Judging the judges:
Evaluating the performance of international gymnastics judges

Hugues Mercier* Christopher Klahn†
February 18, 2017

Abstract

Judging a gymnastics routine is a noisy process, and the performance of judges varies widely. The International Federation of Gymnastics (FIG), in collaboration with Longines and the Université de Neuchâtel, is designing and implementing an improved statistical engine to analyze the performance of gymnastics judges during and after major competitions like the Olympic Games and the World Championships. The engine, called the Judge Evaluation program (JEP), has three objectives: (1) provide constructive feedback to judges, executive committees and national federations; (2) assign the best judges to the most important competitions; and (3) detect bias and outright cheating.

In this article, using data from international competitions held during the 2013-2016 Olympic cycle, we first develop a marking score evaluating the accuracy of the marks given by judges. We then study ranking scores assessing to what extent judges rate gymnasts in the correct order, and explain why we ultimately chose not to implement them. We study outlier detection to pinpoint athletes that were poorly evaluated by judges. Finally, we discuss interesting observations and discoveries that led to recommendations to the FIG.

1 Introduction

Gymnastic judges and judges from similar sports are susceptible to well-studied and often unconscious biases. For instance, Ansorge and Scheer [1] detected a small international bias of gymnastics judges at the 1984 Olympic Games: judges tend to give better marks to athletes from their home country while penalizing close competitors from other countries. Bruine de Bruin [2] observed serial position effects: competitors performing at the end of a competition get better marks than competitors performing at the beginning. Findley and Ste-Marie [3] found a reputation bias in figure skating: judges overestimate the performance of athletes with a good reputation. Boen et al [4] observed a conformity bias in gymnastic judging: open feedback causes judges to adapt their marks to those of the other judges of the panel. These biases are difficult to counter in practice, but fortunately they are reasonably small.

*Université de Neuchâtel, Switzerland, email: hugues.mercier@unine.ch
†Université de Neuchâtel, Switzerland, email: christopher.klahn@unine.ch
Judging is much more about skill and training than bias: it is challenging to evaluate every single aspect of the complex movements that are part of a gymnastics routine, and unsurprisingly nearly all international judges are former gymnasts. Ste-Marie and Flessas et al. compared the ability of judges to detect execution mistakes in gymnastics routines. Novice judges consult their scoring sheet much more often than experienced international judges, thus missing execution errors. Furthermore, international judges are better to detect errors in their peripheral vision.

Even among well-trained judges at the international level, there are significant performance differences. Some judges are simply better than others. For this reason, the International Federation of Gymnastics (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, the current architecture of JEP is outdated. It is hard to maintain, difficult to use by non-expert users, and its conception makes it prone to manipulation errors. More importantly, design errors and undesirable behavior are hard to detect, and it is sometimes adjusted by parties with conflicting viewpoints. Hence, it is not always clear that it evaluates what it ought to evaluate.

1.1 Our contributions

In collaboration with the FIG and Longines, we are currently redesigning and rewriting JEP. The new version, JEP2017+, is a full software stack that handles all the interactions with the databases, and includes a statistical engine as well as a user-friendly front-end to generate statistics, recommendations and judging reports. The main objective of JEP2017+ is to assess, as objectively as possible, the performance of international gymnastics judges. This assessment will provide feedback to judges, executive committees and national federations. It will be helpful, for instance, to evaluate how to train and accredit judges, and to propose corrective measures for judges performing below expectations. It will be used to reward the best judges by selecting them to the most important competitions like the Olympic Games. It can also provide hints about inconsistencies and confusing items in the Codes of Points detailing how to evaluate each apparatus. In uncommon but important circumstances, JEP2017+ can also uncover biased and cheating judges.

This article focuses on the core statistical engine of JEP2017+ and describes how it evaluates judges. Our main objective is to put judge evaluations on a stronger mathematical footing using simple yet rigorous tools. We want the software, once properly trained, to be usable by Longines and the FIG during an entire 4-year Olympic cycle without input from a mathematician or a statistician.

The main tool we develop is a marking score evaluating how accurate the marks given by a judge are. We design the marking score such that it is unbiased with the apparatus/discipline under evaluation, and unbiased with respect to the skill level of the gymnasts. In other words, the main difficulty we overcome is as follows: a parallel bars judge giving 5.3 to an athlete deserving 5.0 must be evaluated more generously than a vault judge giving 9.9 to an athlete deserving 9.6, but how much more? Our analysis includes the five gymnastics disciplines: Artistic Gymnastics, Acrobatic Gymnastics, Aerobic Gymnastics, Rhythmic Gymnastics, and Trampoline. For each of these disciplines we analyzed the marks given by the execution judges at international competitions.

