14TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 6 - 7, 2020 | BOSTON, MA
presented by ESPN

Accuracy and National Bias of Figure Skating Judges: The Good, the Bad and the Ugly

Paper Track: Business of sports
ID 1548706

Abstract

Figure skating has had its share of judging controversies in the last twenty years. The last in line is the suspension of two Chinese judges suspected by the International Skating Union (ISU) of preferential marking in favor of Chinese skaters during the 2018 PyeongChang Olympic Winter Games. In this work we develop novel mathematical techniques to monitor the accuracy and nationalistic bias of figure skating judges. This is fundamental to guarantee a level playing field in this sport. Our analysis reveals systemic nationalistic bias, and although both suspended Chinese judges were undoubtedly biased, they were far from the only ones, nor were they the worst offenders. We also shed light on the current ISU monitoring practices and propose recommendations moving forward.

1. Introduction

Figure skating programs are evaluated by panels of judges, who must evaluate the quality of different elements and components based on precise but sometimes subjective criteria defined in scoring regulations and codes of points. Unlike other sports with similar evaluation systems such as gymnastics, diving and ski jumping, figure skating has had a larger share of judging scandals and controversies. The most notorious occurred during the 2002 Salt Lake City Olympic Winter Games. French Judge Marie-Reine Le Gougne, pressured by the head of the French Federation of Ice Sports, allegedly favored the Russian pair of Elena Berezhnaya and Anton Sikharulidze to the detriment of the Canadian pair of Jamie Salé and David Pelletier, in exchange for a favorable evaluation of the French pair of Marina Anissina and Gwendal Peizerat at the upcoming ice dance competition.

The 2002 judging controversy triggered the development and adoption of the current and more objective judging system. Marks given by judges in the new system were initially kept anonymous to make collusion more difficult, but this introduced new problems including a lack of accountability and the impossibility for third parties to monitor judges. After further controversies at the 2014 Sochi Olympic Winter Games, the ISU abolished judging anonymity in 2016. The most recent controversy occurred during the 2018 PyeongChang Olympic Winter Games. Two Chinese judges, Huang Feng and Chen Weiguang, were accused of favoring Chinese skaters and suspended by the ISU. All these aforementioned events and many others have received significant media exposure, leading to a lot of grumbling about the lack of objectivity of judging in figure skating and the questionable monitoring practices of the ISU.

National bias in figure skating is well-documented in the scientific community [5, 8, 9] and also appears in other sports such as ski jumping, gymnastics, Muay Thai boxing, diving and dressage. Judges have the tendency to favor athletes of the same nationality, while simultaneously penalizing their competitors. National bias can be estimated using various techniques such as sign tests, permutation tests, linear regressions and fixed effects. Zitzewitz [8, 9] showed that national bias in figure skating increases with the importance of the event. He also showed the existence of vote trading between judges of different nationalities.

In this article we develop and apply novel statistical tools to study the accuracy and biases of figure skating judges at the 2018 PyeongChang Olympic Winter Games, with an emphasis on the program component scores. We extend the state of the art in many ways. First, we calculate the intrinsic judging error variability, quantifying precisely the errors made by judges as a function of the performance level. Judges are more accurate for the best performances than for mediocre ones. While prior work focused on national bias, we leverage the intrinsic judging error variability to quantify the accuracy of the judges compared to their peers. This is important because judging is very hard and some judges are measurably much better than others. We then use the intrinsic judging error variability and the accuracy of each judge in the calculation of its national bias. Intuitively this makes sense for two reasons: (1) for the best skaters, judging marks are closer to each other, and a small absolute bias is sufficient to impact the rankings of the athletes; and (2) we can distinguish precise but biased judges from erratic but unbiased judges. We supplement the accuracy and bias measurements with confidence intervals based on the number of observations. The more observations we have, the more confidence we have that a good (or bad) evaluation is representative of the performance of the judge and not due to random factors such as good (or bad) luck. Our main conclusions are:

1. **The best judges are two to three times more accurate than the most erratic judges.**
2. **Nationalistic bias is endemic, and for many judges larger than all the other sources of judging errors.**
3. **Current ISU judge monitoring is utterly inadequate.**
4. **Other well-known issues such as gender and conformity biases are nonexistent.**
5. **The ISU can solve most of its judging problems with mathematically sound long-term monitoring.**

The remainder of this article is organized as follows. We describe the ISU judging system in Section 2. We present our dataset in Section 3. We discuss the ISU monitoring practices in Section 4. We present our methods in Section 5, followed by our results in Section 6. Finally, we discuss the limitations of our approach in Section 7 and conclude with recommendations in Section 8.

2. Current ISU judging system

The current ISU Judging System, also called the Code of Points system, replaced the old 6.0 system in 2004. It has two main parts: the Grade of Execution (GOE) and the Program Component Scores (PCS). Each part is evaluated by a single panel of nine judges[1].

Grade of Execution (GOE). Each technical element of a program (e.g. jump, spin, sequence of steps) has a base value depending on its difficulty. For each of these elements, panel judges evaluate the quality of its execution by giving a mark between -5 to +5 in increments of 1[2]. Judges must mark based on precise criteria defined in the Code of Points. The GOE of the element is the trimmed mean of the middle seven execution marks given by the judges, which is then added to its base value.

Program Component Scores (PCS). Judges must also provide five Program Component Scores: skating skills, transitions, performance, composition and interpretation[3][4] . These component scores superseded the presentation mark of the former 6.0 system. For each component, judges give a mark between 0 and 10 in 0.25 increments, and the final score is the trimmed mean of the middle seven marks. The sum of the five component scores is the PCS of the program[5].

In this work we focus our analysis on the PCS, and leave the GOE for future work.

[1] Lower-level competitions generally have smaller panels.

[2] The range was -3 to +3 in our dataset for the 2018 PyeongChang Olympic Winter Games. The -5 to + 5 range was introduced for the 2018--2019 season.

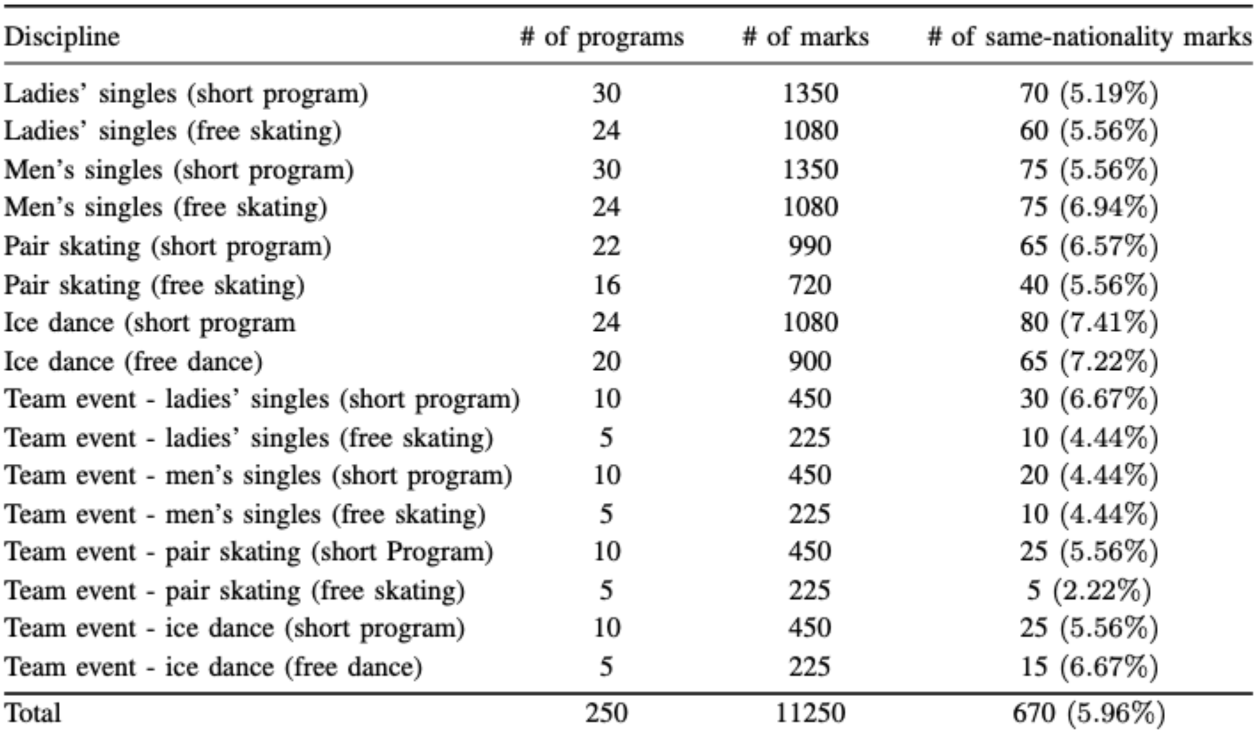
[3] Ice dance has a slightly different setup.

