a strategic voting is quite 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. This might be explained by differences between our and the previous studies in dealing with home advantage in the

[4] For example, Zitzewitz's (2006) findings were summarized and discussed in the *Washington Post*. See, https://www.washingtonpost.com/news/monkey-cage/wp/2014/02/12/how-ski-jumping-gets-olympic-judging-right-and-figure-skating-gets-it-wrong/ (last accessed on 16.10.2020).

[5] This result adds to the previous finding of Elaad et al. (2018) who show the more corrupt the country the higher the probability of a team to 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.

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.[6]

The remainder of the paper is organized as follows: Section 2 describes the institutional settings of ski jumping competitions. The data and descriptive statistics are presented in Section 3. Section 4 presents the empirical strategy. Section 5 presents the baseline results, while Section 6 explores effect heterogeneity. In Section 7, we compare our results with the results in other studies. In Section 8, we offer concluding remarks.

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 to maximize the length of the jump and the style points awarded by the judging panel. There are three different hill sizes (HS) that are used in professional ski jumping events: normal hills (HS 85 m to 109 m), large hills (HS 110 m to 184 m), and flying hills (HS 185 m and larger). Usually, 50 competitors jump in the first round. In flying hills, this 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.[7]

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 meters. This distance is converted to a point value that contributes to the aggregate score. In addition,

[6] 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). Note that in a follow-up study, Zitzewitz (2014) found higher scores for figure skaters whose compatriots were present on the panel. However, the data did not allow him to disentangle nationalistic and compensating biases. In addition, a recent paper by Scholten et al. (2020) also found the existence of nationalistic bias in ski jumping competitions by using only two seasons that included 41 World Cup competitions. However, that paper did not exploit the within-performance variation of scores as was done in Zitzewitz (2006), Sandberg (2018), and our paper. In addition, Scholten et al. (2020) did not investigate the possibility of a compensating bias.

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

there is a subjective performance evaluation by a judging panel. This consists of five judges from five different countries, one of which is always from 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 remaining three 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 professional experts in this task as the international governing body for winter sports, the Fédération Internationale de Ski (FIS), selects only highly skilled individuals for this job. The 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 is allowed and the decision must be entered into the scoring system without any assistance. Moreover, the judging tower where judges are located is constructed in a way to provide optimal conditions to execute the judging task and to guarantee 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.

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. The selected period was chosen because of the introduction of the wind and gate compensation points starting from the 2010/11 season.

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

Table 1: Sample size

No. of ski jumpers	268
No. of ski jumper countries	24
No. of judges	172
No. of judge countries	19
No. of total competitions	203
No. of World Cups	165
No. of Four Hills	28
No. of Nordic World Championships	8
No. of Olympic Games	2
No. of jumps (performances)	15,355
Average no. of jumps per athlete	57.29 (81.70)
Average no. of jumps per athlete and season	17.73 (17.28)
No. of style point scores	76,775
Average no. of scores per judge	446.37 (291.87)
Average no. of scores per judge and season	143.24 (66.41)

Note: Standard deviations are presented in parentheses.

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 Appendix A, the competitions took place in 14 countries. Most of them were held in Norway and Germany, with 38 competitions in each country. German judges were part of the panels in 71% of competitions, followed by Norwegian judges (59%) as the second most frequent country.

Table 2 provides the summary statistics for the full sample as well as subsamples considering whether a judge and jumper are from the same country or not. We see that on average, judges assign a higher score to their compatriots. In 9% of cases, i.e., overall 6,941 evaluations, a judge was a compatriot of the evaluated jumper. This means that in 36% 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.

Table 2: Descriptive statistics

	Ski jumpers		
	All	Compatriot jumpers	Non-compatriot jumpers
Style points			
Mean	17.49	17.61	17.48
(overall SD)	1.07	1.05	1.07
(within-jump SD)	0.31	-	0.31
Min-max	4.0-20.0	5.0-20.0	4.0-20.0
Compatriot on panel			
Mean	0.36	0	0.40
Home event			
Mean	0.14	0.30	0.12
Jumping distance			
Mean	131.62	133.59	131.42
(overall SD)	30.36	31.43	30.24
Min-max	51.0-251.5	55.0-251.5	51.0-251.5
Wind points			
Mean	-0.87	-0.92	-0.87
(overall SD)	8.43	8.30	8.44
Min-max	-34.9-45.7	-34.9-43.4	-34.9-45.7
Gate points			
Mean	0.18	0.18	0.18
(overall SD)	4.54	4.56	4.53
Min-max	-29.4-52.7	-29.4-45.2	-29.4-52.7
Country CPI score (2012-2017)			
Mean	71.01	72.90	70.82
(overall SD)	16.54	13.20	16.83
Min-max	28.33-88.67	28.33-88.67	28.33-88.67
No. of observations	76,775	6,941	69,834

Note: Standard deviations 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 includes information from several sources of the respective and previous years. For additional details on the CPI, see https://www.transparency.org/en/cpi (last accessed on 16.10.2020). Given a very small within-country CPI variation, we use the average CPI score for each country between the years 2012 and 2017.

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, studying the effect of a nationalistic bias on performance evaluation is quite challenging. 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 as both the jumper and the judge come from nations where ski jumping is more popular. However, our data allow us to compare within the same jump the style points of a compatriot judge with the style points of non-compatriot judges. In other words, we compare the scores from different judges who observed the same performance, estimating the following model:

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

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 that receives the value of 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 (e.g., leniency or strictness), which may 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 (Sandberg, 2018; Zitzewitz, 2006). Obviously, any type of compensation (or reciprocation) is not legal and may reinforce bias in evaluations by judging panels. To test the existence of a

compensating bias, according to which judges take into account 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 previously, 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, since jumpers' unobserved ability is likely to affect their performance. However, ability may vary over time, differing over the years due to different preparations between seasons, injuries, or natural decrease in shape that may 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.[8] Since our panel data follows the same jumpers over many years, we are able to follow the approach presented 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:

$$(2)\ style\ points_{jir} = \alpha_1\, compatriot\ jumper_{jir} + \alpha_2\, compatriot\ on\ panel_{jir} + \lambda_{js} + \delta_{is} + \mu_r + X_{ir} + \varepsilon_{jir,}$$

where, $compatriot\ in\ panel_{jir}$, 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_{ir} is our set of controls that includes a dummy variable of whether a jumper competes 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, that were absent before 2010, provide us with the

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

opportunity 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 jump 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), 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 the parentheses. We find that judges assign 0.09 style points more to their compatriot jumpers corresponding to 29% of the within-jump standard deviation (as reported in Table 2). This result is significant at the 1% level.[9]

To test the existence of a compensating bias, according to which judges take into account whether a certain jumper has a compatriot judge on the panel, we cannot use jump fixed effects. Instead, we estimate model (2), but first we follow the approach of Zitzewitz (2006) and Sandberg (2018), who did not use the dummy variable of whether a jumper competes in their home country (Column 2).[10] We find that the $compatriot\ in\ panel$ variable is positive, but not significant at conventional levels (p=0.16), whereas the $compatriot\ jumper$ coefficient slightly increases. However, since 30% 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

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

[10] 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.

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 significant effect on judges' evaluation.