1. www.longines.com
2. The Codes of Points, their appendices and other documents related to rules for all the gymnastics disciplines are publicly available on the FIG website. Consult, for example, http://www.fig-gymnastics.com/site/rules/disciplines/art for Artistic Gymnastics.
Table 1: Composition of the execution panels per discipline. E = Execution judges; R = Reference judges.

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Number of judges (event type 1)</th>
<th>Number of judges (event type 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artistic Gymnastics</td>
<td>5 E + 2 R</td>
<td>4 E</td>
</tr>
<tr>
<td>Acrobatic Gymnastics</td>
<td>4 E + 2 R</td>
<td></td>
</tr>
<tr>
<td>Aerobic Gymnastics</td>
<td>4 E + 2 R</td>
<td></td>
</tr>
<tr>
<td>Rhythmic Gymnastics</td>
<td>5 E + 2 R</td>
<td>5 E</td>
</tr>
<tr>
<td>Trampoline</td>
<td>5 E</td>
<td></td>
</tr>
</tbody>
</table>

held during the 2013-2016 Olympic cycle. This article mostly focuses on Artistic Gymnastics, using 65634 execution judging marks. The approach and techniques for the other disciplines are similar.

Besides allowing us to distinguish between good and bad judges, we also use the marking score as the basic tool to detect outlier evaluations by judges. We design JEP2017+ such that an excellent judge on average will have a lower outlier detection threshold than an erratic judge.

The current iteration of JEP uses a ranking score to evaluate the ranking of the athletes resulting from a judge’s marks. We analyzed different metrics to compare distances between rankings such as the generalized version of Kendall’s τ distance \([9]\). No ranking score achieves the objectives of the FIG, thus we decided not to implement any in JEP2017+.

The remaining of this article is organized as follows. We derive and discuss the marking score in Section 2. In Section 3 we use the marking score to detect outliers. Section 4 discusses ranking scores and why we ultimately left them aside. We present interesting observations and discoveries in Section 5 and conclude in Section 6.

## 2 Marking score

We now derive a marking score to evaluate the performance of gymnastics judges. The marking score must have the following properties. First, it must not depend on the skill level of the athletes evaluated: a judge should not be penalized nor advantaged if he judges an Olympic final with the world’s best 8 gymnasts as opposed to a preliminary round with 200 athletes. Second, it must allow to compare judges across apparatus, disciplines, and competitions.

The execution of a gymnastics routine is evaluated by a panel of judges. The composition of the execution panel, which depends on the discipline and the type of event, is summarized in Table 1. This panel usually includes execution and reference judges, except for some World Cup events and continental championships that only include execution judges. The execution and reference judges have different power and are selected differently, but they all judge the execution of the routines using the same criteria.

After the completion of a gymnastics routine, each judge on the panel evaluates the performance by giving it a score between 0 and 10. The execution evaluation of a gymnastics routine is based on deductions precisely defined in the Code of Points of each apparatus. The score of each judge

---

\(^3\)The Codes of Points, their appendices and other documents related to rules for all the gymnastics disciplines are publicly available on the FIG website. Consult, for example, http://www.fig-gymnastics.com/site/rules/disciplines/art for Artistic Gymnastics.
can thus be compared to a theoretical control score corresponding to the “true” performance of the athlete, provided by video review post-competition. JEP2017+ uses the median mark of each athlete to provide a live analysis during each competition. The marking score of a judge is based on three parameters:

1. The control scores of the athletes
2. The scores given by the judge
3. The apparatus / discipline

Table 2 summarizes the notation used in this section. Let \( s_{a,j} \) be the score of Athlete \( a \) given by Judge \( j \). We base our analysis on 65634 such (athlete, judge) pairs from Artistic Gymnastics international competitions held during the 2013–2016 Olympic cycle. Figure 1 shows the distribution of the random variable \( D_a \triangleq s_{a,j} - c_a \) corresponding to the difference between the mark given by a judge and the control score of the athlete. The good news is that \( D_a \) is unbiased, thus the judges are too severe as often as they are too generous. This unbiasedness remains true with the best athletes, which was initially surprising since a judge cannot give 10.1 to an athlete deserving 9.9 but is explained by the fact that judges are much more accurate for the best athletes. This is bad news because \( D_a \) is highly heteroscedastic, and using \( D_a \) underweights errors made for the best athletes. We initially considered using \( \frac{D_a}{10-c_a} \), which scales the difference by the total deduction, but this swings the pendulum too far and overweights the errors made for the best athletes. This is not as bad since the performance of judges for the athletes competing for medals is the most important, but the resulting marking score is biased with the control score, which does not allow comparisons across competitions and disciplines.

