Research article

Sandro Heiniger and Hugues Mercier*

Judging the judges: evaluating the accuracy and national bias of international gymnastics judges

<https://doi.org/10.1515/jqas-2019-0113>

Received October 30, 2019; accepted June 11, 2021;

published online October 19, 2021

Abstract: We design, describe and implement a statistical engine to analyze the performance of gymnastics judges with three objectives: (1) provide constructive feedback to judges, executive committees and national federations; (2) assign the best judges to the most important competitions; (3) detect bias and persistent misjudging. Judging a gymnastics routine is a random process, and we model this process using heteroscedastic random variables. The developed marking score scales the difference between the mark of a judge and the true performance level of a gymnast as a function of the intrinsic judging error variability estimated from historical data for each apparatus. This dependence between judging variability and performance quality has never been properly studied. We leverage the intrinsic judging error variability and the marking score to detect outlier marks and study the national bias of judges favoring athletes of the same nationality. We also study ranking scores assessing to what extent judges rate gymnasts in the correct order. Our main observation is that there are significant differences between the best and worst judges, both in terms of accuracy and national bias. The insights from this work have led to recommendations and rule changes at the Fédération Internationale de Gymnastique.

Keywords: gymnastics; intrinsic judging error variability; national bias; outlier detection; quantifying accuracy; sports judges.

1 Introduction

Judging a gymnastics routine is a challenging task. Judges must assess the performance of athletes live, without comprehensive technical assistance, surrounded by thousands of cheering spectators, and according to hundreds of instructions specified in scoring regulations. These evaluations anoint world champions or Olympic medalists and all the involved parties – athletes, coaches, fans, officials, sponsors – have a vested interest in having accurate and fair judges.

The two components of judging are fairness and accuracy. The first component, fairness, relates to impartiality and lack of favoritism. Gymnastics judges and judges from similar sports are susceptible to well-studied biases.¹ The most prevalent bias in sports is national bias, which comes in two flavors: judges can favor athletes of the same nationality, and at the same time penalize their competitors. National bias was shown to exist in artistic gymnastics (Ansorge and Scheer 1988; Leskošek et al. 2012) and rhythmic gymnastics (Popović 2000), and in many other sports including figure skating (Campbell and Galbraith 1996; Zitzewitz 2006, 2014), ski jumping (Zitzewitz 2006), Muay Thai boxing (Myers et al. 2006), diving (Emerson, Seltzer, and Lin 2009) and dressage (Sandberg 2018). National bias seems to increase for the most important events with a strong national dimension (Sandberg 2018). The analytical approaches to identify national bias above are manifold: some analyses use sign tests or permutation tests, whereas other models use linear regressions.

Other smaller biases have been observed in gymnastics. Plessner (1999) observed a *serial position bias* in gymnastics: a competitor performing and evaluated last gets better marks than when performing first. An experiment by Boen et al. (2008) found a *conformity bias* in gymnastics: open feedback causes judges to adapt their marks to those of the other judges of the panel. This bias cannot appear during competitions since individual marks are not disclosed. Damisch, Mussweiler, and Plessner (2006) found a

This work was started while S. Heiniger and H. Mercier were at the Université de Neuchâtel, Switzerland.

*Corresponding author: Hugues Mercier, University of Applied Sciences and Arts of Western Switzerland, Neuchâtel, Switzerland, E-mail: hm@huguesmercier.com. <https://orcid.org/0000-0001-9348-4515>

Sandro Heiniger, Swiss Institute for Empirical Economic Research, University of St.Gallen, St.Gallen, Switzerland, E-mail: sandro.heiniger@unisg.ch. <https://orcid.org/0000-0003-4769-8402>

¹ Consult Landers (1970) for an initial comprehensive survey until 1970, and Bar-Eli, Plessner, and Raab (2011) for a recent survey.

sequential bias in artistic gymnastics at the 2004 Olympic Games: the evaluation of a gymnast is likely more generous than expected if the preceding gymnast performed well. Plessner and Schallies (2005) showed in an experiment that still rings judges can make systematic errors based on their viewpoint. Biases observed in other sports might also occur in gymnastics as well. Findlay and Ste-Marie (2004) found a *reputation bias* in figure skating: judges overestimate the performance of athletes with a good reputation. Price and Wolfers (2010) quantified the *racial bias* of NBA officials against players of the opposite race, which was large enough to affect the outcome of basketball games. Interestingly, the racial bias of NBA officials subsequently disappeared, most probably due to the public awareness of the bias from the first study (Pope, Price, and Wolfers 2018). The aforementioned biases are often unconscious and cannot always be entirely eliminated in practice. However, rule changes and monitoring from the Fédération Internationale de Gymnastique (FIG) as well as increased scrutiny induced by the media exposure of major gymnastics competitions make most of these biases reasonably small and/or tempered by mark aggregation.

The second of component of judging is accuracy: it is difficult to evaluate every single aspect of the complex movements that are part of a gymnastics routine. This leads to an inevitable element of subjectivity and randomness in the marks given by each judge. This challenge has been known since at least the 1930s (Zwarg 1935), and there is a large number of studies on the ability of judges to detect execution mistakes in gymnastic routines (Flessas et al. 2015; Pizzera 2012; Pizzera, Möller, and Plessner 2018; Ste-Marie 1999, 2000).² In a nutshell, novice judges consult their scoring sheet much more often than experienced international judges, thus missing execution errors. Furthermore, international judges have superior perceptual anticipation, are more capable of detecting errors in their peripheral vision and, when they are former gymnasts, use sophisticated cognitive judging strategies and leverage their own sensorimotor experiences. Unsurprisingly, nearly all international judges are former gymnasts.

Even among well-trained judges at the international level, differences can still be substantial: some judges are simply better than others. For this reason, the FIG has developed and used the Judge Evaluation Program (JEP) to assess the performance of judges during and after international competitions. The work on JEP was started in 2006 and the tool has grown iteratively since then. Despite its

usefulness, JEP was partly designed with unsound and inaccurate mathematical tools, and was not always evaluating what it ought to evaluate. Surprisingly, there is little scientific work studying how to monitor referees such as sports judges. This is the main objective of this work.

We describe a toolbox to assess, as objectively as possible, the accuracy of international gymnastics judges using a simple yet rigorous methodology. Part of this toolbox is now the core statistical engine of the newest iteration of JEP³ providing feedback to judges, executive committees and national federations. It is used to reward the best judges by selecting them to the most important competitions such as the Olympic Games. The main tool we develop is a *marking score* evaluating the accuracy of the marks given by a judge. We design the marking score so that it is independent of the apparatus/discipline under evaluation and the skill level of the gymnasts. To achieve this, we model the behavior of judges as heteroscedastic random variables using historical data from international and continental gymnastics competitions held during the 2013–2016 Olympic cycle. The standard deviation of these random variables, describing the *intrinsic judging error variability* of each apparatus or discipline, decreases as the performance of the gymnasts improves: judges are more precise judging the best athletes than mediocre ones. This dependence between judging variability and performance quality has never been properly studied in any setting (sport or other).⁴ The marking score for each judge quantifies her/his accuracy as a multiple of the intrinsic judging error variability. We observe significant differences among judges: the average error of the worst judges is typically three times larger than the average error of the best judges. More plainly: the best judges are three times more accurate than the worst judges.

Leveraging the marking score, we show that reference judges, who are hand-picked by the FIG and imparted with more power than regular panel judges, are not better than regular panel judges in the aggregate. We also show that women judges are significantly more accurate than men

² Consult Landers (1970) for an initial comprehensive survey until 1970, and Bar-Eli, Plessner, and Raab (2011) for a recent survey.

³ The current iteration of JEP was developed in collaboration with the FIG and the Longines watchmaker and deployed in 2017. It is a full software stack that handles all the interactions between the databases, our statistical engine, and a user-friendly front-end to generate statistics and judging reports.

⁴ In (Heiniger and Mercier 2019a) we show that the intrinsic judging error variability has the same characteristic heteroscedastic shape in other sports using judging panels and marks within a finite range. Hence, the integration of the marking score could lead to improved monitoring of the accuracy and national bias of judges in all these sports.

judges in artistic gymnastics and in trampoline. The marking score further enables us to detect misjudgements by evaluating outliers with judge-specific thresholds.