Table 3: FE estimates for the judges' style point scores

	(1)	(2)	(3)
Compatriot jumper	0.091***	0.109***	0.094***
	(0.008)	(0.013)	(0.014)
Compatriot on panel		0.018	0.002
		(0.013)	(0.014)
Home event			0.056**
			(0.022)
Jump FE	Yes	No	No
Judge-per-season FE	Yes	Yes	Yes
Jumper-per-season FE	No	Yes	Yes
Competition-round FE	No	Yes	Yes
No. of observations	76,775	76,775	76,775

Note: 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. ***, **, * denote significance at the 1%, 5%, 10% level, respectively.

To test whether our findings on nationalistic bias are driven by extreme judges (outliers), we analyze the data on the level of a single judge. In Figure 1, we present the results of model (1) for each individual judge by using judge fixed effects. The figure shows that 76.7% of judges show a positive nationalistic bias and 49.1% are positive and significant ($p<0.05$), while only 2.5% of judges show a negative and significant nationalistic bias.[11] We therefore conclude that the finding on positive nationalistic bias is not driven by only a few extremely biased judges.

[11] In Appendix B, we present the results of model (1) for each individual judge by using judge-per-season fixed effects. These results show a very similar pattern to that in Figure 1.

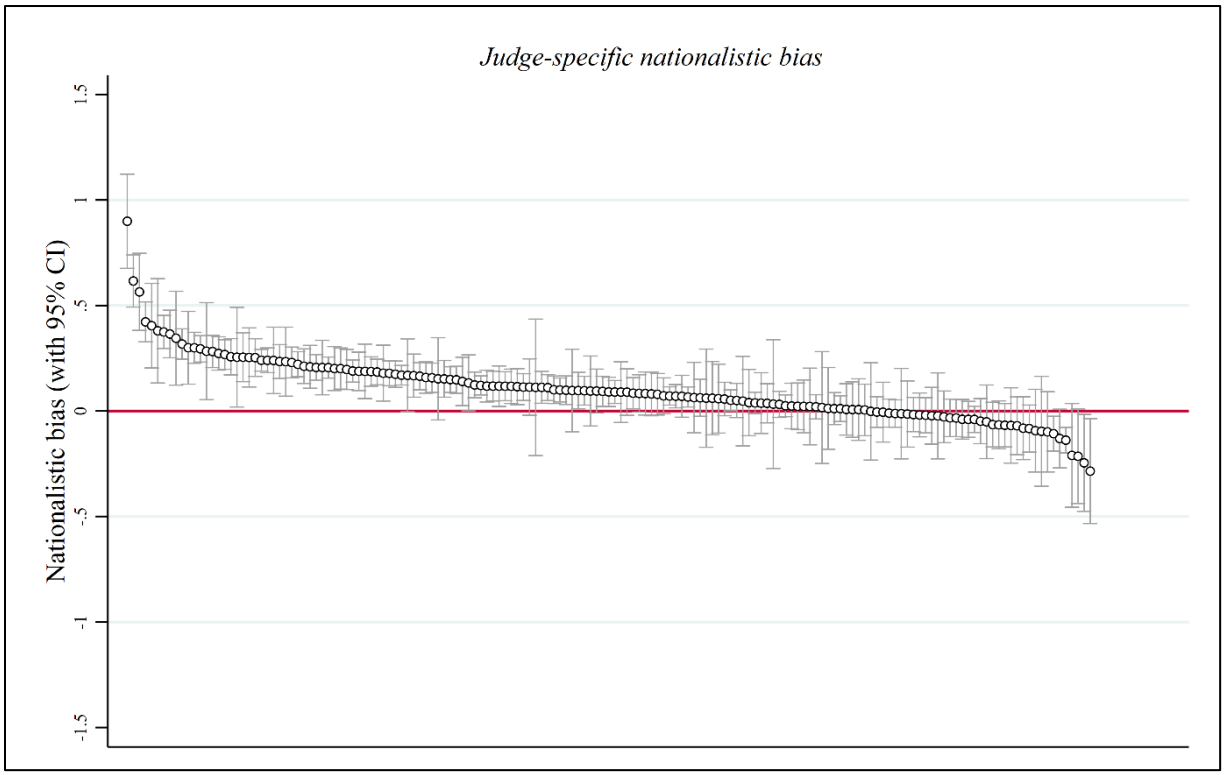


Figure 1: The figure shows the judge-specific nationalistic bias with 95% confidence intervals. The estimates are the judge-specific coefficients from model (1).

Likewise, to test whether our findings on the absence of compensating bias are driven by some abnormal patterns of individual judges, we present the results of model (2) for each individual judge by using judge fixed effects in Figure 2. While 54.7% of judges show a positive compensating bias, only 8.1% are positive and significant ($p<0.05$). The rest (45.3%) show a negative compensating bias with only 7.0% of judges showing a negative and significant compensating bias.[12] Taking together the results in Table 3, Figure 2 and Appendix C, we conclude that compensating bias is not likely to play a significant role in performance evaluations by ski jumping judges.

[12] We observe a similar pattern when using judge-per-season fixed effects, as presented in Appendix C.

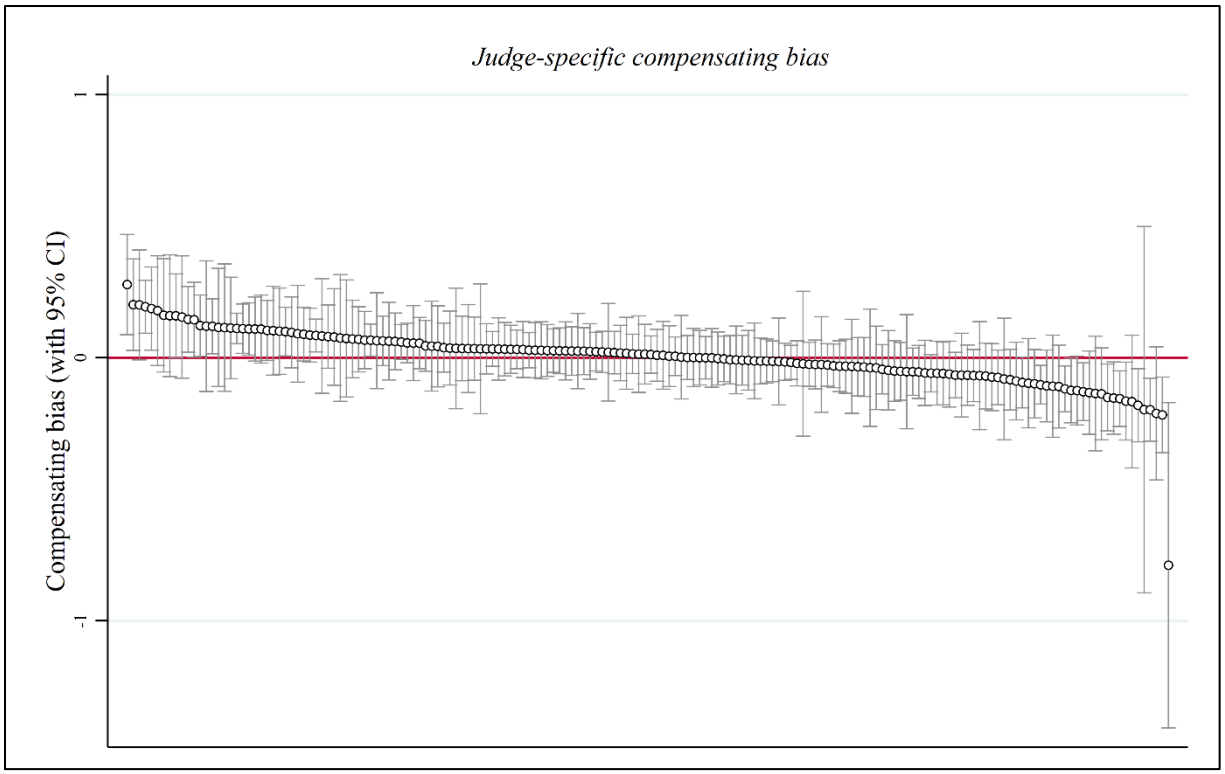


Figure 2: The figure shows the judge-specific compensating bias with 95% confidence intervals. The estimates are the judge-specific coefficients from model (2).