Figures 2 and 3 respectively show the variance and the standard deviation of \( D_a \) as a function of the control score \( c_a \). The variance decreases almost linearly with the control score, except for the best athletes for which it does not converge to zero. The curve in Figure 3 is an exponential weighted least-squares regression of the data. By inspection this is an excellent fit; the outliers correspond to the rare athletes that aborted or completely missed their routine. More formally, the
Figure 1: Distribution of the difference $D_a$ between the judging score and the control score of an athlete.

weighted standard error of the regression, $s \approx 0.001$, is only 1% of the smallest deduction allowed by a judge. We use this exponential equation for our estimator of the standard deviation $\sigma_d(c_a)$.

More generally, the estimator for $\sigma_d(c_a)$ depends on the discipline (or apparatus) $d$ under evaluation and the control score $c_a$ of the athlete, and is given by

$$\sigma_d(c_a) \triangleq \max(\alpha_d + \beta_d e^{\gamma d c_a}, 0.05).$$  \hspace{1cm} (1)

We do not show the results but nevertheless mention that all disciplines and apparatus we analyzed have accurate regressions \(^4\) (the worst weighted standard error of the regressions is $s \approx 0.007$). For some apparatus, there is no athlete with a mark close to 10, and the best fitted curves go to zero before 10. We use $\max(\cdot, 0.05)$ as a fail-safe mechanism should some athletes get much larger marks than expected in the future.

The marking score of Athlete $a$ by Judge $j$ is

$$m_{a,j} \triangleq \frac{s_{a,j} - c_a}{\sigma_d(c_a)}.$$

This expresses the difference as a function of the standard deviation for a specific discipline and control score. The overall marking score for Judge $j$ is given by

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

\(^4\)Some of the regressions are linear instead of exponential.
The overall marking score is the mean squared deviation (or mean squared error). The mean squared error weighs outliers heavily, which is desirable for evaluating judges.

Figure 4 shows the boxplots of the marking scores for all the judges for each apparatus using the regression from Figure 3 (the acronyms are defined in Table 3). The first observation is that there are significant differences between the best and the worst judges. This is not surprising, however the FIG now has a much more precise and objective tool to assess judges. The second observation is that there are significant differences between the apparatus themselves, some of them surprising for the FIG. For instance, pommel horse appears intrinsically much more difficult to judge than vault and floor. We made similar observations for other disciplines. In Rhythmic Gymnastics, group performances are more difficult to judge than individual ones. In trampoline, tumbling is much more difficult to judge than individual trampoline, which in turn is much more difficult to judge than double mini-trampoline.

The differences between apparatus make it challenging for the FIG to qualitatively assess how good the judges are and to clearly convey this information to the judges and national federations. 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 standard deviation \( \sigma_d(c_a) \) for each apparatus (instead of grouping them together) and used the resulting regressions to recalculate the marking scores. The results, presented in Figure 5, now show a good uniformity and make it simple to compare judges from different apparatus to 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 allows to define a single set of quantitative to qualitative thresholds across all the gymnastics disciplines.

We conclude this section by mentioning that the drawback of the marking score is that judges are compared with each other and not based on their objective performance. An apparatus with only outstanding judges will trivially have half of them with a marking score below average\(^5\) but from discussions with the FIG in practice no apparatus has the luxury of having only outstanding judges. We therefore proposed qualitative thresholds based on the fact that most judges are good, and a reward-based approach for the very best ones.

\(^5\)The same is true of an apparatus with only atrocious judges.
Figure 2: Variance of judge difference $D_a$ versus control score $c_a$.

Figure 3: Standard deviation of judge difference $D_a$ versus control score $c_a$. 

\[ \sigma(c_a) = 0.5738 - 0.06131 \cdot e^{0.2157c_a} \quad s = 0.001 \]
Figure 4: Distribution of the overall marking scores per apparatus using one overall formula. The acronyms are explained in Table 3 and the numbers between brackets are the number of judges per apparatus in the data.