We integrate the marking score and the distribution of the intrinsic judging error variability into a national bias model. Contrary to previous regression-based analyses (e.g. Leskošek et al. 2012; Sandberg 2018), this allows us to quantify the national bias not only on a nominal level but also in terms of the intrinsic variability of judging errors. This is essential for evaluating the severity of the bias targeting the best athletes competing for medals. While acrobatic and trampoline judges do not exhibit any national bias in favor of their own athletes in the aggregate, we reveal significant bias in aerobic, artistic and rhythmic gymnastics. At the individual level, the national bias of some judges is two or even three times larger than the intrinsic error variability of an average judge. More plainly once again: the national bias of these judges is two to three times larger than all the sources of errors of an average judge. Even though we cannot infer intent, the magnitude and statistical significance of this bias is too large for the FIG to leave it unaddressed. Fortunately, in only one all-around finals has this national bias plausibly modified the podium in an artistic gymnastics competition. The fact that this did occur so infrequently is a testament to the efforts made by the FIG to avoid same-nationality judges in the finals whenever possible, and to the aggregation mechanisms excluding the worst and best marks from the judging panels.

The FIG ultimately wants to quantify to what extent judges rank gymnasts in the correct order. We therefore compare the marking score to different metrics of distances between rankings such as the generalized version of Kendall's τ distance (Kumar and Vassilvitskii 2010). Depending on how these ranking scores are parametrized, they are either strongly correlated with our marking score and thus redundant, or heavily weighted by a small number of evaluations, which does not provide a comprehensive assessment of judging capabilities. Since no approach was satisfactory, the FIG no longer uses ranks to monitor its judges.⁵

All these observations led to recommendations and changes at the FIG during the course of this work. We proposed concrete measures to improve the training, evaluation, monitoring and accuracy of judges, and to further decrease the impact of national bias, especially in all-around finals in artistic and rhythmic gymnastics where it is difficult to avoid same-nationality evaluations.

The remainder of this article is organized as follows. We present the gymnastics judging system in Section 2 and describe our dataset in Section 3. We discuss the true performance quality, control scores, and the intrinsic judging error variability in Section 4. We derive the marking score in Section 5. We then leverage the intrinsic judging error variability and the marking score to detect outliers in Section 6, and to analyze national bias in Section 7. We compare the marking score to evaluations based on ranks in Section 8 and conclude in Section 9 by presenting general recommendations applicable to all sports with judging systems similar to the one used in gymnastics.

2 Judging in gymnastics

The six main gymnastics disciplines recognized by the Fédération Internationale de Gymnastique (FIG) are artistic gymnastics, rhythmic gymnastics and trampoline, which are Olympic sports and have a world championship every non-Olympic year, aerobic gymnastics and acrobatic gymnastics, which have a world championship held every two years, and the new parkour discipline, which will hold its first world championship in 2022. Gymnastics disciplines have different apparatus and competition formats. Acrobatic gymnastics routines are performed in pairs (men, women and mixed) or in groups (men and women). Aerobic gymnastics features individual routines, mixed pairs and group routines (regardless of gender). Artistic gymnastics is split by gender: men compete on six apparatus (floor exercise, parallel bars, horizontal bar, pommel horse, still rings and vault) and women compete on four (balance beam, floor exercise, uneven bars and vault). Rhythmic gymnastics is only practiced by women. It includes individual routines with one apparatus (ball, club, hoop or ribbon) and group routines with one or two apparatus. Trampoline is split by gender, but men and women compete in the same type of events: individual and synchronized trampoline, double mini-trampoline, and tumbling.

Gymnastics competitions typically consist of a qualifying round followed by a final regrouping of the best qualifiers. A gymnastic routine at the international level is evaluated on the difficulty, artistry and execution components of the performance by panels of judges. In this article we focus on execution and artistry judges. After the completion of a routine by the gymnast, each execution and artistry judge in the panel grades the performance with a mark from 0 to 10 at steps of 0.1. The evaluation of the execution of a gymnastics routine is based on deductions

⁵ The previous iteration of JEP used a rudimentary ranking score.

Table 1: Standard composition of the execution and artistry panels per discipline during the 2013–2016 Olympic cycle.

Discipline	Typical execution panel composition	Typical artistry panel composition
Acrobatic gymnastics	4E + 2R	4E + 2R
Aerobic gymnastics	4E + 2R	4E + 2R
Artistic gymnastics	5E + 2R	
Rhythmic gymnastics	5E + 2R	
Trampoline	5E	

E = execution judges; R = reference judges.

Table 2: Sample size by discipline.

Discipline	Number of performances	Nb. of marks	Marks above 7	Same-nationality marks	Direct competitors
Acrobatic	756	5204	4898	257 (5.3%)	843 (17.3%)
Aerobic	938	6516	6396	200 (3.1%)	757 (11.8%)
Artistic (M)	7652	50,286	46,748	909 (1.9%)	3006 (6.4%)
Artistic (F)	4288	28,410	23,515	522 (2.2%)	1694 (7.2%)
Rhythmic	2841	19,052	17,673	405 (2.3%)	1297 (7.3%)
Trampoline	1986	8550	7278	343 (4.7%)	833 (11.4%)

precisely defined in the Code of Points for each apparatus.⁶ The marks given by the judges are aggregated to generate the final scores and rankings of the gymnasts. The number of judges per panel and the aggregation procedure vary per discipline and competition level.

With the exception of trampoline, the execution and artistry panels can include two additional reference judges who evaluate performances based on the same criteria as the other panel judges, but with increased decision weight, in case their scores strongly diverge from the panel. The aggregation process in artistic and rhythmic gymnastics⁷ works as follows: the execution panel score is the trimmed mean of the middle three of out five execution panel judges, and the reference score is the arithmetic mean of the two reference judges' marks. If the gap between the execution panel score and the reference score exceeds a predefined tolerance threshold, and if the difference between the marks of both reference judges is below a second threshold, then the final execution score of the gymnast is the mean of the execution panel and reference scores, otherwise the execution panel score remains unchanged. This

makes reference judges more powerful than regular panel judges, which exacerbates their mistakes.

Table 1 summarizes the composition of the typical execution and artistry panel for each discipline during the 2013–2016 Olympic cycle.⁸

3 Data

The data, provided by the FIG⁹ and Longines,¹⁰ encompasses 21 international and continental competitions held between 2013 and 2016, including the 2016 Rio Olympic Games. The data is not publicly available because the FIG considers that the individual judging marks are sensitive information. Table 2 shows the size of the dataset by discipline after the subsequent adjustments. The number of marks depends on the number of performances in the dataset and the size of the judging panels. The number of performances in an event is not always equal to the number of gymnasts. For instance, gymnasts who wish

⁶ The Codes of Points, their appendices and other documents related to rules for all the gymnastics disciplines are publicly available at <https://www.gymnastics.sport/site/rules/rules.php>. Competitions in our dataset were ruled by the 2013–2016 Codes of Points.

⁷ Acrobatic and aerobic gymnastics have a similar process for execution and artistry judges.

⁸ The panels do not always follow this typical composition: i.e. the qualifying phases in artistic and rhythmic gymnastics may include four execution judges instead of five, World Cup events and continental championships do not always feature reference judges, aerobic and acrobatic gymnastics competitions can have larger execution panels. We also note that the typical size of some of the panels changed after the 2016 Rio Olympic Games.

⁹ www.gymnastics.sport.

¹⁰ www.longines.com.

to qualify for the vault apparatus finals jump twice, each jump counting as a distinct performance in our analysis.

We analyze artistry and execution marks, including those from reference judges, but exclude marks for the difficulty component. We also exclude synchronized trampoline since its subpanels, each consisting of two judges monitoring one of the two gymnasts, are not amenable to analysis due to their small size. Since we are interested in the raw marks reported by judges, we disregard penalties outside their jurisdiction and post-evaluation aggregation. We do not distinguish between reference and regular panel judges, which are all part of a single and enlarged panel for our analysis. We also restrict the dataset to completed performances with a median panel mark of at least 7.0 to exclude noisy aborted or poorly executed routines, e.g. when gymnasts step outside the trampoline boundaries or fall from the horizontal bar mid-routine. The reported marks are less consistent for those infrequent routines. Furthermore, as aborted routines are of negligible interest for the assessment of judging quality nor targeted for intentional misjudgements, we discard them from the analysis. This excludes 9.9% of the original data points. We also merge the execution and artistry components since both panels exhibit a similar behavior.

Table 2 also shows the number of marks given to athletes of the same nationality as the judge as well as their direct competitors. The share of same-nationality marks is around 2–5%, depending on the discipline. Most same-nationality marks occur in qualifications and all-around finals where they are difficult to avoid as much more athletes of different nationalities are competing than in a single-apparatus final. We define direct competitors as gymnasts ranked immediately ahead or behind an athlete of the same nationality as the judge. A gymnast can have different competitors in qualifying rounds and finals, and the number of direct competitors can be higher than two if multiple gymnasts obtain the same score.