[4] More details about each component can be found at http://www.isuresults.com/results/season1718/owg2018/OWG2018_protocol.pdf .

[5] In ice dance the components do not have the same weight.

3. Dataset

The data from the 2018 PyeongChang Olympic Winter Games comes from the ISU and includes the Program Component Scores of all 250 programs[6]. Table I shows the size of the dataset by event. A panel of nine judges individually evaluate the five program components; this results in 45 marks per program for a total of 11250 observations in our dataset. The judge and the athlete share the same nationality in ≈ 6% of observations.



Discipline	# of programs	# of marks	# of same-nationality marks
Ladies' singles (short program)	30	1350	70 (5.19%)
Ladies' singles (free skating)	24	1080	60 (5.56%)
Men's singles (short program)	30	1350	75 (5.56%)
Men's singles (free skating)	24	1080	75 (6.94%)
Pair skating (short program)	22	990	65 (6.57%)
Pair skating (free skating)	16	720	40 (5.56%)
Ice dance (short program	24	1080	80 (7.41%)
Ice dance (free dance)	20	900	65 (7.22%)
Team event - ladies' singles (short program)	10	450	30 (6.67%)
Team event - ladies' singles (free skating)	5	225	10 (4.44%)
Team event - men's singles (short program)	10	450	20 (4.44%)
Team event - men's singles (free skating)	5	225	10 (4.44%)
Team event - pair skating (short Program)	10	450	25 (5.56%)
Team event - pair skating (free skating)	5	225	5 (2.22%)
Team event - ice dance (short program)	10	450	25 (5.56%)
Team event - ice dance (free dance)	5	225	15 (6.67%)
Total	250	11250	670 (5.96%)

Table I. Sample size by event.

[6] The dataset is available at http://www.isuresults.com/results/season1718/owg2018/OWG2018_protocol.pdf .

4. Current ISU monitoring

The ISU Communication No. 2098[7] describes the ISU internal judging monitoring processes. The approach is based on "deviation points", i.e., the difference between the mark given by a judge and the panel average for each component. A difference in absolute value of less than 1.5 per component is acceptable. For the overall program, a total deviation (the sum of the five component deviations) of up to 7.5 is tolerated. Differences outside these intervals are subject to further evaluation. This approach has many shortcomings:

1. The thresholds appear arbitrary and large: in our dataset they are transgressed only for the evaluation of the short program of Israeli skater Alexei Bychenko by the Finnish judge Pekka Leskinen.
2. Negative and positive component deviations cancel each other when calculating the total deviation, thus the approach cannot detect unbiased but inaccurate judges.
3. The average mark is sensitive to outliers; it underestimates the deviation of a highly biased or inaccurate judge, and overestimates the deviation of other accurate judges on the same panel.
4. It ignores the intrinsic judging error variability: a deviation of 1.5 for a component whose true value is 9.5 is significantly worse than for a component worth 5.0.

Given these shortcomings, which do not allow to properly identify erratic or biased judges, the FIG gives credence to critics saying it applies disciplinary measures arbitrarily[8]. Consider the Disciplinary Commission for the Chinese judge Huang Feng, one of the judges suspended by the ISU. A quick look at the report[9] shows that when considering the deviation points approach of the ISU, he is not even the most biased judge on the panel that led to his suspension. This dubious honor belongs to the German judge Elke Treitz. Figure 1 shows the total net deviation points for the Chinese and German judges for the first four pairs of the short program. The Chinese judge favors the Chinese pair and penalizes the Russian, Canadian and German pairs, whereas the German judge favors the German pair and penalizes the Chinese, Russian and Canadian pairs. We observe that the German judge, while not as generous with German skaters as the Chinese judge is with Chinese skaters, is much harsher with their competitors. Note that none of the individual and overall deviations are close to the 1.5 and 7.5 ISU monitoring thresholds warranting further evaluation.

[7] https://www.isu.org/inside-isu/isu-communications/communications-archives/593-isu-communication-2098/file

[8] Consult, for instance, http://www.globetrottingbyphiliphersh.com/home/2018/7/14/in-sort-of-suspending-a-skating-judge-international-federation-mocks-fans-with-ethical-relativism, https://www.nbcnews.com/storyline/winter-olympics-2018/u-s-judges-give-u-s-skaters-higher-marks-pyeongchang-n850006 and https://www.buzzfeednews.com/article/johntemplon/the-edge.

[9] The report is available at https://www.isu.org/inside-isu/legal/disciplinary-decisions/17359-case-2018-02-isu-vs-huang/file.

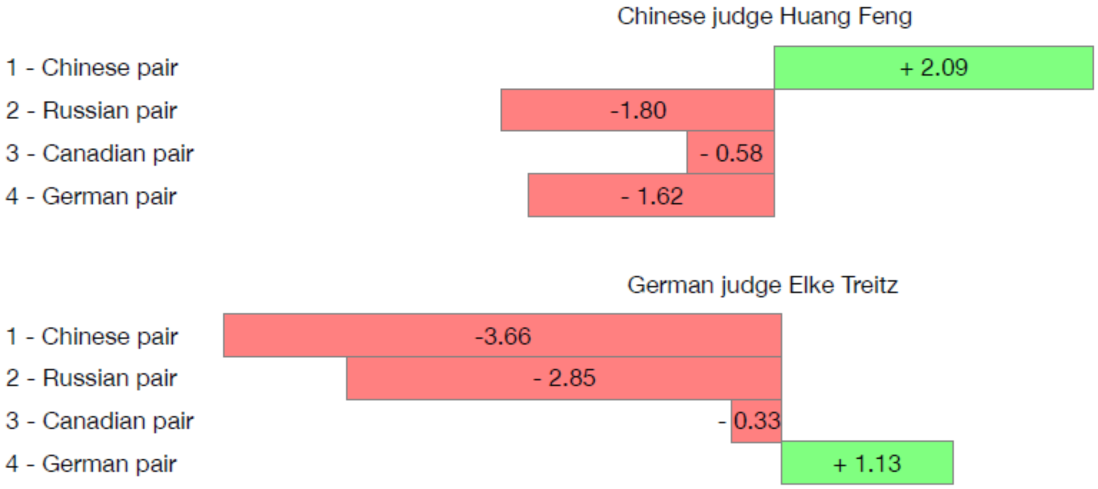


Figure 1 : Total net deviation points (ISU method) for the Chinese and German judges for the first four pairs of the short program (Program Component Scores).