We further explore possible mechanisms underlying the positive and significant home effect. Theoretically, at least two explanations seem plausible. First, judges might be affected by the home crowd and thus bias their decision in favor of the 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 test whether jumpers make longer jumps when competing in their home country, estimating the following model:

$$(3)\ length\ of\ \ jump_{ir} = b_1\ home\ event_{ir} + \alpha_2\ compatriot\ judge\ on\ panel_{ir} + \delta_{is}\ + \mu_r + X_{ir} + \varepsilon_{ir,}$$

where the dependent variable is the length of jump in meters of jumper i in competition round r, $home\ event_{it}$ 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_{ir} is our set of controls that includes the wind and gate compensation points.

Table 4: FE estimates of the effect of competing at home on the length of jump

	(1)	(2)
		Top 30
Home event	1.860***	1.473***
	(0.487)	(0.461)
Compatriot judge on panel	0.072	-0.107
	(0.231)	(0.223)
Jumper-per-season FE	Yes	Yes
Competition-round FE	Yes	Yes
No. of observations	15,355	11,641

Note: The dependent variable is the length of a given jump. In all regressions, 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. ***, **, * denote significance at the 1%, 5%, 10% level, respectively.

In Column 1 of Table 4, we demonstrate that jumpers who compete in their home country have, on average, 1.86 meters longer jumps compared to when competing abroad. We see that similarly to the case with subjective evaluation, the $compatriot\ judge\ on\ panel$ also has no significant relationship with the length of jump, which is an objective type of performance. In Column 2, we restrict the sample to the top 30 jumpers who qualify for the final round. Their performances are decisive for determining the contest winner and the distribution of prize money. We find a similar result compared to Column 1, though with a lower magnitude. Thus, we conclude that jumpers perform better when competing in their home country, which may 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), which is crucial in this technical discipline, which involves complex aerodynamic elements.

6 Effect heterogeneity

In Table 5, we present the results of the effect of a nationalistic bias for different subsamples of the data using model (1). Overall, a nationalistic bias seems to be present in every round, every type of competition and every hill size. The size of this effect is in the range between 0.082 and 0.095 for all the subsamples except for the Olympic Games, whose coefficient is 0.217.

Table 5: 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. ***, **, * denote significance at the 1%, 5%, 10% level, respectively.

To have a more formal test of difference between different subsamples, we estimate model (1) by using all the data and adding interaction terms between $compatriot\ jumper$ and the relevant round/competition type/hill size. We present the results in Table 6. As described previously, competitions consist of two rounds and the final ranking and event winner is determined in the second round. Thus, stakes are higher, and judges may have incentives to increase their nationalistic bias in the second round.[13] The results in Column 1 indicate that if

[13] Note, that we removed data on competitions that had only one round due to bad weather conditions.

at all, the nationalistic bias becomes slightly lower in the second round. However, when using data on the top 30 jumpers who participated in both rounds, no significant difference in the nationalistic bias between the two rounds remain (see Column 2). Thus, we conclude that the change in stakes within a competition does not have a significant effect on nationalistic bias.

Table 6: Event-specific variation of nationalistic bias

	(1)	(2)	(3)	(4)
	Two-round competitions	Top 30		
Compatriot jumper (CJ)	0.116***	0.105***	0.084***	0.092***
	(0.016)	(0.016)	(0.026)	(0.008)
CJ x competition round	-0.016*	-0.012		
	(0.010)	(0.010)		
CJ x large hill			0.010	
			(0.025)	
CJ x flying hill			0.003	
			(0.029)	
CJ x Four Hills				-0.005
				(0.013)
CJ x World Championships				-0.008
				(0.030)
CJ x Olympic Games				0.128
				(0.094)
Jump FE	Yes	Yes	Yes	Yes
Judge-per-season FE	Yes	Yes	Yes	Yes
No. of observations	72,410	55,510	76,775	76,775

Note: The dependent variable is the style points of each individual judge for a given jump. In Column 3, the reference group is normal hill. In Column 4, the reference group is World Cups. Standard errors are three-way clustered at the judge, jumper, and jump level and presented in parentheses. ***, **, * denote significance at the 1%, 5%, 10% level, respectively.

Next, we compare the nationalistic bias between different hill sizes. The competition hills typically have different hill sizes (normal, large, flying) and the importance of the style point score varies across the hill sizes because of a different calculation of the final score. For example, in our data, the share of style points from the final score is 45%, 44%, and only 30%

for normal, large, and flying hills, respectively. This means, that the judges' contribution to the final outcome is less important at flying hill competitions, which may lead to differences in biased behavior. However, the results in Column 3 show that the judges' nationalistic bias, again, does not significantly differ between the different hill sizes.

Finally, we test whether nationalistic bias differs between different types of competitions. Thus, we compare the nationalistic bias at the Olympic Games and the World Championships, which are the most prestigious competitions in professional ski jumping, followed by the Four Hills type of competitions and the ordinary World Cup competitions.[14] According to Sandberg (2018), the nationalistic bias is stronger for competitions with a national character because national identity becomes more salient. Thus, we would expect the highest bias in the Olympic Games and World Championships and the lowest one in the ordinary World Cup events. Although we observe the largest estimate of a nationalistic bias for the Olympic Games in Table 5, the results in Column 4 of Table 6 show no significant difference between the different types of competitions due to large standard errors. However, if the difference between the World Cups, World Championships, and the Four Hills competitions is not significant both statistically and economically, the differential effect of 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 explore possible reasons for large standard errors that can be observed for both Olympic Games-related coefficients in Tables 5 and 6. This can obviously be explained by the lowest number of observations in these competitions. However, one additional explanation can be the variance of judges' countries in terms of their taste for nationalistic favoritism. In general, since the athletes' performances must be evaluated objectively and

[14] Four Hills is a ski jumping event composed of four World Cup events and has taken place in Germany and Austria each year since 1953. Winning all four events in one Four Hills Tournament edition is called the grand slam. For additional information, see https://en.wikipedia.org/wiki/Four_Hills_Tournament (last accessed on 16.10.2020).

without any prejudice, such a favoritism can be described as a corrupt-type of behavior. To illustrate, in Table 7, we present the results of model (1) for the 2014 Sochi Olympic Games and only for subsamples of jumpers from countries whose judges were part of the panel.[15] Because of data constraints, we could neither use judge nor judge-per-season fixed effects. Overall, the results suggest that Russia has the largest nationalistic bias for the 2014 Sochi Olympic Games, which is also the only significant one at the 1% level. This bias is 70% larger than the one estimated for the second highest country's estimator (Switzerland) and 324% larger than the third highest country (Norway). However, we should be very cautious before drawing any conclusions based on such a small number of observations.