4 True performance quality and intrinsic judging error variability

If the *true* performance quality of a gymnastic routine was available, it would be straightforward to compare the marks of each judge against it. In practice, however, the true performance level is unknown and the FIG typically derives post-competition *control scores* with outside judging panels and video reviews. Unfortunately, the FIG does not provide accurate control scores for every performance: the number of control scores and how they are obtained

depends on the discipline and competition. Besides, even when a control score is available, the codes of points might be ambiguous. This still results in an approximation of the true performance, albeit a very good one. Control scores derived post-competition can also be biased, for instance if people deriving them know who the panel judges are, and what marks they initially gave. For all these reasons, in our analysis, we train our model using the median judging mark of each performance as the control score. We use the median because it is less prone to be affected by biased and erratic judges compared to other aggregation measures based on means or trimmed means. Whenever marks by reference judges, superior juries and post-competition reviews are available, we include them with panel judges and take the median mark over this enlarged panel, henceforth increasing the accuracy of our proxy of the true performance quality.

Working with approximations of the true performance level, which remains unknown, has implications for evaluating judges. Even though the median panel mark is a good approach in the aggregate considering that most judging panels contain a majority of good judges, relying on the median for a single performance can be misleading. A large deviation from the aggregate score for a specific performance is not necessarily an indicator of a judging error but can also mean that the judge is accurate but out of consensus with the other inaccurate judges. Since the FIG assures, and our work confirms, that the vast majority of judges at the international level are fair, such random effects are negligible for the longitudinal and aggregate assessments of judges.

Table 3 summarizes the notation we henceforth use to parameterize the *intrinsic judging error variability* and the *marking score*. Let $s_{p,j}$ be the mark of judge j for a performance p whose true quality level λ_p is approximated by the control score $c_p \triangleq \text{median}(s_{p,j})$. The *judging error* $e_{p,j}$ of judge j for performance p can thus be approximated by the *judging discrepancy* $\delta_{p,j} \triangleq s_{p,j} - c_p$. Figure 1 shows the sample standard deviation of the judging error

Table 3: Mathematical notation.

λ_p	True quality level of performance p
c_p	Control score of performance p
$s_{p,j}$	Mark of judge j for performance p
$e_{p,j}$	True judging error of judge j for performance p
$\delta_{p,j}$	Judging discrepancy of judge j for performance p
$m_{p,j}$	Marking score for performance p by judge j
M_j	Marking score of judge j
$\sigma_d(c_p)$	Intrinsic judging error variability of discipline d

Figure 1: Intrinsic judging error variability versus control score in artistic gymnastics.

as a function of the control score for artistic gymnastics (men and women combined). The frequency is the number of observed performances with a given control score in our data, and the fitted curve is an exponential weighted least-squares regressions of the data. The weighted root-mean-square deviation (RMSD) of the regression is 0.009. The main observation is that the magnitude of the judging error is highly heteroscedastic: judges are much more accurate as the performance quality improves. We call the fitted standard deviation of the judging error $\hat{\sigma}_d(c_p)$ the *intrinsic discipline judging error variability*. It quantifies the average error made by an average judge as a function of the performance quality. It models the peculiarities and judging challenges of the discipline. More generally, the estimator for $\hat{\sigma}_d(c_p)$ depends on the discipline (or apparatus) d under evaluation and the control score c_p of the performance, and is given by

$$\hat{\sigma}_d(c_p) \triangleq \max(\alpha_d + \beta_d e^{\gamma_d c_p}, 0.05). \quad (1)$$

Marks close to the perfect score of 10 are very rarely achieved in gymnastics (we have none in our dataset) and $\hat{\sigma}_d(c_p)$ is always smaller in magnitude compared to the distance to the boundaries. We can therefore disregard the mathematical implications of the bounded marking range, i.e., that the standard deviation is bounded by the median of observations on an interval. However, for apparatus such as women's floor exercise the best fitted curves go to zero before a control score of 10. Since athletes might get higher marks than in our original dataset in future competitions, we use $\max(\cdot, 0.05)$ as a fail-safe mechanism to avoid comparing judges' marks to a very low and even negative extrapolated intrinsic error variability in the future. We emphasize that all the disciplines and

apparatus can be well-approximated by the exponential regression and refrain from plotting every apparatus and discipline. Acrobatic gymnastics, for which we do not have as much data, exhibits the largest weighted root-mean-square deviation with $\text{RMSD} \approx 0.085$. The artistry judges in acrobatic and aerobic gymnastics show the same pattern of heteroscedasticity in the judging error as well.

Trampoline illustrates why we discard scores below 7. The left side of Figure 2 displays gymnasts who aborted their routine before completing all their jumps, for instance by losing balance or landing a jump outside the center of the trampoline. The fitted curve based on the completed routines in the cropped dataset, with aborted routines represented with rings instead of filled circles, accurately approximates the empirical standard deviation. The behavior observed in trampoline appears in other sports with aborted routines or low scores such as halfpipe ski and snowboard (Heiniger and Mercier 2019a) and can be modelled with a concave parabola. This, however, decreases the accuracy of the regression for the best performances, which is undesirable for our analysis. We therefore perform the same exponential regression in trampoline to determine the intrinsic judging error variability as for the other disciplines.

5 Marking score

The marking score, quantifying the accuracy of a judge compared to her/his peers, must have the following properties: First, it must not depend on the skill level of the gymnasts evaluated. A judge should not be penalized nor advantaged if he judges an Olympic final with the

Figure 2: Intrinsic judging error variability versus control score in trampoline. The rings indicate aborted routines.

world's best eight gymnasts as opposed to a preliminary round with 200 gymnasts. Second, it must allow judges comparisons across apparatus, disciplines, and competitions. The marking score of a judge is thus based on three parameters:

1. The control scores of the performances
2. The marks given by the judge
3. The intrinsic judging error variability of the apparatus/discipline

The marking score of performance p by judge j is

$$m_{p,j} \triangleq \frac{e_{p,j}}{\hat{\sigma}_d(c_p)} \approx \frac{\delta_{p,j}}{\hat{\sigma}_d(c_p)} = \frac{s_{p,j} - c_p}{\hat{\sigma}_d(c_p)}. \quad (2)$$

It expresses the judging error as a function of the intrinsic discipline judging error variability for the given control score. The overall marking score for judge j is given by

$$M_j \triangleq \sqrt{E[m_{p,j}^2]} = \sqrt{\frac{1}{n} \sum_{p=1}^n m_{p,j}^2}. \quad (3)$$

The marking score quantifies the accuracy of a judge compared to her/his peers. The marking score of a perfect judge is 0, and a judge whose judging error is always equal to the intrinsic judging error variability $\hat{\sigma}_d(c_p)$ has a marking score of 1.0. The mean squared error weights outliers heavily, which is desirable for evaluating judges.

Figure 3A shows the boxplots of the marking scores for all the judges for each apparatus in men's artistic gymnastics using the regression from Figure 1. The acronyms are defined in Table 4. The first observation is the accuracy range among judges: the average judging error of the

Table 4: The artistic gymnastics apparatus and their acronyms.

Acronym	Apparatus
BB	Balance beam (women)
FX	Floor exercise (men and women)
HB	Horizontal bar (men)
PB	Parallel bars (men)
PH	Pommel horse (men)
SR	Still rings (men)
UB	Uneven bars (women)
VT	Vault (men and women)

best judges, whose marking score is ≈ 0.5 , is three times smaller than the average error of the most erratic judges with marking scores of ≈ 1.5 . This is striking considering that our dataset only includes the most important international competitions with well-recognized judges. The second observation is that there are significant differences between apparatus. Pommel horse, for instance, is intrinsically more difficult to judge accurately than vault and floor exercise. The FIG confirms that the alternative, i.e., that judges in pommel horse are less competent than judges in men's vault or men's floor exercise, is highly unlikely considering that most judges officiate on all apparatus. The differences between floor and vault on one side and pommel horse on the other side were previously observed in punctual competitions (Atikovic et al. 2011; Bučar et al. 2012; Pajek et al. 2013). Note that the better accuracy of vault judges does not make it easier to rank the gymnasts since many gymnasts execute the same jumps at a similar performance level.

A highly desirable feature for the marking score is to be comparable between apparatus and disciplines, which proves difficult with one overall formula. We thus estimated the intrinsic judging error variability $\hat{\sigma}_d(c_p)$ for each apparatus (instead of grouping them together) and used the resulting regressions to recalculate the marking scores. The results for men's artistic gymnastics, presented in Figure 3B, now show a good uniformity and make it simpler to compare judges from different apparatus with each other. A pommel horse judge with a marking score of 0.9 is average, and so is a vault judge with the same marking score. This has allowed us to define a single set of quantitative to qualitative thresholds applicable across all the gymnastics apparatus and disciplines.