Figure 5: Distribution of the overall marking scores using an individual formula per apparatus. The acronyms are explained in Table 3 and the numbers between brackets are the number of judges per apparatus in the data. The red dots are the deciles.
Table 3: The artistic apparatus and their acronyms.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Apparatus</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>Balance beam</td>
</tr>
<tr>
<td>FX</td>
<td>Floor exercise</td>
</tr>
<tr>
<td>HB</td>
<td>Horizontal bar</td>
</tr>
<tr>
<td>PB</td>
<td>Parallel bars</td>
</tr>
<tr>
<td>PH</td>
<td>Pommel horse</td>
</tr>
<tr>
<td>SR</td>
<td>Still rings</td>
</tr>
<tr>
<td>UB</td>
<td>Uneven bars</td>
</tr>
<tr>
<td>VT</td>
<td>Vault table</td>
</tr>
</tbody>
</table>

3 Outlier detection

The current iteration of JEP includes a tool that flags judges suspected of bias in favor of athletes of their own country. These suspicious evaluations are signaled to the FIG and are usually followed by a video review of the relevant routines post-competition. In JEP2017+, we use the marking score to detect (athlete, judge) outlier pairs, that is, judging marks that are unreasonably high or unreasonably low. Figure 6 like Figure 1 shows the differences $D_a$ between judging scores and control scores for Artistic Gymnastics. Differences more than two standard deviations ($2 \cdot \sigma_d(c_a)$) away from the control score are marked in red. The problem with this approach is that a bad judge has a lot of outliers, and a good judge none. This is not what the FIG wants, because an erratic judge can be unbiased and a precise judge can be biased.

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

$$|D_a| > \max(2 \cdot \sigma(c_a) \cdot M_j, 0.1).$$

We use $\max(\cdot, 0.1)$ to ensure that a difference of 0.1 is never an outlier. The results are shown in Figure 7. Eq. 4 flags $\approx 5\%$ of the marks, which is slightly more than what would be expected for a normal distribution. 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. Furthermore, other flags like outliers from the same country, and large outliers that are three standard deviations away from the control score, can be added easily. It also confirms that blatant cheating by judges increasing or decreasing marks by large values is rare.

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 it has. The FIG can have a close look at outliers, even for precise judges with good marking scores.
Figure 6: Distribution of the difference $D_a$ between the judging score and the control score. Dots in red are more than two standard deviations ($2 \cdot \sigma_d(c_a)$) away from the control score.

Figure 7: Distribution of the difference $D_a$ between the judging score and the control score. Dots in red are more than $2 \cdot \sigma_d(c_a) \cdot M_j$ away from the control score.
4 Ranking score

In gymnastics, the ranking of the gymnasts is determined by their scores, which are themselves calculated from the judges’ marks. The current iteration of JEP uses a ranking score to evaluate to what extent judges rank the best athletes in the right order. In a vacuum this makes sense: the FIG wants to select the right athletes for the finals, and award the medals to the very best athletes in the correct order. In practice providing an objective assessment of the judges based on the order in which they rank the best athletes is problematic, and we ultimately recommended to the FIG to move away from this approach.

Definition 4.1 Let $X = \{a, b, c, \ldots, n\}$ be a set of $n$ athletes. A ranking on $X$ is a sequence $r = a_1 a_2 a_3 \ldots a_n$, $a_i \neq a_j \ \forall i, j \in \{1, \ldots, n\}$ of all the elements of $X$ that defines a weak order $\succeq$ on $X$. Alternatively, a ranking can be noted as $r = (r_a, r_b, r_c, \ldots)$, where $r_a$ is the rank of Athlete $a$, $r_b$ is the rank of Athlete $b$, and so on.

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 $\tau$ distance and Spearman’s footrule, both of which are within a constant fraction of each other [10]. In recent years, Kumar and Vassilvitskii [9] 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 4.2 Let $r$ be a ranking of $n$ competitors. Let $w = (w_1, \ldots, w_n)$ be a vector of element weights, $\delta = (\delta_1, \ldots, \delta_n)$ a vector of position swap costs where $\delta_i$ is the cost of swapping elements at positions $i-1$ and $i$. Let $p_1 = 1$ and for $1 < i \leq n$ define $p_i = \sum_{j=1}^{i-1} \delta_j$. We define the mean cost of interchanging positions $i$ and $r_i$ by $\bar{p}(i) = \frac{p_i - p_{i-1}}{r_i}$. Finally, let $D : \{1, \ldots, n\} \times \{1, \ldots, n\}$ be a non-empty metric and interpret $D(i, j) = D_{ij}$ as the cost of swapping element $i$ and $j$. The generalized Kendall’s $\tau$ distance is

$$K^\tau = K^\tau_{w,\delta,D}(r) = \sum_{s>t} w_s w_t \bar{p}_s \bar{p}_t D_{st}[r_s < r_t]. \tag{5}$$

Note that $K^\tau$ is the distance between $r$ and the identity ranking $id = (1, 2, 3, \ldots)$. 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 $r^2$.