Table 7: Judge-specific variation of nationalistic bias at the 2014 Sochi Olympic Games

Country	No. of obs.	Coefficient	Standard error	*p*-Value
France	10	-0.125	0.090	0.396
Italy	20	0.125*	0.045	0.071
Norway	75	0.150	0.082	0.128
Switzerland	40	0.375*	0.174	0.083
Russia	55	0.636***	0.134	0.005
Austria	70	0.047	0.071	0.536

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) without judge-per-season fixed effects for the performances of all ski jumpers from the respective countries at the 2014 Sochi Olympic Games. Standard errors are two-way clustered at the judge and jump level. ***, **, * denote significance at the 1%, 5%, 10% level, respectively.

To test more carefully whether (and why) any nationalistic bias might vary by country, we use all data and explore the relationship between the countries' specific nationalistic bias and the Corruption Perceptions Index (CPI). First, in Figure 3, we present the results of 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

[15] See Appendix D for the full description of judges and jumpers in the 2014 Sochi Olympic Games.

highest nationalistic bias estimator (0.22). Out of the 12 countries, Norway (0.00) and Finland (0.01) are the only two countries whose coefficients are negligibly small, both economically and statistically.[16]

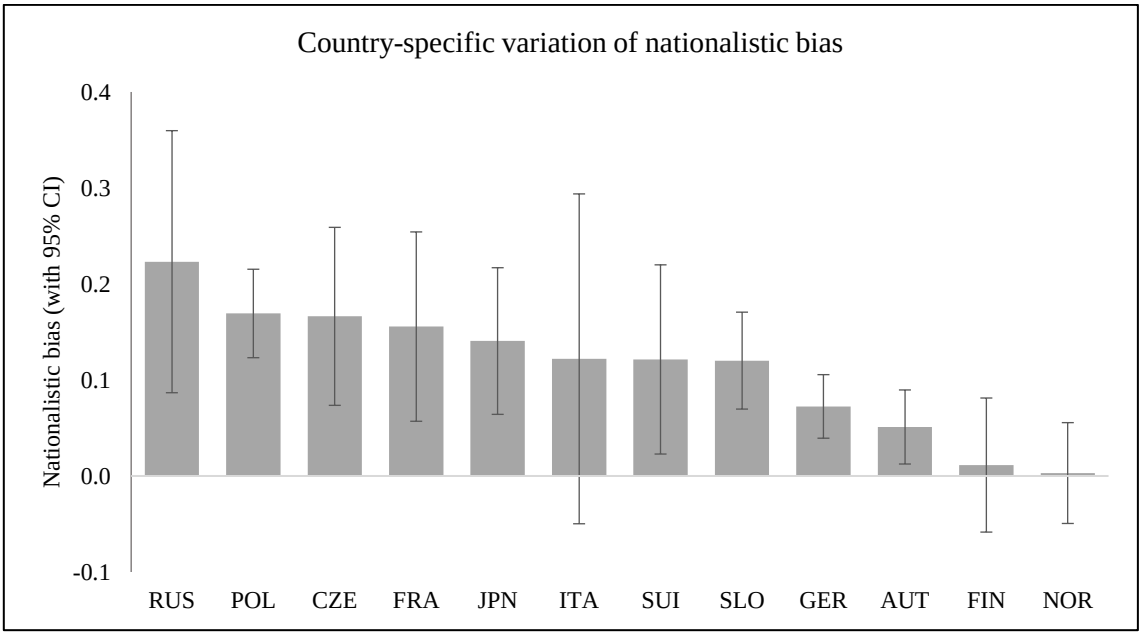


Figure 3: The figure shows the average nationalistic bias with 95% confidence intervals of judges when they evaluate performances of their compatriot jumpers. The estimates are based on subsample estimations of model (1) without judge-per-season fixed effects for the performances of all ski jumpers from 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.

In Figure 4, we demonstrate a negative relationship between the nationalistic bias and the countries' CPI score for the performances of all ski jumpers from the respective countries. In other words, the higher the CPI (less corrupt country), the lower the nationalistic bias.[17]

[16] Note that Italian judges participated in only 21% of competitions compared to 59% and 53% of Norwegian and Finnish judges, respectively. For additional details, see Appendix A.

[17] Note that we could not estimate the average nationalistic bias for three countries in our dataset. These countries had too few ski jumper performance observations, which resulted in no variation or too few clusters. Sweden had only two jumps from two ski jumpers; Slovakia had two jumps from one ski jumper, and Romania had three jumps from two ski jumpers.

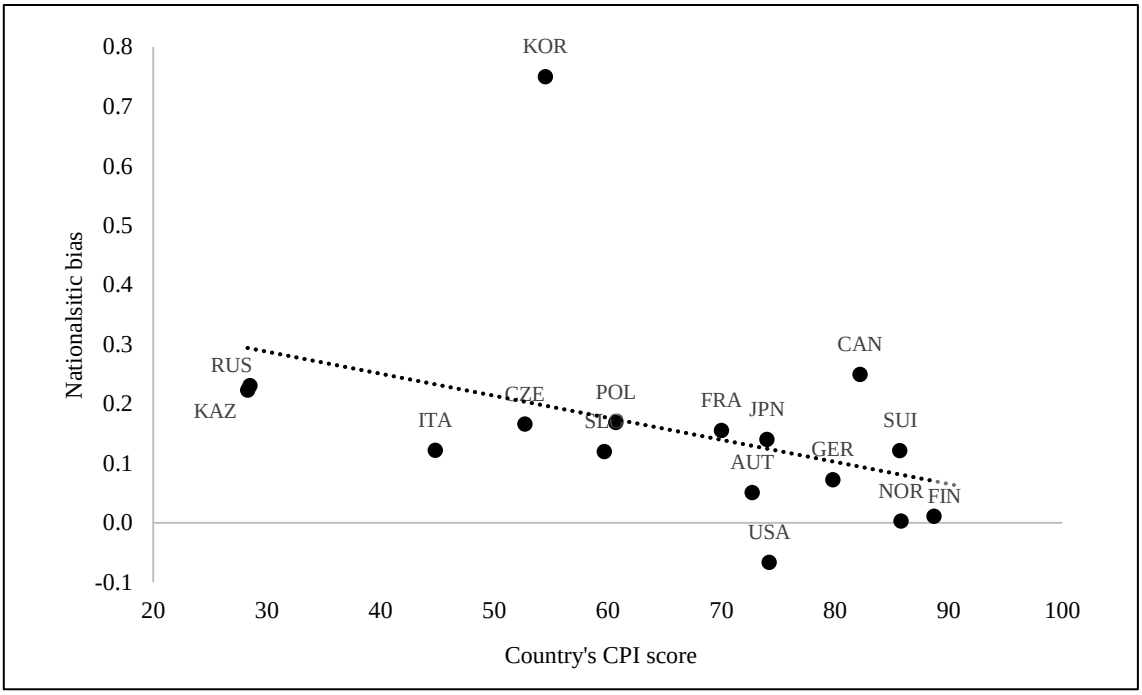


Figure 4: The figure shows the average nationalistic bias of compatriot judges from different countries relative to the country's CPI score. The estimates are based on subsample estimations of model (1) without judge-per-season fixed effects for the performances of all ski jumpers from the respective countries. The dashed line depicts the linear relationship between the size of bias and the CPI score.