We use the same approach for women's artistic gymnastics and every other gymnastics discipline. We do not report all the results, but they are strikingly similar to those in men's artistic gymnastics and lead to a comparable

Figure 3: Distribution of the overall marking scores per men's artistic gymnastic apparatus using (A) one overall formula; (B) an individual formula per apparatus. The acronyms are defined in Table 4, and the numbers between brackets are the number of judges per apparatus in the dataset.

measure of individual judging accuracy across disciplines when adjusted for the notable differences at the apparatus level. For instance, group routines in rhythmic and aerobic gymnastics are more difficult to judge than individual routines. The trampoline disciplines exhibit the largest discrepancies: tumbling is much more difficult to judge than individual trampoline, which in turn is much more difficult to judge than double mini-trampoline.

We must mention two marking score caveats. First, comparing judges with each other and not based on their objective performance may be problematic. An apparatus with only outstanding judges will trivially have half of them with a marking score below the median, and the same would be true of an apparatus with only mediocre judges. However, from our observations and discussions with the FIG, the difference between the most and least accurate judges is similar for all apparatus and disciplines, and the marking score provides valuable information. Second, since the median panel score is an approximation of the true performance quality, marking scores only based on small events such as finals with eight gymnasts are subject to more errors and randomness. However, in the aggregate over hundreds and even thousands of evaluations per judge, these random effects do fade out and the verdict is irrefutable: for every apparatus and discipline the best judges, over many years, are systematically two to three times more accurate than the worst judges.

6 Applications of the marking score

6.1 Gender discrepancies: women are more accurate judges than men

In artistic gymnastics, men apparatus are almost exclusively evaluated by men judges and women apparatus are almost exclusively evaluated by women judges. When calculating the marking score according to Eq. (3) but with one overall formula for the intrinsic judging variability as shown in Figure 3, the marking scores for women apparatus are lower than those of men apparatus. The average woman evaluation is $\approx 15\%$ better than the average man evaluation. More formally, we ran a one-sided Welch's t -test with the null-hypothesis that the mean of the marking scores of men is smaller than or equal to the mean marking score of women. We obtained a p -value of 10^{-15} , leading to the rejection of the null-hypothesis.

A first hypothesis that can explain this difference is that in artistic gymnastics men judges have more complex

judging tasks and less training. The formation and accreditation process is different for men and women judges. Men, who must judge six apparatus, receive less training per apparatus than women, who must only judge four. Furthermore, men routines include 10 different technical elements to evaluate, whereas women routines only include eight. To obtain more insight, we compared women and men judges in trampoline, which has mixed judging panels and the same accreditation and evaluation processes for both genders. Men and women judges in trampoline receive the same training and execute the same judging tasks. The results are shown in Figure 4A. The difference observed in artistic gymnastics is less pronounced but remains in trampoline: women judge more accurately than men. We suspect that an important contributor of this judging gender discrepancy in gymnastics is the larger pool of women practicing the sport, which increases the likelihood of having more good women judges at the top of the pyramid since nearly all judges are former gymnasts from different levels. As an illustration, a 2007 Survey from USA Gymnastics reported four times more women gymnasts than men gymnasts in the USA (Gymnastics Member Club 2008). A 2004 report from the ministère de la Jeunesse, des Sports et de la Vie Associative reported a similar ratio in France (STAT Info 2004). Accurate information on participation per gender is difficult to come by, but fragmentary results indicate a similar participation gender imbalance in trampoline (Silva et al. 2017).

On a different note, we did not observe any mixed-gender bias in trampoline, i.e., judges are not biased in favor of same-gender athletes. This in opposition to other sports such as handball where gender bias by referees led to transgressive behaviors (Souchon et al. 2004).

Figure 4: Distribution of marking scores for (A) trampoline execution judges by gender; (B) panel versus reference execution judges in artistic gymnastics.

6.2 Reference judges are not more accurate

At each competition, execution judges are randomly selected from a set of accredited judges submitted by national federations. In contrast, reference judges are hand-picked by the FIG, and the additional power granted to them is based on the assumption that execution judges are sometimes inaccurate or biased. To test this assumption, we compared the marking scores of the execution panel and reference judges. The results for artistic gymnastics are shown in Figure 4B.¹¹ Although this is obvious by inspection, a two-sided Welch's t -test returned a p -value of 0.18 and we could not reject the null-hypothesis that both means are equal. We ran similar tests for the other gymnastics disciplines and in all instances reference judges are either statistically indistinguishable from the execution panel judges, or less accurate in the aggregate.

6.3 Outlier detection

We can use the marking score to signal judging marks that are unlikely high or low. While large differences to the control score as a multiple of the intrinsic judging error variability $\hat{\sigma}_d(c_p)$ already raise questions, the problem with this simple approach is that an erratic judge has a lot of outliers, and an accurate judge none. This is not necessarily relevant because an erratic judge can be unbiased and a precise judge can be dishonest. Instead of using the same standard deviation for all the judges, we scale the standard deviation by the overall marking score of each judge, and flag the judging scores that satisfy

$$|\hat{e}_{p,j}| > \max(2.5 \cdot \hat{\sigma}_d(c_p) \cdot M_j, 0.1). \quad (4)$$

We use $\max(\cdot, 0.1)$ to ensure that a difference of 0.1 from the control score is never an outlier. Eq. (4) flags approximately 2% of the marks and the results are shown in Figure 5. The advantage of the chosen approach is that it compares each judge to herself/himself, that is, it is more stringent for precise judges than for erratic judges. The disadvantage of the chosen approach is that one might think that a judge without outliers is good, which is false. The marking score and outlier detection work in tandem: a judge with a bad marking score is erratic, thus bad no matter how many outliers he has. It is important to note that we cannot automatically infer conscious bias, chicanery or cheating from

Figure 5: Distribution of the judging errors in artistic gymnastics. Dots in red are more than $2.5 \cdot \hat{\sigma}_d(c_p) \cdot M_j$ away from the control score. To improve the visibility, we aggregate the points on a 0.1×0.1 grid and shift the outliers (red dots) by 0.05 on both axes.

an outlier mark. A flagged evaluation can be a bad but honest mistake, indicate that a judge is out of consensus with the other judges who might be wrong at the same time, or caused by a data entry error. This information is nevertheless useful. The FIG systematically reviews these outliers after the most important international competitions and can sanction judges. For instance, judges with large positive same-nationality outliers are prohibited from judging at the Olympic Games.

7 National bias analysis

We leverage the intrinsic judging error variability and the marking score to study the national bias of international gymnastics judges during the 2013–2016 Olympic cycle. As opposed to prior work, we express the national bias as a multiple of the intrinsic judging error variability. This implicitly takes into account that as the performance quality improves, a small nominal bias becomes larger compared to the other sources of error and sufficient to affect the ranking of the performances. Furthermore for the best athletes, a judge with a large nominal national bias will likely be suspected of intentional misjudgment, while even a small nominal national bias can remain very large compared to the other sources of judging errors and impact the ranking of the athletes. We believe that judges are, deliberately or unconsciously, aware of this, and that the national bias follows in magnitude this intrinsic variability. We elaborate our regression model in Section 7.1, followed by the results of our analysis in Sections 7.2 and 7.3, and the impact of the national bias on rankings in Section 7.4.

¹¹ In Figure 4B, judges have at least one marking score per apparatus for which they evaluated gymnasts. A judge has two marking scores on a single apparatus when appearing on the regular execution panel and on the reference panel for different events.

7.1 Regression model

The score of a judge j for performance p can be modeled by

$$s_{p,j} = \lambda_p + (\mu_j + \beta_{SN} \cdot \mathbb{1}_{SN} + \beta_{COMP} \cdot \mathbb{1}_{COMP}) \cdot \hat{\sigma}_d(c_{p,-j}) + \epsilon_{p,j} \quad (5)$$

where λ_p is the true but unknown performance quality, μ_j is the general tendency of a judge who consistently applies the judging regulations too harshly or too generously,¹² and $\epsilon_{p,j}$ is a judge/performance-specific random error-term. The national bias of a judge in favor of an athlete of the same nationality is integrated into the model with the binary indicator $\mathbb{1}_{SN}$ and the parameter β_{SN} , determining the extent of the bias. National bias against direct competitors of same-nationality athletes, who are determined as explained in Section 3, are included analogously with $\mathbb{1}_{COMP}$ and β_{COMP} . We assume that the general tendency μ_j and the biases $(\beta_{SN}, \beta_{COMP})$ are heteroscedastic like the other sources of error: intentional and unintentional misjudgements are smaller for the best athletes. Being too far off the panel score is very suspicious and is immediately noticed by officials. We therefore suppose that judges do adapt their scoring behaviour to not stand out too much, no matter if their bias is intentional or not. We thus assess the individual general tendency μ_j and the national bias $(\beta_{SN}, \beta_{COMP})$ as a multiple of the intrinsic judging error variability $\hat{\sigma}_d(c_{p,-j})$. Note that for the analysis of national bias, the control score $c_{p,-j}$ is determined as the median panel score excluding judge j . This removes any possible national or personal bias from the control score.¹³