In the context of our gymnastics setting these generalizations are natural: swapping the gold and silver medalists should be evaluated more harshly than inverting the ninth and tenth best athletes, 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.

To test the relevance of ranking scores as a measurement of the judges’ performances, we ran several simulations to compare them to our marking score. As an example for this article, we use data from the qualification phase of the men’s floor exercise at the 2013 European Championships in Moscow. We first calculate the control scores $c_1, c_2, \ldots, c_8$ of the best eight athletes. We then simulate the performance of 1000 average judges $j \in \{1, 2, \ldots, 1000\}$ by randomly creating, for each
Table 4: Parameters of the ranking scores for our two simulations.

<table>
<thead>
<tr>
<th>Parameter set</th>
<th>( w_i )</th>
<th>( \delta_p )</th>
<th>( D_{aj} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>( \frac{1}{p} )</td>
<td>(</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 8: Ranking score vs marking score for 1000 synthetic average judges and the first set of parameters from Table 4 for the ranking score.

of them, eight marks \( s_{1,j}, s_{2,j}, \ldots, s_{8,j} \) for the eight athletes using a normal distribution with mean \( c_a \) and standard deviation \( \sigma_d(c_a) \) for \( a \in \{1, 2, \ldots, 8\} \). We then calculate, for each judge, the marking score and a ranking score. This set of 1000 synthetic average judges is simulated twice, each time with a ranking score based on Eq. (5) but with a different set of parameters. Table 4 shows the two sets of parameters used.

Figures 8 and 9 show the ranking score with respect to the marking score of the 1000 judges for the two parameter sets. The figures illustrate that the correlation between the ranking score and the marking score varies widely depending on the chosen parameters. In Figure 8, the position swap costs and the element swap costs vary. The position swap costs increase the importance of having the correct order as we move towards the gold medalist. The element swap costs improve the evaluations as the marks get closer to each other, which brings the ranking score closer to the marking score. In fact, the high correlation between the marking score and the ranking score means that both largely measure the same thing. The parameters used in Figure 9 are those of the original version of Kendall’s \( \tau \) distance. In this instance there is almost no correlation between the marking score and the ranking score, thus we penalize good but unlucky judges that make mistakes at the wrong place, and reward erratic but lucky judges.
Figure 9: Ranking score vs marking score for 1000 synthetic average judges and the second set of parameters from Table 4 for the ranking score.

The marking score already achieves our objectives. It is based on the theoretical performances of the athletes and reflects bias and cheating, as this involves changing the marks up or down for some of the athletes. Furthermore, a theoretical judge who ranks all the athletes in the correct order but is either always too generous or too strict is not a good judge because he/she does not apply the Codes of Points properly. From these observations, the ranking score is either redundant with the marking score, or too uncorrelated to be of any practical value. We thus decided to build JEP2017+ without a ranking score.

5 Observations and discoveries

During the course of this work we made interesting and sometimes surprising observations and discoveries. This section summarizes some of them.

5.1 Trampoline

Trampoline was the most puzzling discipline to tackle. Figure 10 shows the standard deviation $\sigma_d(c_a)$ of the judge distance $D_a$ as a function of the control score $c_a$. This strange behavior is due to the considerable number of athletes that aborted their routine before completing all their jumps, for instance by losing balance and landing a jump outside the center of the trampoline.
Figure 10: Standard deviation of judge difference $D_a$ versus control score $c_a$ for Trampoline.