To probe more deeply into the relationship between the CPI and nationalistic bias, we take one step beyond Zitzewitz (2006), who only reported a simple negative cross-country correlation between the CPI and judges' nationalistic bias without estimating its effect in a regression analysis. We add on his analyses and estimate model (1) by using all the data and interacting between the CPI and *compatriot jumper*. In Column 1 of Table 8, we see a negative sign of the interaction term, suggesting that the nationalistic bias is higher the more corrupt the country is (the lower the CPI). 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% of the within-jump standard deviation of style points' evaluation. In Column 2, we exclude South Korea as an extreme outlier (see Figure 4) and three countries (Sweden, Slovakia, and Romania) that only had few ski jumper performance observations (see Appendix A for more details). Our results are robust to exclusion of these countries.

Table 8: Nationalistic bias and the Corruption Perceptions Index (CPI)

	(1)	(2)
Compatriot jumper (CJ)	0.230***	0.228***
	(0.050)	(0.050)
CJ x CPI score	-0.002***	-0.002***
	(0.001)	(0.001)
Jump FE	Yes	Yes
Judge-per-season FE	Yes	Yes
No. of observations	76,775	76,505

Note: The dependent variable is the style points of each individual judge for a given jump. Standard errors are three-way clustered at the judge, jumper, and jump level and presented in parentheses. In Column 2, we exclude observations from four countries: KOR as an extreme outlier and SWE, SVK, and ROU due to too few ski jumper performance observations. ***, **, * denote significance at the 1%, 5%, 10% level, respectively.

In addition, since Russia had the highest estimator of 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 Games and not by Russia per se. To obviate this concern, we remove the data of the Olympic Games and perform analyses similar to those in Figures 3 and 4 and Table 8. The results presented in Appendixes E-G 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 (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

Given that we ask a question similar to several previous studies, and even use the same sport as one of those studies, it is natural to compare the results. This is particularly important because of the increasing voices in scientific society regarding the need for replication studies (Ioannidis and Doucouliagos, 2013; Open Science Collaboration, 2015) and crowdsourced

research (Silberzahn and Uhlmann, 2015; Silberzahn et al., 2018). Thus, we compare the magnitude of nationalistic and compensating biases in our paper with the biases in ski jumping, figure skating, and dressage, as reported in Zitzewitz (2006) and Sandberg (2018), respectively.

First, we compare between the baseline results on nationalistic bias in our study, as presented in Column 1 of Table 3, with the results of the baseline specifications in Zitzewitz (2006) on ski jumping (Line 1 of Panel A in Table 4) and figure skating (Line 1 of Panel B in Table 4), and with results in Sandberg (2018) on dressage (Column 1 of Table 3). Given the different scale of scores between the sports, we cannot compare the estimators of nationalistic bias. To make a credible comparison, we will therefore compare the estimators of nationalistic bias as a share of the within- and overall-jump standard deviations of style points.

In Table 9, we see that our estimator of nationalistic bias is equal to 8% of the overall standard deviation and to 29% of the within-standard deviation. We see that in the other ski jumping-related study (Zitzewitz, 2006), the nationalistic bias estimator as a share of the standard deviations is higher. The possible reason for such a difference is that Zitzewitz (2006) used data only on the year 2002, which was the Olympic year. As such, his study is concentrated much more around the Olympic Games, which may be less representative since the Olympic Games only take place once in four years. This would not be a problem if the bias in the Olympic Games did not differ from the other competitions. However, similar to our results, the nationalistic bias in the Olympic Games in Zitzewitz (2006) is the highest (0.258) and close to our estimator (0.217). Thus, in both papers, the Olympic Games increase the average estimator of nationalistic bias in ski jumping. However, in our case, the share of the Olympic Games is only 1% of the overall number of observations, which is much lower than in the case of Zitzewitz (2006), where the share of the observations in the Olympic Games is 13.3% of the

overall number of observations.[18] In addition, our dataset includes information on the wind and gate compensation points that were absent before 2010. These wind and gate compensation points provide us with the opportunity to better control for the objective quality of the jump.

Table 9: Comparison of nationalistic bias among studies and sports

	(1)	(4)	(5)	(6)
	Our estimations	Zitzewitz (2006)		Sandberg (2018)
Sport	Ski jumping	Ski jumping	Figure skating	Dressage
Reference	Table 3, Column 1	Table 4, Panel A, Line 1	Table 4, Panel B, Line 1	Table 3, Column 1
Nationalistic bias coefficient	0.09	0.15	0.17	0.36
Overall SD of performance score	1.07	1.09	1.28	5.02
Within-SD of performance score	0.31	0.33	0.35	1.51
Relative bias to overall SD	0.08	0.14	0.13	0.07
Relative bias to within SD	0.29	0.45	0.49	0.24

Contrary to Zitzewitz's (2006) data on ski jumping, Sandberg's (2018) data have a wider coverage in terms of years and competitions. More specifically, her paper uses more than 90,000 observations from seven years of competitions that do not include the Olympic Games. In that regard, our data (76,775 observations from seven years) is much more similar to Sandberg (2018) than to Zitzewitz (2006). This might be the reason why our results are closest to the results in dressage, whose coefficient of nationalistic bias accounts for 7% and 24% of the overall and within-jump standard deviations, respectively.[19]

When comparing our results to figure skating (Zitzewitz, 2006), we find a similar deviation to that of ski jumping. In that case, the possible reason is a high variation between the

[18] For example, in estimations without Olympic Games presented in Appendix G, the magnitude of nationalistic bias is lower compared to our Table 8, which included data on 2014 Sochi Olympic Games.
[19] When estimating model (1) with judge fixed effects as in Sandberg (2018), our results remain the same.

disciplines. For example, the more subjective ice dancing discipline accounts for one third of the data on figure skating. More importantly, it has the highest estimator of nationalistic bias, which is 33% higher than non-ice dancing disciplines (see Table 5 in Zitzewitz, 2006). The author states that "biases are larger where scoring is more subjective, as it is for ice dancing, where skaters do not have as many mandatory deductions for falls, and for artistic impression as opposed to technical merit scores" (Zitzewitz, 2006, p. 79). That is in line with a recent paper by Joustra et al. (2020) who found a significant advantage of later performances in female gymnastics, which is likely to be driven by the existence of subjective evaluation only in female competitions considering artistry.

Turning to compensating bias, we showed that close to 85% of the judges do not exhibit any significant bias, which is supported by the highly insignificant estimator whose magnitude is also very close to zero. Moreover, the estimator of nationalistic bias is not affected by the inclusion of compensating bias when controlling for home advantage. In that regard, our results deviate from the findings in figure skating and dressage and somewhat from the findings of Zitzewitz (2006) on ski jumping, none of which included the home variable (though they reported that their results are robust to exclusion of home participants). More specifically, in several specifications, Zitzewitz (2006) finds a negative and significant compensating bias in ski jumping. However, when he uses jump observable characteristics as covariates, his estimator of compensating bias loses half of its magnitude and becomes statistically insignificant, (Lines 4b and 5b of Panel A in Table 4), which is similar to our finding.

Our results deviate from the findings on positive indirect bias in figure skating (Zitzewitz, 2006). The possible reason for this is the difference in institutional settings that relate to truncation and exposure of scores. The point is that figure skating uses a random truncation of scores, which according to Emerson (2007) may add noise to the results. In addition, judge anonymity was introduced in 2002, according to which it was not possible to match the score

and the identity of the judge. Such an anonymity could also have an effect on the willingness to be involved in strategic voting.