To restrict the number of variables in the regression model, we proxy the true performance quality λ_p by the median panel mark. The same justifications and concerns discussed in earlier sections for the intrinsic judging error variability and the marking score still apply: the median aggregation penalizes a fair judge if the panel is predominantly composed of imprecise or biased judges, but remains a valid approximation in the aggregate. We furthermore determine the general judge tendency μ_j beforehand as

$$\hat{\mu}_j = \frac{1}{n} \sum_p \frac{s_{p,j} - c_{p,-j}}{\hat{\sigma}_d(c_{p,-j})} \quad (6)$$

which is the average judging error expressed as a multiple of $\hat{\sigma}_d(c_{p,-j})$. Finally, we model the error term $\epsilon_{p,j}$ of a specific judge j as a normal random variable with mean zero and standard deviation $\hat{\sigma}_d(c_{p,-j}) \cdot M_j$, i.e., the

intrinsic discipline judging error variability multiplied by the judge-specific marking score. The term $\hat{\sigma}_d(c_{p,-j}) \cdot M_j$ is the *intrinsic judging error variability* of a specific judge in discipline d . We group the known and predetermined values into a single outcome variable $d_{p,j}$ which represents the judging error corrected for the judge-specific tendency. The final model is

$$\begin{aligned} d_{p,j} &\equiv s_{p,j} - c_{p,-j} - \hat{\mu}_j \cdot \hat{\sigma}_d(c_{p,-j}) \\ &= (\beta_{SN} \cdot \mathbb{1}_{SN} + \beta_{COMP} \cdot \mathbb{1}_{COMP}) \cdot \hat{\sigma}_d(c_{p,-j}) + \epsilon_{p,j} \\ &\text{where } \epsilon_{p,j} \sim \mathcal{N}(0, \hat{\sigma}_d^2(c_{p,-j}) \cdot M_j^2) \end{aligned} \quad (7)$$

and the parameters can be estimated with the generalized least squares method (Aitken 1936). We can use Eq. (7) to estimate the national bias $(\beta_{COMP}, \beta_{SN})$ by apparatus, by discipline, by nationality and by judge.

7.1.1 Constrained model versus full model

Note that with a very large dataset we can also estimate the model given by Eq. (5) directly, without estimating $c_{p,-j}$ and μ_j *a priori*. This requires the estimation of a fixed effect for every judge (μ_j) and every performance (λ_p), leading to a high-dimensional model. Table 5 compares the estimates of β_{SN} and β_{COMP} for the constrained and full models, split per competition stage for each discipline in our dataset. We observe that the results almost coincide. Two other variants, where we calculate only one of the two parameters *a priori* and include the other in the model, give similar results. Hence, the estimated national bias is robust to the choice of the model, and we restrict the subsequent analysis in the rest of this section to the constrained model from Eq. (7), for which we determine the control scores $c_{p,-j}$ and the judge tendencies μ_j beforehand.

7.2 National bias by discipline

Table 6 shows the outcome of the general linear model specified by Eq. (7). The results are split by discipline and stage of the event. While ‘All gymnasts’ encompasses all performances in the dataset, ‘Top 8 finalists’ only includes the top eight gymnasts in the final stage of a competition (apparatus and all-around finals), for whom a national bias would be particularly undesirable. Because the general linear model includes the functional heteroscedasticity variable $\hat{\sigma}_d(c_{p,-j})$, the estimated parameters β_{SN} and β_{COMP} are to be interpreted as a multiple of $\hat{\sigma}_d(c_{p,-j})$. For instance, $\beta_{SN} = 0.5$ means that the bias level in favor of same-nationality gymnasts is half the intrinsic discipline judging

¹² μ_j is generally very small, i.e., below 0.02, however for some judges this deviation reaches 0.2 points.

¹³ For the intrinsic judging error variability $\hat{\sigma}_d(c_{p,-j})$, removing judge j does not notably alter the results of the previous sections.

Table 5: Regression results by discipline and competition stage for the full model (Eq. (5)) and the restricted model (Eq. (7)).

		All gymnasts		Top 8 finalists	
		Restricted model Eq. (7)	Full model Eq. (5)	Restricted model Eq. (7)	Full model Eq. (5)
Acrobatic	β_{SN}	0.05 (0.05)	0.04 (0.06)	0.05 (0.24)	0.26 (0.30)
	β_{COMP}	−0.03 (0.03)	−0.05 (0.04)	−0.07 (0.07)	0.06 (0.21)
Aerobic	β_{SN}	0.23 (0.06)**	0.28 (0.07)**	0.41 (0.14)**	0.37 (0.16)*
	β_{COMP}	−0.02 (0.03)	−0.06 (0.05)*	−0.02 (0.05)	−0.13 (0.12)
Artistic (M)	β_{SN}	0.39 (0.03)**	0.45 (0.04)**	0.53 (0.10)**	0.68 (0.12)**
	β_{COMP}	−0.01 (0.02)	0.01 (0.03)	−0.03 (0.03)	−0.04 (0.10)
Artistic (F)	β_{SN}	0.26 (0.04)**	0.27 (0.05)**	0.52 (0.13)**	0.62 (0.16)**
	β_{COMP}	−0.07 (0.02)**	−0.11 (0.04)**	−0.01 (0.04)	−0.05 (0.13)
Rhythmic	β_{SN}	0.32 (0.04)**	0.36 (0.05)**	0.18 (0.17)	0.27 (0.19)
	β_{COMP}	−0.05 (0.02)*	−0.08 (0.04)*	−0.02 (0.05)	−0.01 (0.17)
Trampoline	β_{SN}	−0.06 (0.05)	−0.10 (0.06)	0.02 (0.12)	0.10 (0.16)
	β_{COMP}	−0.04 (0.03)	−0.02 (0.05)	−0.10 (0.06)	−0.02 (0.13)

Significance code: $p < 0.05^*$, $p < 0.01^{**}$. Values denote point estimates, and standard errors are in brackets.

Table 6: Regression results by discipline.

		All gymnasts			Top 8 finalists		
		Estimate (se)	t-stat.	p-value	Estimate (se)	t-stat.	p-value
Acrobatic	β_{SN}	0.05 (0.05)	0.98	0.327	0.05 (0.24)	0.19	0.848
	β_{COMP}	−0.03 (0.03)	−1.14	0.252	−0.07 (0.07)	−1.08	0.281
Aerobic	β_{SN}	0.23 (0.06)	3.56	0.000**	0.41 (0.14)	2.86	0.004**
	β_{COMP}	−0.02 (0.03)	−0.70	0.482	−0.02 (0.05)	−0.40	0.689
Artistic (M)	β_{SN}	0.39 (0.03)	12.32	0.000**	0.53 (0.10)	5.40	0.000**
	β_{COMP}	−0.01 (0.02)	−0.61	0.544	−0.03 (0.03)	−1.27	0.205
Artistic (F)	β_{SN}	0.26 (0.04)	6.22	0.000**	0.52 (0.13)	3.96	0.000**
	β_{COMP}	−0.07 (0.02)	−2.98	0.003**	−0.01 (0.04)	−0.41	0.684
Rhythmic	β_{SN}	0.32 (0.04)	7.08	0.000**	0.18 (0.17)	1.07	0.285
	β_{COMP}	−0.05 (0.02)	−1.99	0.047*	−0.02 (0.05)	−0.53	0.595
Trampoline	β_{SN}	−0.06 (0.05)	−1.10	0.271	0.02 (0.12)	0.16	0.870
	β_{COMP}	−0.04 (0.03)	−1.26	0.207	−0.10 (0.06)	−1.59	0.111

Significance code: $p < 0.05^*$, $p < 0.01^{**}$. Estimated parameters (β_{SN} , β_{COMP}) indicate the national bias as a multiple of the estimated intrinsic discipline judging error variability $\hat{\sigma}_d(c_{p,-j})$. To obtain the nominal bias for a given performance quality $c_{p,-j}$, multiply the estimated parameter with $\hat{\sigma}_d(c_{p,-j})$.

error variability for a specific performance level. The results reveal that in aerobic, artistic and rhythmic gymnastics, judges in the aggregate mark same-nationality gymnasts significantly higher than the other panel judges, whereas national bias is not a systemic issue in acrobatic gymnastics and trampoline. In aerobic and artistic gymnastics in particular, national bias is even more pronounced for finalists than during earlier stages of competitions, in other words judges bend the rules further when it counts. This does not necessarily imply that the nominal bias is higher for the best gymnasts since the intrinsic judging error

variability decreases as the performance level improves, but instead that the magnitude of the national bias compared to the other sources of judging errors increases for the best athletes.