Even though Huang Feng received a letter of warning from the ISU Officials Assessment Commission three weeks before the 2018 Olympics, it is not clear why he and his compatriot Chen Weiguang were singled out by the ISU. To the best of our knowledge, Elke Treitz did not receive any blame, nor did Finnish judge Pekka Leskinen, who was beyond the thresholds but had the benefit of being only incompetent since there was no Finnish skater at the Olympics. Our statistical analysis in the next sections provides more detailed insight.

5. Methods

In this section we first calculate the intrinsic judging error variability in figure skating, which we then leverage to quantify the accuracy and biases of judges. This is an extension of the work initially started in gymnastics [3, 4, 6]. We emphasize once more that our analysis focuses on the Program Component Scores.

5.1. Intrinsic judging error variability

Using the data, we first estimate the intrinsic judging error variability $\hat{\sigma}(l_p)$, modeling the error made by an average judge as a function of the performance level l_p. Since the true performance level l_p of the component is unknown, we approximate it by the median panel mark, which is an excellent proxy over our large dataset. The results are shown in Figure 2. For instance, when evaluating a component whose true quality is 9.0, judges make an average error of 0.272. The Root Mean Square Error (RMSE) is very low, indicating that our weighted exponential regression is an excellent fit. We observe that judges are more accurate when evaluating the best performances, which is in line with other sports except dressage [3].

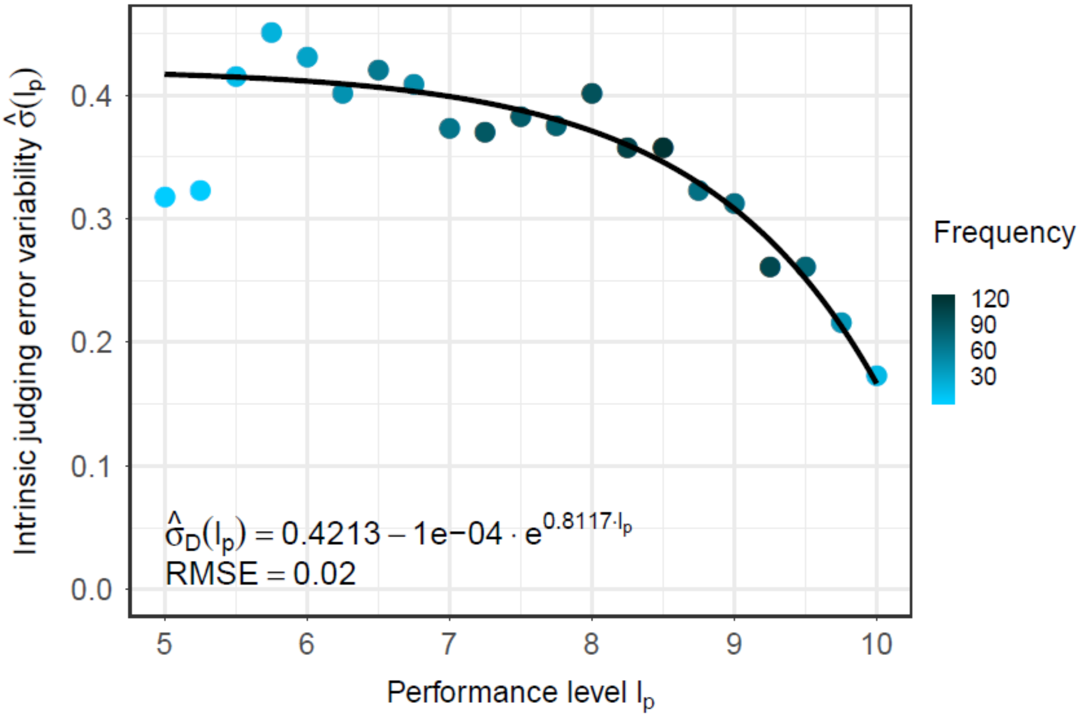


Figure 2 : Intrinsic judging error variability $\hat{\sigma}(l_p)$, as a function of the performance level l_p.

5.2. Marking score quantifying accuracy

We calculate, for each judge j and performance p, a marking score $m_{p,j}$ quantifying her/his accuracy as a function of the intrinsic judging error variability $\hat{\sigma}(l_p)$. It is given by

$$m_{p,j} \triangleq \frac{\widehat{e_{p,J}}}{\hat{\sigma}(l_p)} = \frac{s_{p,j} - l_p}{\hat{\sigma}(l_p)}$$

where $s_{p,j}$ is the mark of judge j for performance p and $\widehat{e_{p,J}} \triangleq s_{p,j} - l_p$ is the discrepancy (error) of judge j for performance p.

The overall marking score M_j of judge j is given by

$$M_j \triangleq \sqrt{\frac{1}{n}\sum_{p=1}^{n} m_{p,j}^2}.$$

The marking score expresses the accuracy of a judge compared to his peers as a multiple of the intrinsic judging error variability: a perfect judge whose scores are always equal to the true performance level has a marking score of 0, and an average judge with a judging error always equal to the intrinsic judging error variability has a marking score of 1.

Since the errors $\widehat{e_{p,J}}$ made by judges are random processes, we can complement our assessment of their accuracy with confidence intervals (CI). More precisely, the overall marking score is the sum of n squares of independent standard normal random variables $m_j \sim N(0,1)$ which follows a χ_n^2 distribution:

$$n \cdot M_j^2 = n \cdot \frac{1}{n}\sum_{p=1}^{n} m_{p,j}^2 \sim \chi_n^2$$

The boundaries of a confidence interval at the α level are

$$\mathrm{CI}_{M_j}(1-\alpha) = \left[\sqrt{\frac{n}{\chi_{1-\alpha/2,n}^2}}M_j ; \sqrt{\frac{n}{\chi_{\alpha/2,n}^2}}M_j\right].$$

As judges evaluate more athletes, their confidence intervals narrow, which decreases the probability that their good (or bad) marking score is due to random factors such as good (or bad) luck.

5.3. Nationalistic bias

We estimate the national bias with the regression

$$\widehat{e_{p,j}} \triangleq s_{p,j} - l_p = \beta_{\mathrm{SN}} \cdot \mathbb{1}_{\mathrm{SN}} \cdot \widehat{\sigma}(l_p) + \epsilon_{p,j}$$

where

- $\epsilon_{p,j} \sim \mathcal{N}\left(\widehat{\mu_j} \cdot \widehat{\sigma}(l_p), \widehat{\sigma^2}(l_p) \cdot M_j^2\right)$ is a normally distributed random error term with mean $\widehat{\mu_j} \cdot \widehat{\sigma}(l_p)$ and variance $\widehat{\sigma^2}(l_p) \cdot M_j^2$;
- $\mu_j \triangleq \frac{1}{n}\sum_{p=1}^{n} m_{p,j} = \frac{1}{n}\sum_{p=1}^{n}\frac{\widehat{e_{p,j}}}{\widehat{\sigma}(l_p)}$ is the general tendency expressing whether judge j is overall too generous or too severe;
- $\mathbb{1}_{SN}$ an indicator variable equal to 1 if and only if the judge and the athlete have the Same Nationality;
- β_{SN} is the amount of Same-Nationality bias estimated from the regression.

Our national bias β_{SN} is the propensity of judges to favor same-nationality athletes and has four characteristics. First, the bias is expressed as a multiple of the intrinsic judging error variability $\widehat{\sigma}(l_p)$, which makes sense intuitively because for the best athletes a small absolute bias can have a large effect on the final rankings. From Figure 2 an absolute bias of 0.3 for a performance level of 9.5 is much worse than for a performance of 7.5. Second, it leverages the marking score of each judge to differentiate between erratic but unbiased judges from precise but biased judges. Third, it takes into accounts the general tendency μ_j so that a generous judge is not perceived as being as biased, and vice-versa for a severe judge. Finally, it leverages the intercept of the linear regression, which allows to flag judges who are biased by increasing the marks of their own athletes while simultaneously penalizing their competitors.