Finally, our results differ from the results of Sandberg (2018), who, similarly to figure skating, showed a positive indirect bias in dressage competitions. Beyond the differences that relate to inclusion of the home variable, an additional possible reason for the differences between results may stem from the fact that dressage competitions have a rule that 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% among the judges. Thus, it is possible that experienced dressage judges anticipate the nationalistic bias of their panel members and act accordingly, showing compensating bias to ensure consistency. This is different in ski jumping, where the truncation mechanism provides no incentive either for compensation, because the extreme votes are excluded, or for consistency, because there is no such 5% rule.

8 Conclusion

This paper was inspired by two streams in the literature. The first is on nationalistic bias in subjective evaluations by international experts (Zitzewitz, 2006; Sandberg, 2018). The second is on 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).

Unlike previous findings, our results show no evidence of strategic voting, according to which judges assign significantly different scores to jumpers whose compatriots are present on the judging panel. However, in line with previous literature, our results indicate that well-trained professional experts are prone to nationalistic bias more than a decade after that bias was first illustrated in similar settings. It suggests that in-group favoritism is a very strong

feature of human behavior, especially in countries with a high prevalence of corruption in the institutional environment.

It is important to note that our results were obtained from fully observable sports competitions. Therefore, such an in-group favoritism may 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. That 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 may lead to an increased nationalistic favoritism.

Acknowledgements

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 performance evaluation in professional ski jumping. We also thank the participants of the 10th Annual Meeting of the European Sport Economics Association in Liverpool, 42nd Meeting of the Norwegian Association of Economists in Ås, Reading Online Sport Economics Seminar, and the 1st Molde Sports Analytics workshop in Molde University College for their helpful comments and suggestions. All errors are our own.

References

Andrews, T.J., Smith, R.K., Hoggart, R.L., Ulrich, P.I., Gouws, A.D., 2019. Neural correlates of group bias during natural viewing. Cerebral Cortex 29 (8), 3380-3389.

Barnett, V., 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 (1), 39-50.

Barr, A., Serra, D., 2010. Corruption and culture: An experimental analysis. Journal of Public Economics 94 (11-12), 862-869.

Efferson, C., Lalive, R., Fehr, E., 2008. The coevolution of cultural groups and ingroup favoritism. Science 321 (5897), 1844-1849.

Elaad, G., Krumer, A., Kantor, J., 2018. Corruption and sensitive soccer games: Cross-country evidence. The Journal of Law, Economics, and Organization 34 (3), 364-394.

Emerson, J.W., 2007. Chance, on and off the ice. Chance 20 (2), 19-21.

Faltings, R., Krumer, A., Lechner, M., 2019. Rot-Jaune-Verde. Language and favoritism: Evidence from Swiss soccer, University of St. Gallen, School of Economics and Political Science. Working paper. Retrieved from: http://ux-tauri.unisg.ch/RePEc/usg/econwp/EWP-1915.pdf

Fédération Internationale de Ski (FIS), 2017a. Rules for the FIS ski jumping world cup (men), edition 2017/2018. Oberhofen, CH.

Fédération Internationale de Ski (FIS), 2017b. The international ski competition rules (ICR), Book III, Ski jumping. Oberhofen, CH.

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

Garicano, L., Palacios-Huerta, I., Prendergast, C., 2005. Favoritism under social pressure. Review of Economics and Statistics 87 (2), 208-216.

Gächter, S., Schulz, J.F., 2016. Intrinsic honesty and the prevalence of rule violations across societies. Nature 531 (7595), 496-499.

Genakos, C., Pagliero, M., 2012. Interim rank, risk taking, and performance in dynamic tournaments. Journal of Political Economy 120 (4), 782-813.

Genakos, C., Pagliero, M., Garbi, E., 2015. When pressure sinks performance: Evidence from diving competitions. Economics Letters 132, 5-8.

Harb-Wu, K., 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., Doucouliagos, C., 2013. What's to know about the credibility of empirical economics?. Journal of Economic Surveys 27 (5), 997-1004.

Joustra, S.J., Koning, R.H., Krumer, A., 2020. Order effects in elite gymnastics. De Economist, forthcoming.

Koning, R.H., 2011. Home advantage in professional tennis. Journal of Sports Sciences 29 (1), 19-27.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349 (6251), aac4716.

Page, K., Page, L., 2010. Alone against the crowd: Individual differences in referees' ability to cope under pressure. Journal of Economic Psychology 31 (2), 192-199.

Pope, B.R., Pope, N.G., 2015. Own-nationality bias: Evidence from UEFA Champions League football referees. Economic Inquiry 53 (2), 1292-1304.

Pope, D.G., Price, J., Wolfers, J., 2018. Awareness reduces racial bias. Management Science, 64(11), 4988-4995.

Price, J., Remer, M., Stone, D.F., 2012. Subperfect game: Profitable biases of NBA referees. Journal of Economics & Management Strategy 21 (1), 271-300.

Price, J., Wolfers, J., 2010. Racial discrimination among NBA referees. The Quarterly Journal of Economics 125 (4), 1859-1887.

Sandberg, A., 2018. Competing identities: A field study of in-group bias among professional evaluators. The Economic Journal 128 (613), 2131-2159.

Scholten, H., Schneemann, S., Deutscher, C., 2020. The impact of age on nationality bias and cultural proximity bias: Evidence from ski jumping. Journal of Institutional and Theoretical Economics, forthcoming.

Shayo, M., Zussman, A., 2011. Judicial ingroup bias in the shadow of terrorism. The Quarterly Journal of Economics 126 (3), 1447-1484.

Silberzahn, R., Uhlmann, E.L. (2015). Crowdsourced research: Many hands make tight work. Nature News 526 (7572), 189-191.

Silberzahn, R., Uhlmann, E.L., Martin, D.P., Anselmi, P., Aust, F., Awtrey, E., ..., Carlsson, R. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. Advances in Methods and Practices in Psychological Science 1 (3), 337-356.

Spierdijk, L., Vellekoop, M., 2009. The structure of bias in peer voting systems: Lessons from the Eurovision Song Contest. Empirical Economics 36 (2), 403-425.

Sumner, W.G., 1906. Folkways: A Study of the Sociological Importance of Usages, Manners, Customs, Mores, and Morals. Boston: Ginn and Company.

Waguespack, D.M., Salomon, R., 2015. Quality, subjectivity, and sustained superior performance at the Olympic Games. Management Science 62 (1), 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 (2), 166-183.

Zitzewitz, E., 2006. Nationalism in winter sports judging and its lessons for organizational decision making. Journal of Economics & Management Strategy 15 (1), 67-99.