The most severe bias appears in men's artistic gymnastics, where judges give gymnasts of the same country an average bonus of almost half the intrinsic discipline judging error variability. The best men artistic gymnasts in the world typically get marks between 8.5 and 9.5 depending on the apparatus. The national bias $\beta_{SN} = 0.53$ for the top men finalists corresponds to a nominal bias

of between 0.05 and 0.10 points depending on the apparatus, or 7–10% of the total deductions of the performance. Considering the narrow gaps between the best gymnasts, this is a worrying discovery, both in relative and in absolute terms. We point out again that this is for an average judge; the most biased judges are significantly worse! Even though we observed a few judges with a statistically significant bias against direct competitors of same-nationality athletes, Table 6 shows that in the aggregate this bias, if it exists, is certainly very small. This is in line with prior research (Ansorge and Scheer 1988; Popović 2000).

7.3 National bias by nationality and judge

We can apply the same general linear regression model from Eq. (7) to estimate national bias by nationality and by judge and identify whether specific countries or judges are particularly prone to national bias. In the subsequent analysis we focus only on the positive bias toward same-nationality gymnasts which we have shown to be much larger than the penalization of direct competitors. Figure 6 shows the estimated national bias per country against the number of same-nationality marks in men's artistic gymnastics. The national bias is once again expressed as a multiple of the intrinsic discipline judging error variability $\hat{\sigma}_d(c_{p,-j})$. The figure also shows the weighted empirical cumulative distribution function (wECDF) by the number of same-nationality marks. The distribution of the estimated parameter is centered around the magnitude of the national bias $\beta_{SN} = 0.39$ and for about 90% of countries the point estimates are positive. The scatterplot further

Figure 6: Estimated national bias versus same-nationality marks by nationality in men's artistic gymnastics together with the weighted empirical cumulative distribution function (wECDF) of the estimations. The statistically significant estimations ($p < 0.01$) and ($p < 0.001$) are highlighted.

highlights results with p -values $p < 0.01$ and $p < 0.001$ located in the top right quadrant of the figure. As we simultaneously test 66 countries, the p -values have to be interpreted with caution as multiple testing concerns apply. Despite this disclaimer, our results are unambiguous that national bias is certainly present for some countries. Using a Holm–Bonferroni correction (Holm 1979), 6 out of 66 countries still yield a statistically significant national bias estimate at a 0.05 level. We observe similar results in women's artistic gymnastics and rhythmic gymnastics. Note that in figure skating Zitzewitz (2006) found that the less transparent countries from the 2001 Transparency International's Corruption Perceptions Index had larger nationalistic biases. We did not observe this in our gymnastics dataset.

Figure 7 shows the estimated national bias per judge against the number of same-nationality marks in men's artistic gymnastics. The scatterplot again shows the wECDF and the p -values below 0.01 and 0.001. The point estimates are positive for about 80% of judges. Figure 7 exhibits a similar shape as Figure 6, but with fatter tails and a larger dispersion of the bias. For a judge to exhibit a low p -value, she/he must have shown either a very large bias, or a smaller bias but supported by more data points. With individual bias estimated for 264 judges, the statistical significance is again to be treated carefully, and false positives are certainly possible. Applying once again a Holm–Bonferroni correction (Holm 1979), only one judge exhibits a statistically significant national bias estimate at the 0.05 level. This reduction is a consequence of the large number of tests and the small number of same-nationality performances for most individual judges. However, there are judges with

Figure 7: Estimated national bias versus same-nationality marks by judge in men's artistic gymnastics together with the weighted empirical cumulative distribution function (wECDF) of the estimations. p -values below ($p < 0.01$) and ($p < 0.001$) are highlighted. Checked observations are South Korean judges.

very small p -values whose estimated national bias is two to three times larger than the intrinsic discipline judging error variability. In other words, their national bias is two and even three times larger than all the sources of errors of an average, fair judge. Even though we cannot infer intent without other evidence such as compromising communications or bribes, this is worrying, and whether this bias is conscious or not, or caused by other improbable random causes, is irrelevant: the FIG expects its best judges to be unbiased and applies a “better safe than sorry” precaution principle by prohibiting potentially biased judges from officiating at the Summer Olympics.

Looking at the individual judge results, we must point out that it can be misleading to infer systemic national bias, or lack thereof, based on the average national bias of a country. To illustrate this, we highlight in Figure 7 the estimated national bias of all South Korean men’s artistic gymnastics judges, which varies from judge to judge. Despite this, a per country analysis remains relevant, because a country with a high overall national bias might be a sign of undue pressure put on judges to inflate the marks of their best gymnasts.

7.4 Impact of national bias on rankings

The prevalent bias in favor of same-nationality gymnasts naturally raises the question of its impact on the competitions’ rankings. We study this further and focus on the apparatus and all-around finals in artistic gymnastics. We apply the scoring aggregation procedure defined in the Codes of Points, and study ranking distortions by calculating the rankings with and without the marks of same-nationality judges. Whenever we observe a same-nationality mark, we discard it and calculate the trimmed mean of the remaining panel marks. If a discarded mark comes from a reference judge, the average reference score simply becomes the mark of the second reference judge.

In the apparatus finals, 4 out of 740 performances in our dataset include a same-nationality evaluation. For one of the four gymnasts, the final score of the performance is boosted by 0.267 points due to the presumable national bias of a reference judge. The FIG and the other athletes were lucky in this instance because the gymnast in question finished last and far behind the other finalists, so this scoring discordance had no effect on the ranking. In most other instances a difference of this magnitude would have tweaked the position of the gymnast by a few ranks. This example further illustrates the danger of granting more power to small panels of reference judges. In the all-around finals, evaluations by same-nationality judges are more difficult to avoid and occur in 330 out of 1990 performances

in our dataset. In 42% of these all-round finals, removing the potentially biased same-nationality marks changes the final ranking of the gymnasts, including one podium.

8 Marking score versus ranking scores

The ranking of the gymnasts is determined by their scores, which are themselves aggregated from the marks given by the judges. All similar sports follow the same approach, although in figure skating, until 2004, the final ranking of the athletes was based on the ordinal ranking of each judge. The old iteration of JEP used a rudimentary *ranking score* to evaluate to what extent judges ranked the best athletes in the right order. In a vacuum this makes sense: swapped medalists at the Olympic Games and World Championships are a major source of scandal for sports such as gymnastics. The FIG wants to select the most deserving gymnasts for the finals, award the medals in the correct order, and expects accurate judges to rank athletes correctly. In this section we show that providing an objective assessment of the judges based on the order in which they rank the best athletes is problematic, and we recommended that the FIG stops using this approach.

The mathematical comparison of rankings is closely related to the analysis of voting systems and has a long and rich history dating back to the work of Ramon Llull in the 13th century. Two popular metrics on the set of weak orders are Kendall’s τ distance (Kendall 1938) and Spearman’s footrule (Spearman 1904), both of which are within a constant fraction of each other (Diaconis and Graham 1977). In recent years, Kumar and Vassilvitskii (2010) generalized these two metrics by taking into account element weights, position weights, and element similarities. Their motivation was to find the ranking minimizing the distance to a set of search results from different search engines.

Definition 8.1. Let r be a ranking of n competitors. Let $w = (w_1, \dots, w_n)$ be a vector of element weights. Let $\delta = (\delta_1, \dots, \delta_n)$ be a vector of position swap costs where $\delta_1 \triangleq 1$ and δ_i is the cost of swapping elements at positions $i-1$ and i for $i \in \{2, 3, \dots, n\}$. Let $p_i = \sum_{j=1}^i \delta_j$ for $i \in \{1, 2, \dots, n\}$. We define the mean cost of interchanging positions i and r_i by $\bar{p}(i) = \frac{p_i - p_{r_i}}{i - r_i}$. Finally, let $D: \{1, \dots, n\} \times \{1, \dots, n\}$ be a non-empty metric and interpret $D(i, j) = D_{ij}$ as the cost of swapping elements i and j . The generalized Kendall’s τ distance (Kumar and Vassilvitskii 2010) is

$$K'^* = K'_{w, \delta, D}(r) = \sum_{s > t} w_s w_t \bar{p}_s \bar{p}_t D_{st} [r_s < r_t]. \quad (8)$$

Note that K'^* is the distance between r and the identity ranking $id = (1, 2, 3, \dots)$. To calculate the distance between two rankings r^1 and r^2 , we calculate $K'(r^1, r^2) = K'_{w, \delta, D}(r^1 \circ (r^2)^{-1})$, where $(r^2)^{-1}$ is the right inverse of r^2 . These generalizations are natural for evaluating gymnastics judges: swapping the gold and silver medalists should be evaluated more harshly than inverting the ninth and tenth best gymnasts, but swapping the gold and silver medalists when their marks are 9.7 and 9.6 should be evaluated more leniently than if their marks are 9.7 and 8.7. All this can be configured by the parameters w_i, δ_i, D_{ij} . Setting $w_i = \delta_i = D_{ij} = 1$ yields the original version of Kendall's τ distance (Kendall 1938), which simply counts the number of bubble sort swaps required to transform one ranking into the other. Defining the element swap costs by $D_{ij} = |c_i - c_j|$ decreases the penalty of swaps as the marks get closer to each other; in particular, swapping two gymnasts with the same control score $c_i = c_j$ incurs no penalty. Altering the position swap costs to $\delta_i = \frac{1}{i}$ increases the importance of having the correct order as we move towards the gold medalist.