5.4. Gender bias

Using the same method, we can estimate the gender bias β_{SG} by solving the regression

$$\widehat{e_{p,j}} \triangleq s_{p,j} - l_p = \beta_{\mathrm{SG}} \cdot \mathbb{1}_{\mathrm{SG}} \cdot \hat{\sigma}(l_p) + \epsilon_{p,j}$$

where $\epsilon_{p,j} \sim \mathcal{N}\left(\widehat{\mu_j} \cdot \hat{\sigma}(l_p), \widehat{\sigma^2}(l_p) \cdot M_j^2\right)$ and the indicator variable $\mathbb{1}_{\mathrm{SG}}$ takes the value 1 if and only if the judge and the athlete have the Same Gender.

6. Main Results

6.1. Marking score quantifying accuracy

Figure 3 shows the marking scores of the judges with 95% confidence intervals. We observe huge differences in judging accuracy: the average error of the best judges is two to three times smaller than the average error of the most erratic judges. The precision of Czech judge Richard Kosina and Slovakian judge Kvetoslava Matejova stand out, and so does the erraticness of Australian judge Elizabeth Ryan and, once again, Finnish judge Pekka Leskinen. The confidence intervals have similar widths because judges evaluated a similar number of performances.

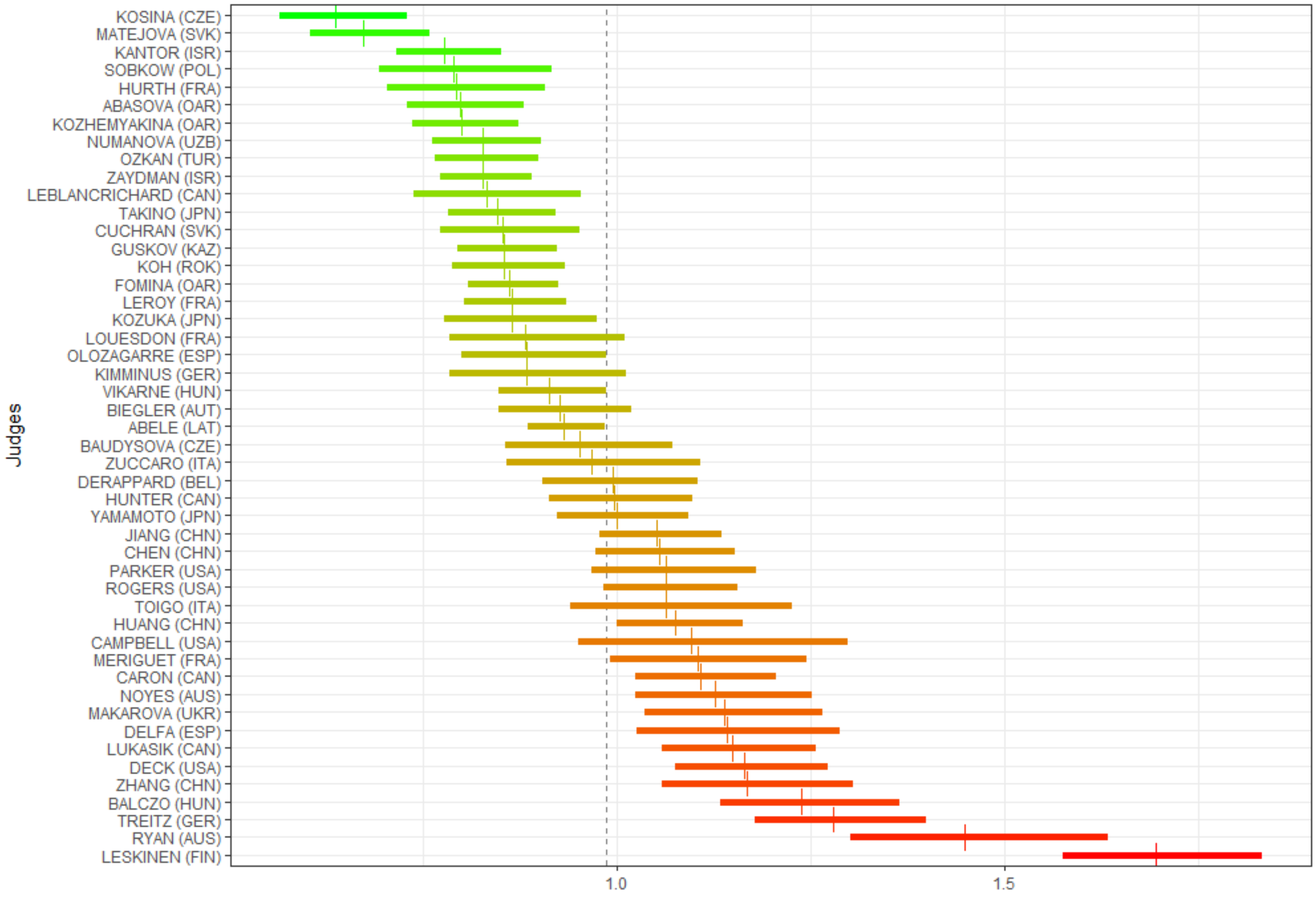
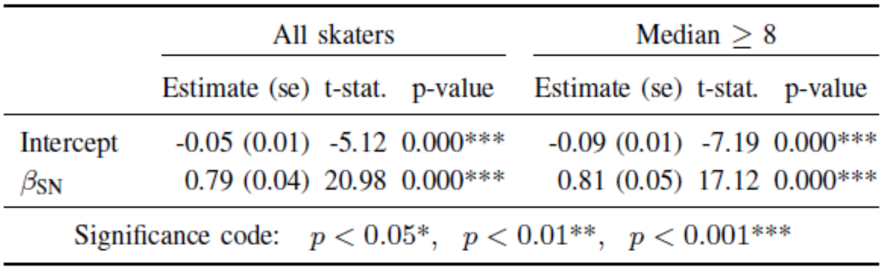


Figure 3 : Marking score of each judge with 95% confidence intervals.

6.2. Nationalistic bias

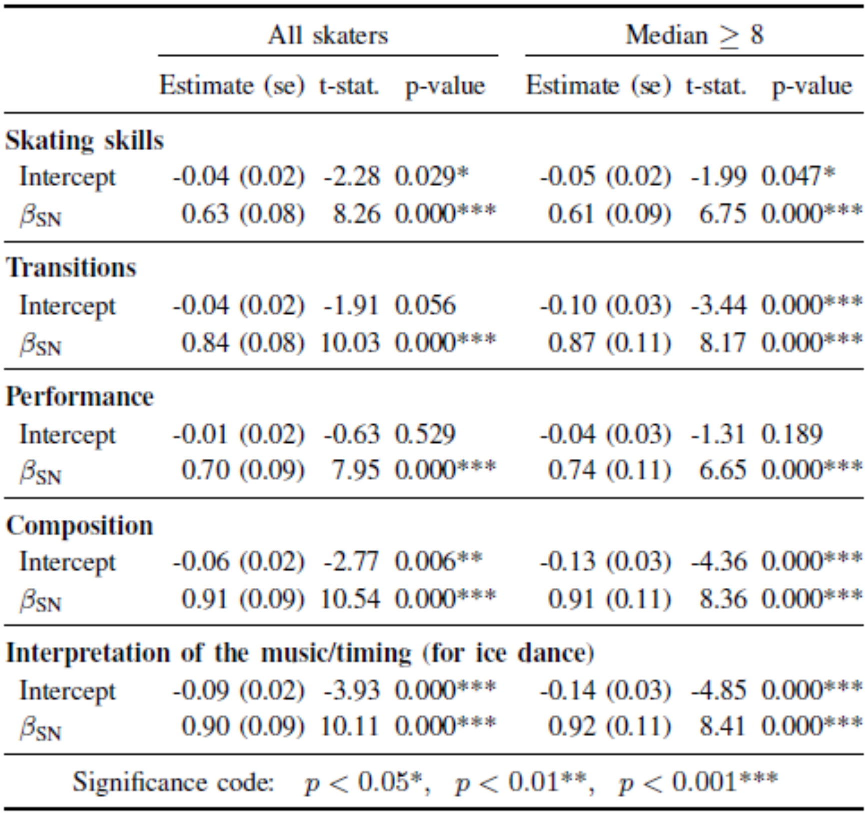
We express nationalistic bias as a multiple of the intrinsic judging error variability $\hat{\sigma}(l_p)$. Table II shows that the overall nationalistic bias is $\beta_{SN} = 0.79$, and statistically significant ($p \ll 0.001$). Overall, national bias is almost as important as all the sources of error of an average unbiased judge. This is two to three times higher than in artistic and rhythmic gymnastics, and obviously worse than trampoline judges who overall are unbiased [4]. Snowboard judges at the 2018 PyeongChang Olympic Winter Games were also unbiased. Table II also shows $\beta_{SN} = 0.81$ for the best performances with a median mark above 8.0. In other words, as the performance level improves, the absolute bias decreases, but remains constant or increases slightly when compared to the intrinsic judging error variability. This makes sense: for outstanding performances a large absolute bias is highly suspicious while a small absolute bias is sufficient to impact the rankings. Table III further shows that the national bias is important and statistically significant for all five components.