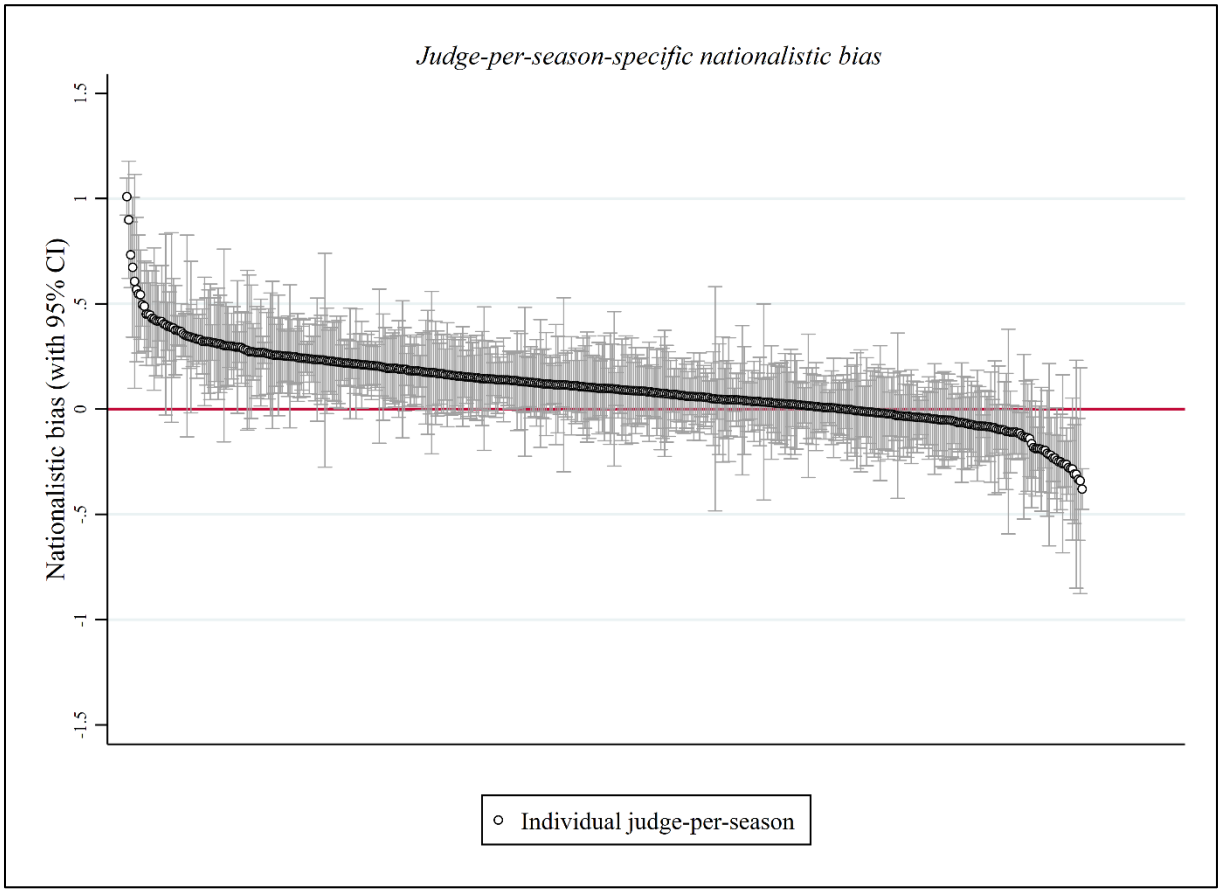
Zitzewitz, E., 2014. Does transparency reduce favoritism and corruption? Evidence from the reform of figure skating judging. Journal of Sports Economics 15(1), 3-30.

Appendix A: Frequencies of countries by groups of jumpers, judges, and competitions

Country name	Country code	Jumpers	Jumps	Judges	Competitions	Judges in competitions
Austria	AUT	27	2077	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	2098	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	2037	12	38	120 (59%)
Poland	POL	20	1666	8	16	79 (39%)
Romania	ROU	2	3	4	0	8 (4%)
Russia	RUS	18	688	6	8	38 (19%)
Slovenia	SLO	27	1774	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 competitions	

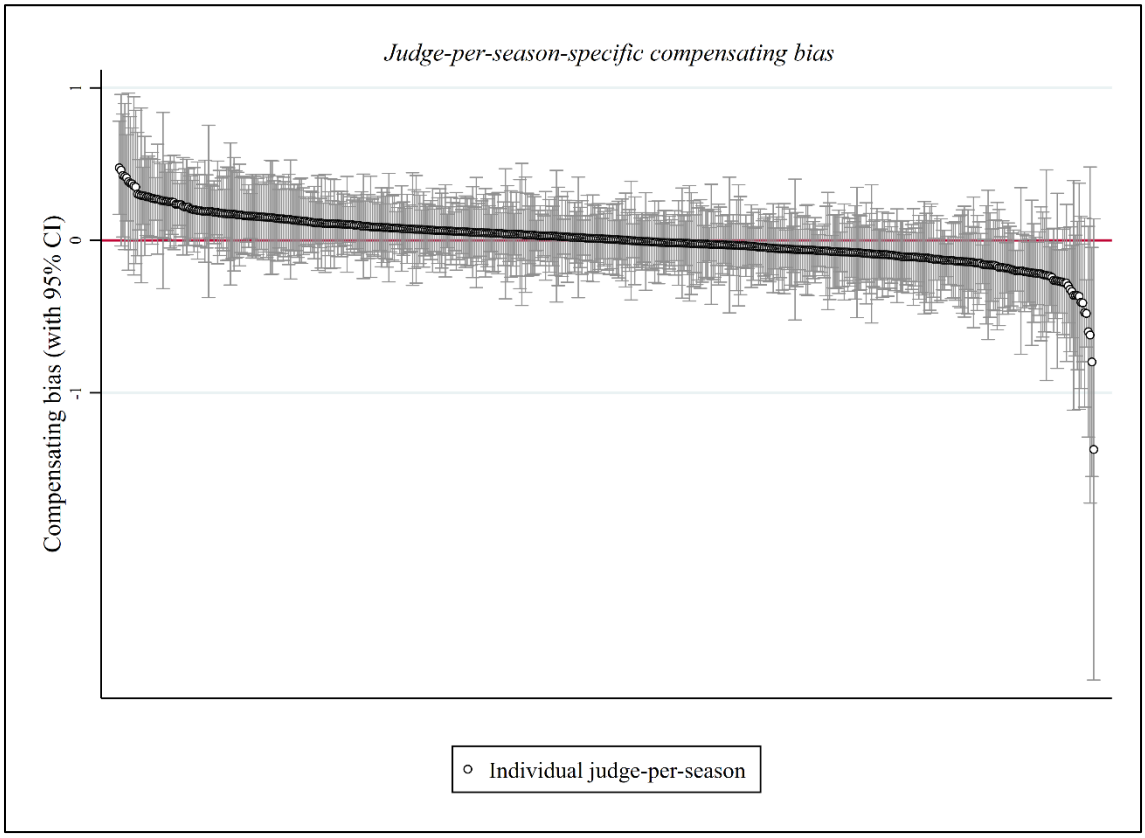
Note: The last column states the number of competitions in which the respective country has a judge on the panel. There are five judges in each competition. This is also presented as percentage share based on the total number of competitions in parentheses.

Appendix B: Judge-per-season-specific nationalistic bias



Note: The figure shows the judge-per-season-specific nationalistic bias with 95% confidence intervals. The estimates are the judge-per-season-specific coefficients from estimating model (1).