To test the relevance of ranking scores as a measurement of judging accuracy, we ran several simulations with the aforementioned parameterizations and compared them to our marking score. As an example for this article, we use the men's floor exercise finals at the 2016 Rio Olympic Games. We first calculate the control scores $c_1, c_2, \dots, c_8$ of the eight finalists from the marks given by the seven execution judges (five panel judges and two reference judges). We then simulate the performance of 1000 average judges $j \in \{1, 2, \dots, 1000\}$ by randomly creating, for each of them, eight marks $s_{1,j}, s_{2,j}, \dots, s_{8,j}$ for the eight finalists using a normal distribution with mean c_p and standard deviation $\hat{\sigma}_d(c_p)$ for $p \in \{1, 2, \dots, 8\}$. We then calculate, for each judge, the marking score as well as three ranking scores based on Eq. (8). The basic Kendall's τ distance results in a correlation 0.40 between the marking score and the ranking score. Setting the swap penalty costs $D_{ij} = |c_i - c_j|$ increases the correlation to 0.50 and both, to some extent, measure the same thing. Further varying the position swap costs with $\delta_i = \frac{1}{i}$ lowers the correlation to 0.35, but we now penalize accurate judges who unluckily make mistakes at the wrong place, and reward erratic judges who somehow get the podium in the correct order.

It is unclear how to parameterize the ranking score; it is either redundant with the marking score, or too uncorrelated to be of any practical value. The marking score already achieves our objectives. It is based on the actual performances of the gymnasts over hundreds of

performances for each judge and reflects bias and cheating, as these involve changing the marks up or down for some of the performances. Furthermore, the FIG is adamant that a theoretical judge who ranks all the gymnasts in the correct order but is either always too generous or too strict is not a good judge because she/he does not apply the Codes of Points properly. From these observations, the FIG stopped using ranking scores to monitor the accuracy of its judges.

9 Recommendations and conclusion

We put the evaluation of international gymnastics judges on a strong mathematical footing using a rigorous yet simple methodology. This has led to a better assessment of current judges, and will improve judging in the future. The main novelty of our approach over prior work is that we systematically leverage and integrate the intrinsic judging error variability into our models. This intrinsic variability is heteroscedastic and can be accurately modeled with weighted least-squares exponential regressions using historical data. We develop a marking score that evaluates the accuracy of the marks given by judges as a multiple of the intrinsic judging error variability. The marking score can be used across disciplines, apparatus and competitions. We also quantify national bias as a multiple of the intrinsic judging error variability. Intuitively this approach makes sense: judges are very accurate when evaluating the best gymnasts, thus even a small nominal error or bias can have a large impact on the rankings.

Our main observation is that there are significant differences between the best and the worst judges; this in itself is not surprising, but we can now quantify this much more precisely than in the past. In every discipline, the best judges are systematically two to three times more accurate than the worst judges, and there are a few but highly biased judges whose bias in favor of their own athletes is two to three times higher than all the sources of errors of an average fair judge. The FIG can use this information to assign the best judges to the most important competitions, bestow judging awards, provide feedback to all judges so that they can improve, and get rid of highly inaccurate or biased judges.

During the course of this work, we made interesting and sometimes surprising observations and discoveries that led to recommendations and rule changes at the FIG. These recommendations, which we present next, also apply to the other sports with similar judging systems.

9.1 Monitor judging accuracy and bias longitudinally

There are clear benefits in monitoring the accuracy and bias of judges over long periods of time. Many international judges officiate for decades, and as the number of observations increases, so does our confidence that their good (or bad) evaluations is representative of their ability to judge and not due to random effects such as good (or bad) luck.

9.2 Analyze small panels cautiously

Judging evaluations based on a few marks in single events such as finals with eight gymnasts are subject to more randomness than large panels and longitudinal evaluations. This is exacerbated during live events because accurate video reviews are not yet available. A high marking score or a large bias for a specific performance is not necessarily a judging error but can also mean that the judge is accurate but out of consensus. The FIG relies on outside panels and superior juries to obtain quick feedback during competitions, but our analysis shows that like for reference judges, these observers are in the aggregate equal or worse than regular panel judges, and giving them additional power is problematic. Discrepancies between these observers and the regular panel should be viewed with circumspection. The best the FIG can do in this circumstance is to add these outside marks to the regular panel to increase its robustness until accurate control score are available post-competition.

9.3 Improve the collection of control scores

The Technical Committee (TC) of each discipline calculates control scores post-competition using video reviews. Each TC uses a different number of members and different aggregation techniques: sometimes members verbally agree on a score and other times they take the average. Furthermore the FIG cannot guarantee the accuracy and unbiasedness of the TC members: some of them might be friends with judges they are evaluating and know what marks they initially gave. Clear guidelines describing the acquisition process of these control scores must be put in place. This is paramount to guarantee the accuracy of judging evaluations on a per routine and per competition basis.

9.4 Get rid of the increased power imparted to reference judges

Having additional judges selected by the FIG is an excellent idea because it increases the size of the panels, thus

making them more robust. However, reference judges should not be granted more power. They are not better in aggregate, and the small size of the reference panels further increases the likelihood that their errors and biases have greater consequences on the final scores and rankings of the gymnasts. Starting in 2022, reference judges will be replaced by additional regular panel judges.

9.5 Retrain the model periodically

Since athletes and judges improve, and since Codes of Points are revised every four years, the intrinsic judging error variability can and should be recalibrated periodically. We recently recalibrated the models for artistic, rhythmic and trampoline gymnastics before the 2021 Tokyo Summer Olympics.

9.6 Aggregate marks more aggressively

We recommend to remove more high and low marks from judging panels before aggregating them with a more aggressive trimmed mean. Starting in 2022, the best and worst two marks will be removed from the execution panels for all the gymnastics disciplines except trampoline, which already uses the median. The resulting aggregation will be the average of the middle three marks in artistic gymnastics, and the average of the middle two marks in aerobic, acrobatic and rhythmic gymnastics.

9.7 Review training and accreditation procedures

The brevet classification of gymnastics judges is based on theoretical and practical examinations, with increasingly stringent thresholds for the higher categories. In men's artistic gymnastics¹⁴ the theoretical examination for the execution component consists in the evaluation of 30 routines, 5 per apparatus. Our statistical engine is much more precise than the FIG examinations because it tracks judges longitudinally in real conditions over thousands of evaluations. Our dataset is dominated by Category 1 judges, and even at this highest level it shows significant differences among judges. We recommend to include the longitudinal marking score of judges to the certification process. In light of our gender analysis, we also recommend that the FIG and its technical committees thoroughly review their processes to select, train and evaluate men judges in artistic gymnastics and trampoline.

¹⁴ <https://www.gymnastics.sport/site/rules/rules.php>.

For future work, we could refine our work on national bias. In particular, we could investigate complex judging behavior in gymnastics such as vote trading and compensation effects revealed in figure skating, ski jumping and dressage (Sandberg 2018; Zitzewitz 2006, 2014). Finally, in subsequent work (Heiniger and Mercier 2019a), we show that the intrinsic judging error variability in other sports with panels of judges awarding marks within a finite range has a similar heteroscedastic shape. This is the case in artistic swimming, diving, dressage, karate, figure skating, freestyle skiing and snowboard, and ski jumping. The methodology presented in this article could provide long-term monitoring and an improved national bias analysis in all these sports.