	All skaters			Median ≥ 8		
	Estimate (se)	t-stat.	p-value	Estimate (se)	t-stat.	p-value
Intercept	-0.05 (0.01)	-5.12	0.000***	-0.09 (0.01)	-7.19	0.000***
β_{SN}	0.79 (0.04)	20.98	0.000***	0.81 (0.05)	17.12	0.000***

Significance code: $p < 0.05$*, $p < 0.01$**, $p < 0.001$***

Table II. Overall nationalistic bias expressed as a multiple of the intrincic judging error variability $\hat{\sigma}(l_p)$.



	All skaters			Median ≥ 8		
	Estimate (se)	t-stat.	p-value	Estimate (se)	t-stat.	p-value
Skating skills						
Intercept	-0.04 (0.02)	-2.28	0.029*	-0.05 (0.02)	-1.99	0.047*
β_{SN}	0.63 (0.08)	8.26	0.000***	0.61 (0.09)	6.75	0.000***
Transitions						
Intercept	-0.04 (0.02)	-1.91	0.056	-0.10 (0.03)	-3.44	0.000***
β_{SN}	0.84 (0.08)	10.03	0.000***	0.87 (0.11)	8.17	0.000***
Performance						
Intercept	-0.01 (0.02)	-0.63	0.529	-0.04 (0.03)	-1.31	0.189
β_{SN}	0.70 (0.09)	7.95	0.000***	0.74 (0.11)	6.65	0.000***
Composition						
Intercept	-0.06 (0.02)	-2.77	0.006**	-0.13 (0.03)	-4.36	0.000***
β_{SN}	0.91 (0.09)	10.54	0.000***	0.91 (0.11)	8.36	0.000***
Interpretation of the music/timing (for ice dance)						
Intercept	-0.09 (0.02)	-3.93	0.000***	-0.14 (0.03)	-4.85	0.000***
β_{SN}	0.90 (0.09)	10.11	0.000***	0.92 (0.11)	8.41	0.000***
Significance code: $p < 0.05$*, $p < 0.01$**, $p < 0.001$***						

Table III. Nationalistic bias per component expressed as a multiple of the intrinsic judging error variability $\hat{\sigma}(l_p)$.

Figure 4 shows the national bias per judge with 95% confidence intervals. Confidence intervals narrow as the number of same-nationality observations increases. Figure 5 also shows the national bias per judge, this time against the number of same-nationality observations. Statistically significant biases ($p \ll 0.001$) are in dark blue. Table IV provides more details about the most biased judges.

Our main observation is that the most biased judges have $\beta_{\mathrm{NB}} \approx 2 \cdot \hat{\sigma}(l_p)$, thus their national bias is twice higher than all other sources of error of an average unbiased judge! We also observe that the Chinese judges suspended by the ISU, Chen Weiguang and Huang Feng, do not stand out among the numerous judges with a statistically significant bias. An interesting judge is Elena Fomina from Russia, who is mostly biased for the best athletes, and does so less by giving better marks to her own athletes than by giving lower marks to everybody else (Table IV).

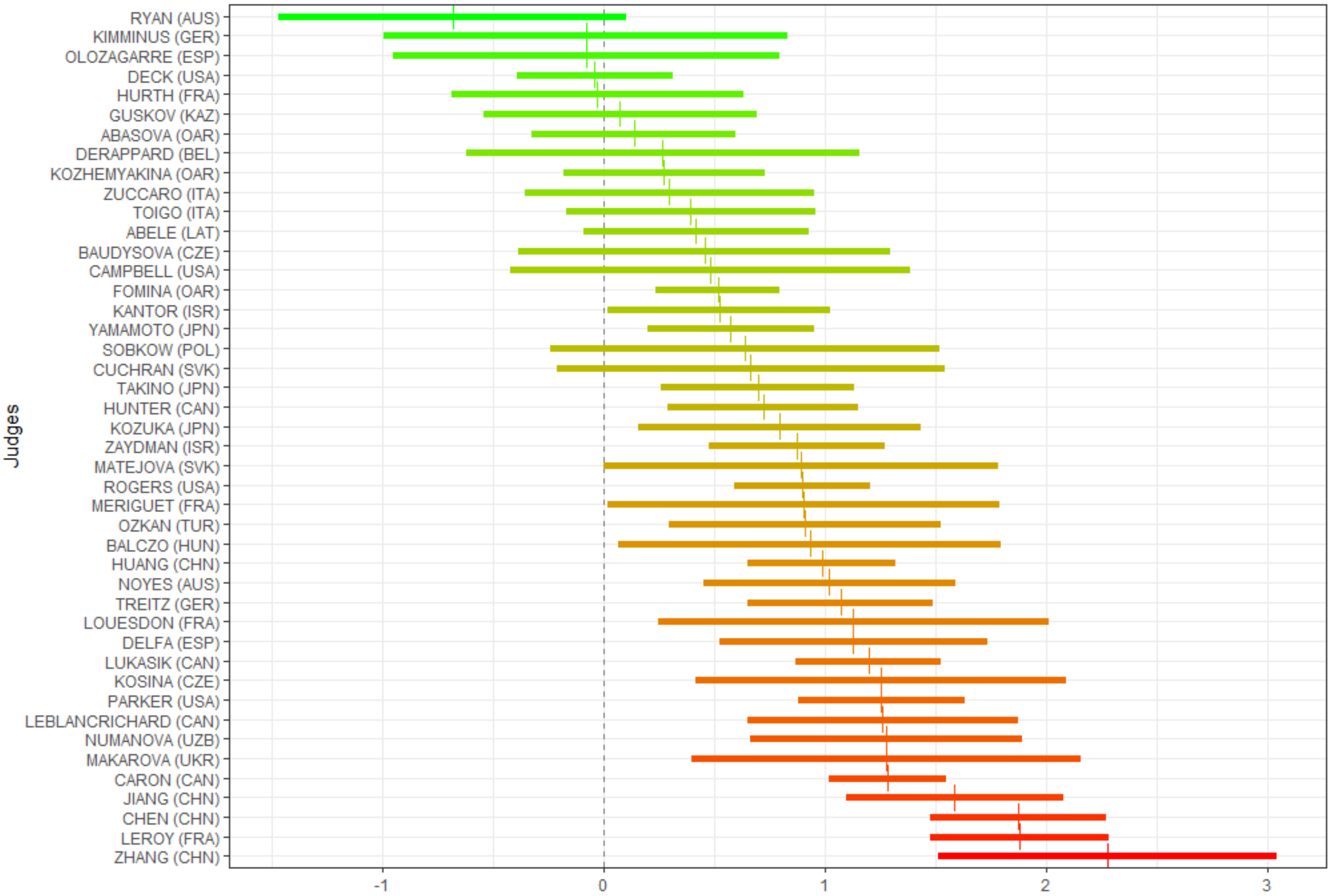


Figure 4 : Nationalistic bias as a multiple of the intrinsic judging error variability $\hat{\sigma}(l_p)$ with 95% confidence interval.