Figure 11: Standard deviation of judge difference $D_a$ versus control score $c_a$ for Trampoline. The rings indicate aborted routines. Data from synchronized trampoline is removed.
We solved the problem by fitting the curves based on the completed routines. The results are shown in Figure 11 with aborted routines represented using rings instead of filled circles. Despite the bad fit for low control scores, the weighted standard error of the regression is excellent ($s \approx 0.002$). We note that Figure 11 excludes the data from synchronized trampoline, for which the judge evaluations were bad. The problem is that synchronized trampoline panels are partitioned in two halves, each evaluating one of the athletes, and we did not have the control score of each athlete. When calculating the marking score for trampoline judges, the marks of athletes that did not complete their exercise may or may not be taken into account. The estimator generously evaluates judges when athletes do not complete their routine, which results in a slightly improved overall marking score.

5.2 Reference judges

In Artistic Gymnastics’ most important competitions, the execution score is first calculated with the trimmed mean of the five execution panel judges. This score is then compared to the reference score consisting of the mean score of the two reference judges. If the gap between the execution panel score and the reference score exceeds a predefined tolerance, the final score of the athlete is the mean of the execution panel and reference scores. This makes reference judges dangerously powerful.

![Figure 12: Distribution of marking scores for Artistic Gymnastics execution panel and reference judges.](image)

At each competition, execution judges are randomly selected from a set of accredited judges submitted by the 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 incompetent or corrupt. 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 12. 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 worse. 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, we strongly recommended that the FIG does not grant more power to reference judges. They are not better, and the small size of the reference panels further increases the likelihood that the errors they make have greater consequences.

5.3 Battle of the sexes

In Artistic Gymnastics, men apparatus are evaluated by men judges and women apparatus are evaluated by women judges. Figure 4, besides showing the differences between apparatus, also shows that the marking scores for women apparatus are lower than those of men apparatus. Figure 13 formalizes this observation by directly comparing the marking scores of men and women judges. 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 $10^{-15}$ at the 95% confidence level, thus leading to the rejection of the null-hypothesis.

![Figure 13: Distribution of marking scores for Artistic Gymnastics split by sex.](image)

Two hypotheses can explain this difference: either women gymnasts are easier to evaluate than men, or women judges are better than men. The FIG originally thought that the women Codes
of Points were more difficult to apply, which does not seem to be the case from the analysis. Furthermore, we can observe from Figure 4 that judges for women’s vault table are better than judges for men’s vault table, and that judges for women’s floor exercise are better than judges for men’s floor exercise. This can hardly be explained by the differences in the Codes of Points or the intrinsic nature of the exercises.

We suspect that the difference is explained by two factors. First, the formation and accreditation process is different for men and women judges. Men, who must judge six apparatus, receive less training than women, who only must judge four. Some men judges also have a (maybe unjustified) reputation of laissez-faire, which contrasts with the precision required from women judges. Second, the pool of female gymnasts is larger than the pool of male gymnasts, which increases the probability of having more good female judges at the top of the pyramid. We are currently investigating this further by comparing the marking scores for men and women judges in other disciplines, such as Trampoline, where judging panels are mixed.

6 Conclusion

We put the evaluation of international gymnastics judges on a strong mathematical footing using robust yet simple tools. This has led to a better assessment of current judges, and will improve judging in the future. Our main contribution is a marking score that evaluates the accuracy of the marks given by judges. The marking score can be used across disciplines, apparatus and competitions. Its calculation is based on the standard deviation, estimated from prior data, of the difference between the marks given by judges and the theoretical control scores of the athletes. It can and should be calibrated at the beginning of every Olympic cycle with data from the previous cycle. The marking score is the central piece of our outlier detection technique highlighting evaluations far above or below what is expected from each judge.

For future work we would like to calculate marking scores for artistic judges in Acrobatic and Aerobic gymnastics, and compare them to the execution judges. We also want to study long-term bias and collusion of judges in favor of their own athletes and against athletes directly competing with them, which is the Holy Grail of the FIG. Finally, we would like to adapt the techniques developed for gymnastics to other sports like figure skating, diving and snowboarding, where performances are evaluated on a finite numerical scale based on a set of rules and deductions.

Acknowledgements

This work is the result of fruitful interactions and discussions with the other project partners. We would like to thank André Gueisbuhler, Nicolas Buompane, Les Fairbrother, Sylvie Martinet and Rui Vinagre from the FIG, Benoit Cosandier, Jose Morato, Christophe Pittet, Pascal Rossier and Fabien Voumard from Longines, and Pascal Felber, Rolf Klappert and Claudia Nash from the Université de Neuchâtel.

6 A 2007 Survey from USA Gymnastics reported four times more female gymnasts than male gymnasts in the USA. Consult https://usagym.org/pages/memclub/news/winter07/diversity.pdf for details.
References