Acknowledgments: This work is the result of fruitful interactions and discussions with the other project partners. We would like to thank Nicolas Buompane, Steve Butcher, Les Fairbrother, André Gueisbuhler, Sylvie Martinet and Rui Vinagre from the FIG, Benoit Cosandier, Jose Morato, Christophe Pittet, Pascal Rossier and Fabien Voumard from Longines, and Pascal Felber, Christopher Klahn, Rolf Klapert and Claudia Nash from the Université de Neuchâtel. We would also like to thank the anonymous referees for their careful reviews of the manuscript and their insightful comments and suggestions. In particular, we thank them for encouraging us to compare the constrained and full models for the national bias analysis. A preliminary version of this work was presented at the 2017 MIT Sloan Sports Analytics Conference, and subsequent iterations published at (Heiniger and Mercier 2019b; Mercier and Heiniger 2019).

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: This work was partly funded by Longines.

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

- Aitken, A. C. 1936. "IV. — On Least Squares and Linear Combination of Observations." *Proceedings of the Royal Society of Edinburgh* 55: 42–8.
- Ansorge, C. J., and J. K. Scheer. 1988. "International Bias Detected in Judging Gymnastic Competition at the 1984 Olympic Games." *Research Quarterly for Exercise & Sport* 59 (2): 103–7.
- Atikovic, A., S. D. Kalinski, S. Bijelić, and N. A. Vukadinović. 2011. "Analysis Results Judging World Championships in Men's Artistic Gymnastics in the London 2009 Year." *SportLogia* 7 (2): 95–102.
- Bar-Eli, M., H. Plessner, and M. Raab. 2011. *Judgement, Decision Making and Success in Sport*. Wiley-Blackwell.
- Boen, F., K. V. Hoye, Y. V. Auweele, J. Feys, and T. Smits. 2008. "Open Feedback in Gymnastic Judging Causes Conformity Bias Based on Informational Influencing." *Journal of Sports Sciences* 26 (6): 621–8.
- Bučar, M., I. Čuk, J. Pajek, M. Kovač, and B. Leskošek. 2012. "Reliability and Validity of Judging in Women's Artistic Gymnastics at University Games 2009." *European Journal of Sport Science* 12 (3): 207–15.
- Campbell, B., and J. W. Galbraith. 1996. "Nonparametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments." *The Statistician* 45: 521–6.
- Damisch, L., T. Mussweiler, and H. Plessner. 2006. "Olympic Medals as Fruits of Comparison? Assimilation and Contrast in Sequential Performance Judgments." *Journal of Experimental Psychology: Applied* 12 (3): 166–78.
- Diaconis, P., and R. L. Graham. 1977. "Spearman's Footrule as a Measure of Disarray." *Journal of the Royal Statistical Society* 39 (2): 262–8.
- Emerson, J. W., M. Seltzer, and D. Lin. 2009. "Assessing Judging Bias: An Example from the 2000 Olympic Games." *The American Statistician* 63 (2): 124–31.
- Findlay, L. C., and D. M. Ste-Marie. 2004. "A Reputation Bias in Figure Skating Judging." *Journal of Sport & Exercise Psychology* 26: 154–66.
- Flessas, K., D. Mylonas, G. Panagiotaropoulou, D. Tsopani, A. Korda, C. Siettos, A. D. Cagno, I. Evdokimidis, and N. Smyrnis. 2015. "Judging the Judges' Performance in Rhythmic Gymnastics." *Medicine & Science in Sports & Exercise* 47 (3): 640–8.
- Gymnastics Member Club (2008). *Diversity Study*. USA Gymnastics Member Club Online Newsletter. <https://usagym.org/pages/memclub/news/winter07/diversity.pdf>. Version Winter 2008. (accessed January 08, 2019).
- Heiniger, S., and H. Mercier. 2019a. *Judging the Judges: A General Framework for Evaluating the Performance of International Sports Judges*. ArXiv e-prints. arXiv: 1807.10055 [stat.AP]. Also Available at <https://arxiv.org/abs/1807.10055>.
- Heiniger, S., and H. Mercier. 2019b. *National Bias of International Gymnastics Judges during the 2013-2016 Olympic Cycle*. ArXiv e-prints. arXiv: 1807.10033 [stat.AP]. Also Available at <https://arxiv.org/abs/1807.10033>.
- Holm, S. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65–70.
- Kendall, M. G. 1938. "A New Measure of Rank Correlation." *Biometrika* 30 (1/2): 81–93.
- Kumar, R., and S. Vassilvitskii. 2010. "Generalized Distances Between Rankings." In *Proceedings of the 19th International Conference on World Wide Web. WWW'10*, 571–80. Raleigh. ISBN 978-1-60558-799-8. New York: Association for Computing Machinery.
- Landers, D. M. 1970. "A Review of Research on Gymnastic Judging." *Journal of Health, Physical Education, Recreation* 41 (7): 85–8.
- Leskošek, B., I. Čuk, J. Pajek, W. Forbes, and M. Bučar-Pajek. 2012. "Bias of Judging in Men's Artistic Gymnastics at the European Championship 2011." *Biology of Sport* 29 (2): 107–13.
- Mercier, H., and S. Heiniger. 2019. *Judging the Judges: Evaluating the Performance of International Gymnastics Judges*. ArXiv

- e-prints. arXiv: 1807.10021 [stat.AP]. Also Available at <https://arxiv.org/abs/1807.10021>.
- Myers, T. D., N. J. Balmer, A. M. Nevill, and Y. A. Nakeeb. 2006. "Evidence of Nationalistic Bias in MuayThai." *Journal of Sports Science and Medicine* 5 (CSSI): 21–7.
- Pajek, M. B., I. Čuk, J. Pajek, M. Kovač, and B. Leskošek. 2013. "Is the Quality of Judging in Women Artistic Gymnastics Equivalent at Major Competitions of Different Levels?" *Journal of Human Kinetics* 37 (1): 173–81.
- Pizzera, A. 2012. "Gymnastic Judges Benefit from Their Own Motor Experience as Gymnasts." *Research Quarterly for Exercise & Sport* 83 (4): 603–7.
- Pizzera, A., C. Möller, and H. Plessner. 2018. "Gaze Behavior of Gymnastics Judges: Where Do Experienced Judges and Gymnasts Look While Judging?" *Research Quarterly for Exercise & Sport*: 1–8. <https://doi.org/10.1080/02701367.2017.1412392>.
- Plessner, H. 1999. "Expectation Biases in Gymnastics Judging." *Journal of Sport & Exercise Psychology* 21: 131–44.
- Plessner, H., and E. Schallies. 2005. "Judging the Cross on Rings: A Matter of Achieving Shape Constancy." *Applied Cognitive Psychology* 19: 1145–56.
- Pope, D. G., J. Price, and J. Wolfers. 2018. "Awareness Reduces Racial Bias." *Management Science* 64 (11): 4988–95.
- Popović, R. 2000. "International Bias Detected in Judging Rhythmic Gymnastics Competition at Sydney-2000 Olympic Games." *Facta Universitatis – Series: Physical Education and Sport* 1 (7): 1–13.
- Price, J., and J. Wolfers. 2010. "Racial Discrimination Among NBA Referees." *Quarterly Journal of Economics* 125 (4): 1859–87.
- Sandberg, A. 2018. "Competing Identities: A Field Study of In-Group Bias Among Professional Evaluators." *The Economic Journal* 128 (613): 2131–59.
- Silva, M.-R. G., R. S. Rocha, P. Barata, and F. Saavedra. 2017. "Gender Inequalities in Portuguese Gymnasts between 2012 and 2016." *Science of Gymnastics Journal* 9 (2): 191–200.
- Souchon, N., G. Coulomb-Cabagno, A. Traclet, and O. Rascle. 2004. "Referees' Decision Making in Handball and Transgressive Behaviors: Influence of Stereotypes About Gender of Players?" *Sex Roles* 51 (7/8): 445–53.
- Spearman, C. 1904. "The Proof and Measurement of Association between Two Things." *American Journal of Psychology* 15 (1): 72–101.
- STAT Info. 2004. *Bulletin de la Mission statistique du ministère de la Jeunesse, des Sports et de la Vie Associative*. <http://www.sports.gouv.fr/IMG/archives/pdf/statinfo04-07.pdf>. Version November 2004. (accessed January 08, 2019).
- Ste-Marie, D. M. 1999. "Expert–Novice Differences in Gymnastic Judging: An Information-Processing Perspective." *Applied Cognitive Psychology* 13 (3): 269–81.
- Ste-Marie, D. M. 2000. "Expertise in Women's Gymnastic Judging: An Observational Approach." *Perceptual & Motor Skills* 90: 543–6.
- Zitzewitz, E. 2006. "Nationalism in Winter Sports Judging and its Lessons for Organizational Decision Making." *Journal of Economics and 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.
- Zwarg, L. F. 1935. "Judging and Evaluation of Competitive Apparatus or Gymnastic Exercises." *The Journal of Health and Physical Education* 6 (1): 23–49.