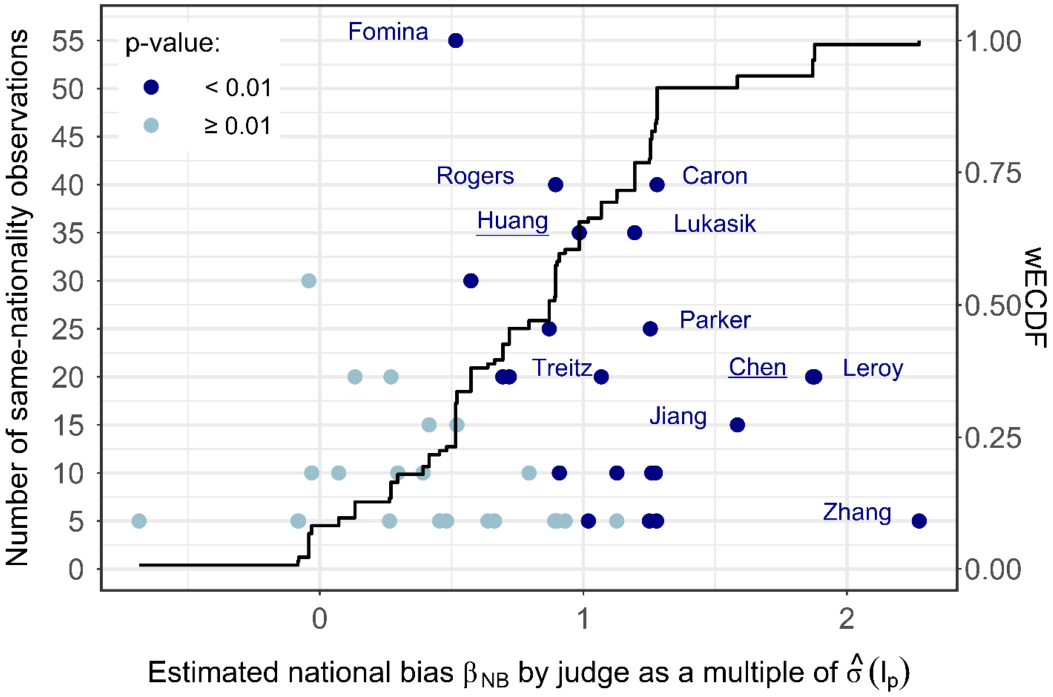
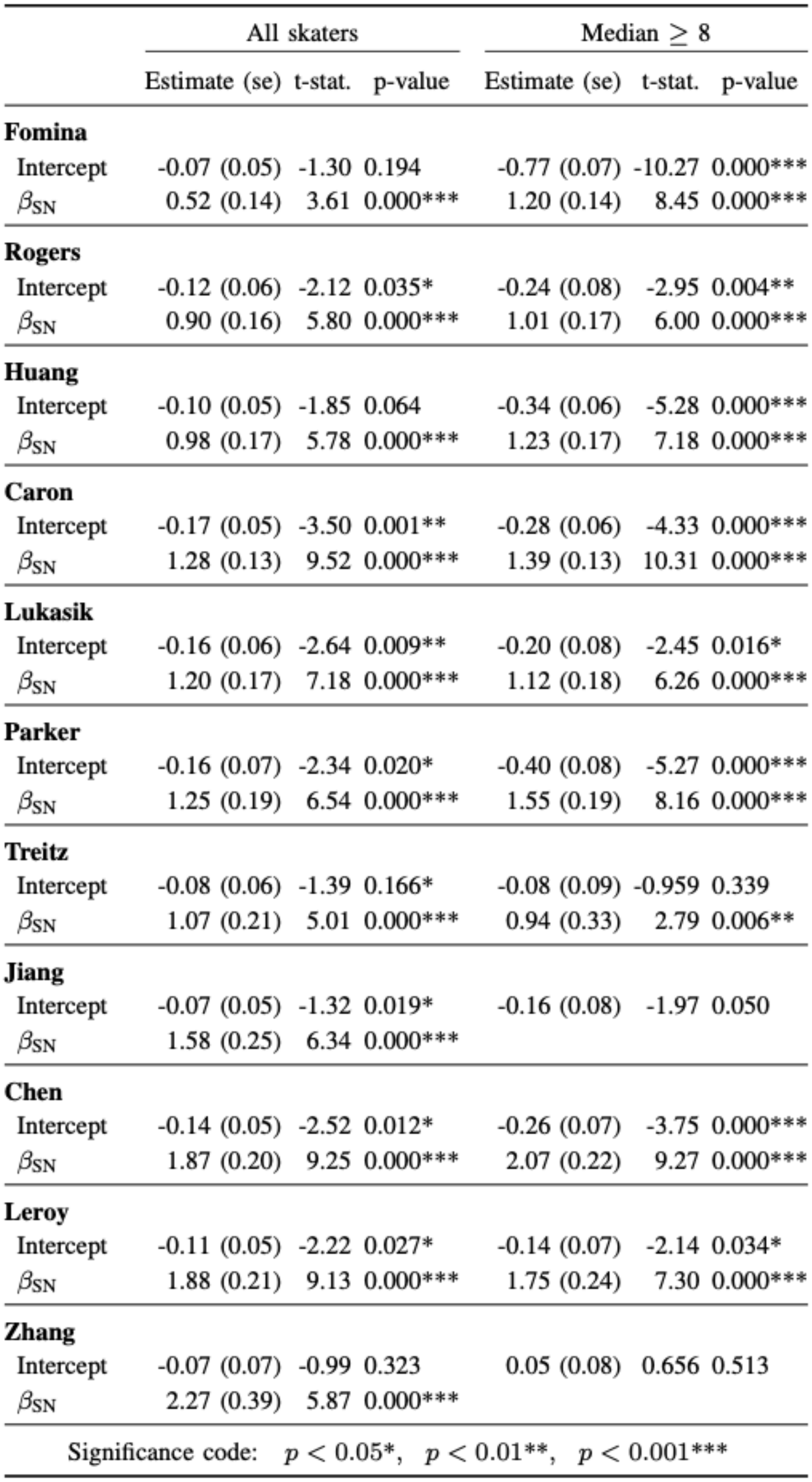


Figure 5 : National bias per judge, with the weighted empirical cumulative distribution function (wECDF) of the estimations. The national bias is expressed as a multiple of the intrinsic judging error variability $\hat{\sigma}(l_p)$.