Appendix C: Judge-per-season-specific compensation bias



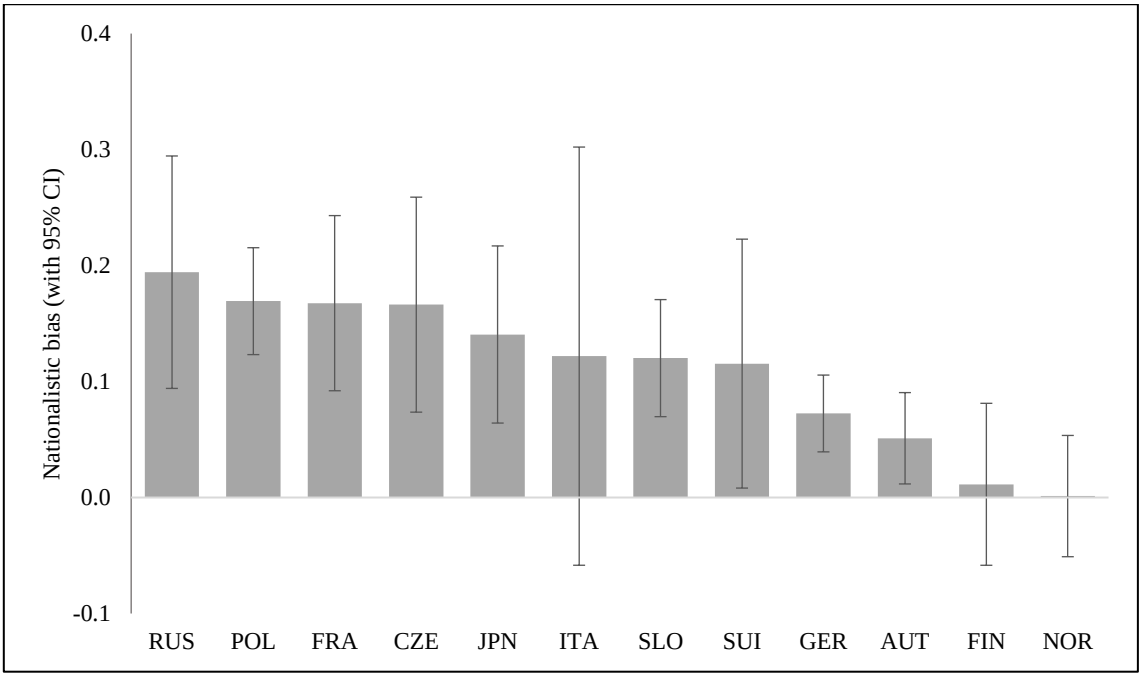
Note: The figure shows the judge-per-season-specific compensating bias with 95% confidence intervals. The estimates are the judge-per-season-specific coefficients from estimating model (2).

Appendix D: Frequencies of countries by groups of jumpers, jumps, and judges in the 2014 Sochi Olympic Games

Country name	Country code	Jumpers	Jumps	Judges	Judges in competitions
Austria	AUT	4	14	1	1 (50%)
Bulgaria	BUL	1	1	0	0 (0%)
Canada	CAN	3	5	0	0 (0%)
Czech Republic	CZE	5	15	0	0 (0%)
Estonia	EST	1	2	0	0 (0%)
Finland	FIN	4	12	0	0 (0%)
France	FRA	1	2	1	2 (100%)
Germany	GER	5	13	0	0 (0%)
Italy	ITA	1	4	1	2 (100%)
Japan	JPN	5	16	0	0 (0%)
South Korea	KOR	3	5	0	0 (0%)
Norway	NOR	4	15	1	2 (100%)
Poland	POL	5	14	0	0 (0%)
Russia	RUS	5	11	1	2 (100%)
Slovenia	SLO	4	13	0	0 (0%)
Switzerland	SUI	2	8	1	1 (50%)
USA	USA	4	5	0	0 (0%)
Total	17	57	155	6	2 competitions

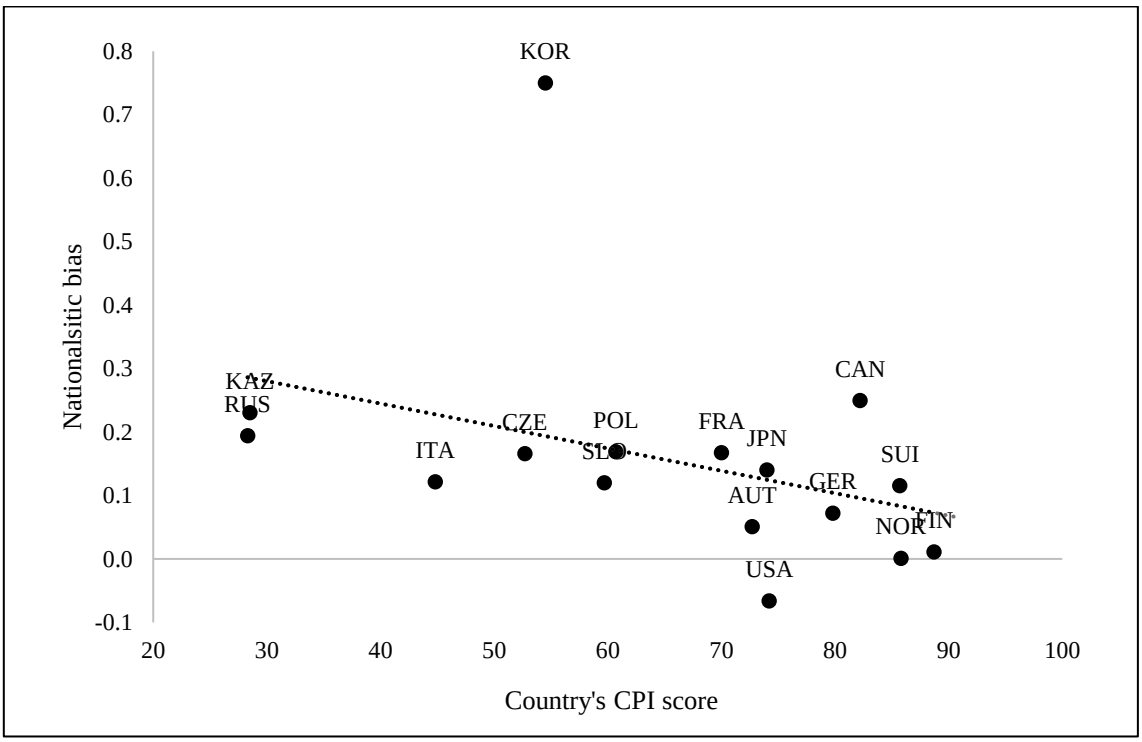
Note: The Olympic Games include two competitions for individual ski jumpers on normal and large hills.

Appendix E: Country variation of nationalistic bias without Olympic Games



Note: The figure shows the average nationalistic bias with 95% 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, excluding the Olympic Games. The 12 countries are those with the most performance observations. The order of countries is based on the size of nationalistic bias.

Appendix F: Relationship between the nationalistic bias and the Corruption Perceptions Index (CPI) without Olympic Games



Note: The figure shows the average nationalistic bias of compatriot judges from different countries relative to the country's CPI score, excluding Olympic Games data. 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 dashed line depicts the linear relationship between the size of bias and the CPI score.

Appendix G: Country-specific variation of nationalistic bias without Olympic Games

	(1)	(2)
Compatriot jumper (CJ)	0.220***	0.218***
	(0.046)	(0.046)
CJ x CPI score	-0.002***	-0.002***
	(0.001)	(0.001)
Jump FE	Yes	Yes
Judge-per-season FE	Yes	Yes
No. of observations	76,000	75,755

Note: The dependent variable is the style points of each individual judge for a given jump. Standard errors are three-way clustered at the judge, jumper, and jump level and presented in parentheses. In Column 2, we exclude observations from four countries: KOR as an extreme outlier and SWE, SVK, and ROU due to too few ski jumper performance observations. ***, **, * denote significance at the 1%, 5%, 10% level, respectively.