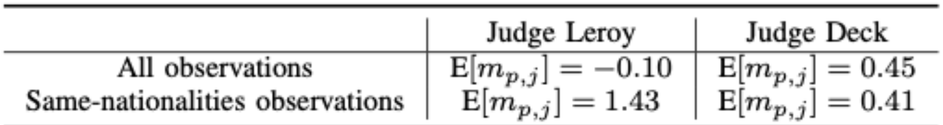


	All skaters			Median $\geq$ 8		
	Estimate (se)	t-stat.	p-value	Estimate (se)	t-stat.	p-value
Fomina						
Intercept	-0.07 (0.05)	-1.30	0.194	-0.77 (0.07)	-10.27	0.000***
β_{SN}	0.52 (0.14)	3.61	0.000***	1.20 (0.14)	8.45	0.000***
Rogers						
Intercept	-0.12 (0.06)	-2.12	0.035*	-0.24 (0.08)	-2.95	0.004**
β_{SN}	0.90 (0.16)	5.80	0.000***	1.01 (0.17)	6.00	0.000***
Huang						
Intercept	-0.10 (0.05)	-1.85	0.064	-0.34 (0.06)	-5.28	0.000***
β_{SN}	0.98 (0.17)	5.78	0.000***	1.23 (0.17)	7.18	0.000***
Caron						
Intercept	-0.17 (0.05)	-3.50	0.001**	-0.28 (0.06)	-4.33	0.000***
β_{SN}	1.28 (0.13)	9.52	0.000***	1.39 (0.13)	10.31	0.000***
Lukasik						
Intercept	-0.16 (0.06)	-2.64	0.009**	-0.20 (0.08)	-2.45	0.016*
β_{SN}	1.20 (0.17)	7.18	0.000***	1.12 (0.18)	6.26	0.000***
Parker						
Intercept	-0.16 (0.07)	-2.34	0.020*	-0.40 (0.08)	-5.27	0.000***
β_{SN}	1.25 (0.19)	6.54	0.000***	1.55 (0.19)	8.16	0.000***
Treitz						
Intercept	-0.08 (0.06)	-1.39	0.166*	-0.08 (0.09)	-0.959	0.339
β_{SN}	1.07 (0.21)	5.01	0.000***	0.94 (0.33)	2.79	0.006**
Jiang						
Intercept	-0.07 (0.05)	-1.32	0.019*	-0.16 (0.08)	-1.97	0.050
β_{SN}	1.58 (0.25)	6.34	0.000***			
Chen						
Intercept	-0.14 (0.05)	-2.52	0.012*	-0.26 (0.07)	-3.75	0.000***
β_{SN}	1.87 (0.20)	9.25	0.000***	2.07 (0.22)	9.27	0.000***
Leroy						
Intercept	-0.11 (0.05)	-2.22	0.027*	-0.14 (0.07)	-2.14	0.034*
β_{SN}	1.88 (0.21)	9.13	0.000***	1.75 (0.24)	7.30	0.000***
Zhang						
Intercept	-0.07 (0.07)	-0.99	0.323	0.05 (0.08)	0.656	0.513
β_{SN}	2.27 (0.39)	5.87	0.000***			

Significance code: $p < 0.05$*, $p < 0.01$**, $p < 0.001$***

Table IV. Bias of most biased judges expressed as a multiple of the intrinsic judging error variability $\hat{\sigma}(l_p)$.

We can also combine the marking scores in Figure 3 and the national bias Figure 4. For instance, French judge Anthony Leroy is in general rather accurate ($M_j \approx 0.86$) but highly biased in favor of French skaters ($\beta_{SN} \approx 1.62$), whereas American judge Deveny Deck is erratic ($M_j \approx 1.17$) but unbiased ($\beta_{SN} \approx -0.05$). Table V provides additional details about both judges, showing that Deck is overall too lenient, but in equal fashion for his own athletes and the other skaters.

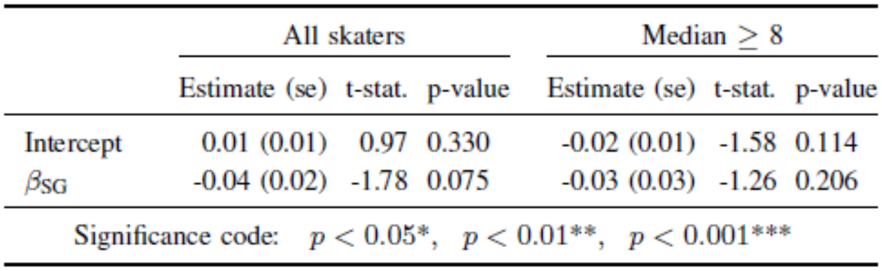


	Judge Leroy	Judge Deck
All observations	$\mathrm{E}[m_{p,j}] = -0.10$	$\mathrm{E}[m_{p,j}] = 0.45$
Same-nationalities observations	$\mathrm{E}[m_{p,j}] = 1.43$	$\mathrm{E}[m_{p,j}] = 0.41$

Table V. Comparison of judges Anthony Leroy and Deveny Deck.

6.3. Gender studies

We did not observe any gender bias, as shown in Table VI. Men (and women) judges do not judge men and women skaters differently. This is similar to what Sandberg concluded in dressage [7]. The amount of national bias is also comparable for both men and women judges. However, Figure 6 shows that women judges are, in the aggregate, 7% more accurate than men judges. This is similar to what we observed in artistic gymnastics and trampoline [6], and we conjecture that this is due to the larger pool of female figure skaters[10], leading to more women who want and are eventually qualified to become judges.



	All skaters			Median ≥ 8		
	Estimate (se)	t-stat.	p-value	Estimate (se)	t-stat.	p-value
Intercept	0.01 (0.01)	0.97	0.330	-0.02 (0.01)	-1.58	0.114
β_{SG}	-0.04 (0.02)	-1.78	0.075	-0.03 (0.03)	-1.26	0.206
Significance code: $p < 0.05^*$, $p < 0.01^{**}$, $p < 0.001^{***}$						

Table VI. Gender bias expressed as a multiple of the intrinsic judging error variability $\hat{\sigma}(l_p)$.

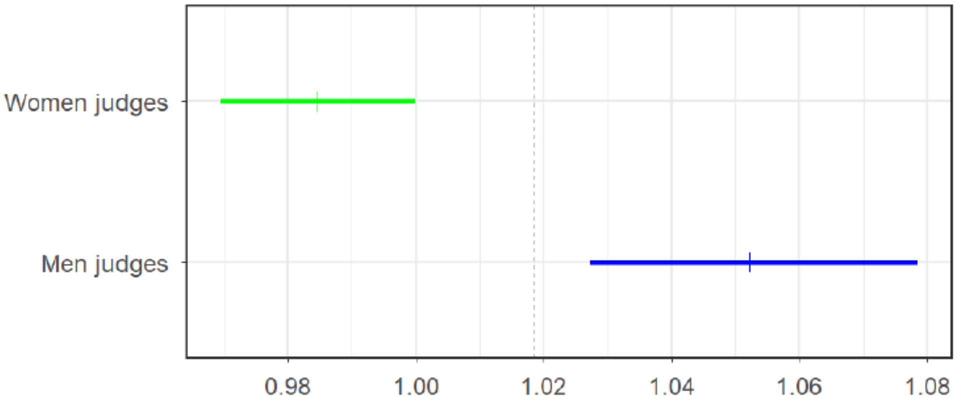


Figure 6 : Overall marking score for men and women judges with 95% intervals.

[10] For instance more than 70% of U.S. figure skating members are women (https://www.usfsa.org/content/FactSheet.pdf).

6.4. Effect of the starting number, or lack thereof

Many "laboratory" experiments reported a conformity bias in figure skating [5] and other sports such as gymnastics [2] and artistic swimming [1]. In these studies, open feedback causes judges to adapt their marks to those of the other judges of the panel.

To test whether this applies to our dataset, we first define the adjusted error of judge j for performance p as

$$d_{p,j} \triangleq \frac{s_{p,j} - l_p - \widehat{\mu}_j \cdot \widehat{\sigma}(l_p)}{\widehat{\sigma}(l_p) \cdot M_j}.$$

We then estimate the root mean square of the adjusted judging error as a function of the starting number # as

$$\mathrm{RMS}_{\mathrm{order}}(\#) = \sqrt{\sum_{\substack{\text{observations } (p,j) \\ \text{of starting order } \#}} d_{p,j}^2}$$

$\mathrm{RMS}_{\mathrm{order}}$ is expressed as a multiple of the intrinsic judging error variability, and thus adjusted for the performance quality. It is also adjusted for the general tendency and accuracy of the judges. The result appears in Figure 7 and indicate that the variability of the adjusted judging errors is not dependent on the starting number. More plainly: conformity bias, if it exists, is dwarfed by the intrinsic error variability of the judges and their national bias.

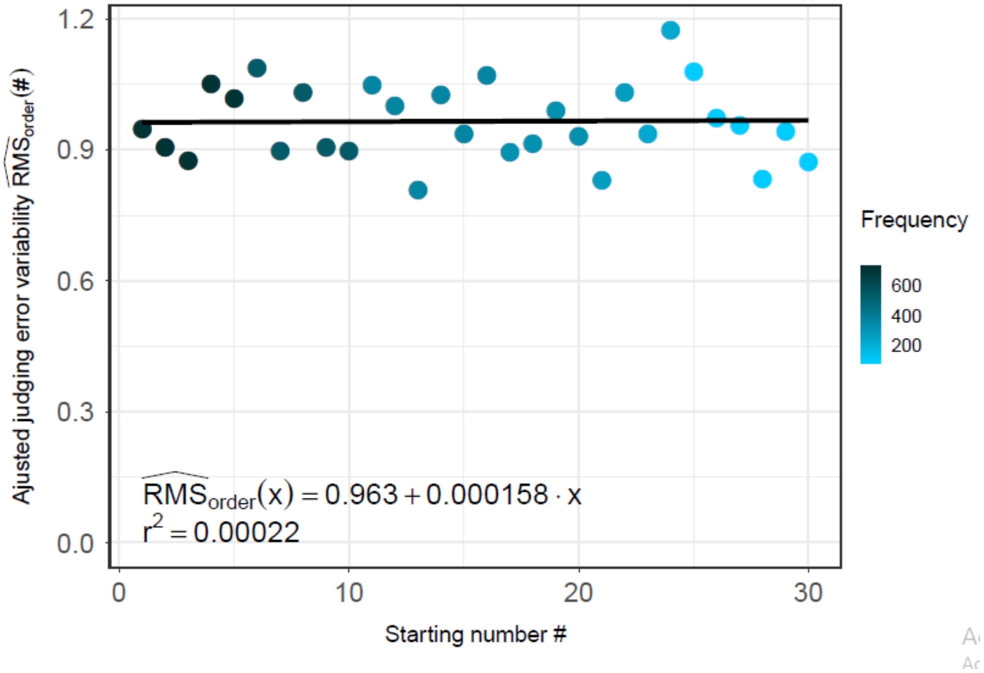


Figure 7 : Adjusted judging error variability as a function of the starting number.

7. Limitations of our approach

The main weakness of our approach is the assumption that the median of all the nine panel marks is a good estimation of the real value of the athlete's performance. This has two consequences. First, it does not take into account compensation effects. If other-nationality judges counter same-nationality evaluations by systematically awarding lower marks themselves, our model will overestimate national bias. Zitzewitz [9] showed that the opposite was true in figure skating: having a same-nationality judge on the panel slightly increased the marks given by other judges. The most probable cause is vote trading, the other judges hoping to get something else in return. If this is still true, our analysis slightly underestimates national bias. Second, and more importantly, a judge whose mark deviates from the others for a single performance might be correct but out of consensus with other incorrect or colluding judges, or might have made an honest judging error. As mentioned in prior work [4], a single outlier must be viewed with circumspection. However, as the number of observations increases, so does our confidence that high marking scores and large national biases are representative of the poor performance of judges and not due to bad luck.

Another limitation of our approach is that we estimate national bias directly and do not consider more complex mechanisms such as vote trading. Consider the big misjudging from the Finnish judge Pekka Leskinen "against" the Israeli skater Alexei Bychenko. This evaluation contributes to the very bad marking score of judge Leskinen, but without further information we must assume that it is due to a misguided personal dislike of the performance. Suspected of being at the root of the 2002 Winter Olympics figure skating scandal, vote trading is hard to study objectively, and is far from the most important judging issue facing the ISU.

8. Recommendations moving forward

In this article we studied the accuracy and national bias of figure skating judges at the 2018 PyeongChang Olympic Winter Games using novel and sound statistical tools. Our analysis reveals systemic national bias and large differences in accuracy among judges, mostly caused by the inadequate monitoring currently done by the ISU.

Despite these discouraging results, there are many positive developments. The new judging system with a precise code of points is a significant improvement over the old 6.0 system based on ordinal rankings, because it provides precise and objective judging criteria. Even the program component scores, while more subjective, are a significant improvement over the old presentation mark. The marks given by judges are public once again, which facilitates monitoring by third parties. The aggregation mechanisms removing the best and worst marks for each element are also sound. We are convinced that most judging problems and controversies in figure skating could be solved by properly monitoring judges longitudinally using sound statistical tools.

Nationalistic bias is higher in figure skating than in other similar sports simply because judges can get away with it. Proper long-term monitoring can get rid of it. As mentioned in prior work [4], our analysis cannot infer intent. This being said intent does not matter: good judges should be unbiased, and biased judges should not be allowed to officiate, no matter whether this bias is conscious or not.

Judging accuracy is harder to tackle for the simple reason that evaluating a figure skating program is very hard. Even among the best trained judges at the international level, some judges are simply better than others. With long-term monitoring, we can identify and reward the best judges and provide constructive feedback to less accurate judges so that they can improve. Phasing out inaccurate judges has an additional benefit: the more accurate a judge is over the long run, the harder it is for this judge to cheat or exhibit a large bias without being caught.

References

[1] Yves Vanden Auweele et al. “Judging Bias in Synchronized Swimming: Open Feedback Leads to Nonperformance-Based Conformity”. In: Journal of Sport and Exercise Psychology 26.4 (2004), pp. 561–571.

[2] Filip Boen et al. “Open feedback in gymnastic judging causes conformity bias based on informational influencing”. In: Journal of Sports Sciences 26.6 (2008), pp. 621–628.

[3] Sandro Heiniger and Hugues Mercier. “Judging the Judges: A General Framework for Evaluating the Performance of International Sports Judges”. In: ArXiv eprints (Aug. 2019). arXiv: 1807.10055 [stat.AP]. URL: https://arxiv.org/abs/1807.10055.

[4] Sandro Heiniger and Hugues Mercier. “National Bias of International Gymnastics Judges during the 2013-2016 Olympic Cycle”. In: ArXiv e-prints (Aug. 2019). arXiv: 1807.10033 [stat.AP]. URL: https://arxiv.org/abs/1807.10033.

[5] Jungmin Lee. “Outlier Aversion in Subjective Evaluation: Evidence From World Figure Skating Championships”. In: Journal of Sports Economics 9.2 (2008), pp. 141–159.

[6] Hugues Mercier and Sandro Heiniger. “Judging the Judges: Evaluating the Performance of International Gymnastics Judges”. In: ArXiv e-prints (2019). arXiv: 1807.10021 [stat.AP]. URL: https://arxiv.org/abs/1807.10021.

[7] Anna Sandberg. “Competing Identities: A Field Study of In-group Bias Among Professional Evaluators”. In: The Economic Journal 128.613 (2018), pp. 2131–2159.

[8] Eric Zitzewitz. “Does transparency reduce favoritism and corruption? Evidence from the reform of figure skating judging”. In: Journal of Sports Economics 15.1 (2014), pp. 3–30.

[9] Eric Zitzewitz. “Nationalism in winter sports judging and its lessons for organizational decision making”. In: Journal of Economics & Management Strategy 15.1 (2006), pp. 67–99.