AN EFFICIENT METHOD TO EVALUATE AND ENHANCE SPORT JUDGES' PERFORMANCE DURING COMPETITION: A CASE STUDY IN ACROBATIC GYMNASTICS

Juan Antonio León-Prados[1] & Carmen Ramos[2]

[1] Physical Performance & Sport Research Center. Pablo de Olavide University, Seville, Spain.
[2] Statistic and Operational Research Department. Social and Communication Sciences Faculty, Cadiz University, Cadiz, Spain

Original article *DOI: 10.52165/sgj.16.3.387-400*

Abstract

Who judges the gymnastics judges? How do we measure their accuracy and concordance? How do we know if the judging process is fair? The superior jury has this responsibility. However, they normally lack time to provide effective feedback during competitions.
Using macro functions to process statistical and mathematical statements, we designed and validated an automated Excel-based tool called the Automatic Acrobatic Gymnastics Judges Individual Report Tool to evaluate judges' performances quickly and easily during competition, automatically creating and exporting individual reports showing each judge's accuracy and concordance performance on a daily basis, rather than after the competition is ended.
We present empirical data for 76 experienced international judges evaluating acrobatic gymnastics routines in four major official events. A total of 1240 individual reports were analyzed and sent confidentially to the judges during the competition, and 952 were analyzed to evaluate whether this feedback was effective in improving judges' performance during the competition.
The tool provides efficient and easily understood evidence-based feedback on acrobatic gymnastics judges' performance during competition, quickly and automatically creating, analyzing and sending individualized information to judges, thus helping with specific Technical Committee scoring control tasks during competitions. We suggest that judges' performances remain high or are enhanced after receiving daily evaluation during major competition events.

Keywords: *Excel-based tool, Acrobatic Gymnastic, Judges, Evaluation, ACROAJIR®*

INTRODUCTION

Judges are often criticized and sometimes undervalued in sports. In subjectively assessed sports, judges may collude, giving higher scores to their own athletes and lower scores to others. To prevent this, federations implement various strategies, such as automatically eliminating the highest and lowest scores or involving a referee judge (Gambarelli et al., 2012). Gymnastics judges must observe and assess the quality of performances, often processing large amounts of information (Dosseville et al., 2014). Their scores can be influenced by factors such as their viewing position (Dallas et al., 2011; Plessner & Schallies, 2005), serial position bias (Plessner & Schallies, 2005; Fasold et al., 2012; De Bruin, 2005), conformity bias

(Auweele et al., 2004; Boen et al., 2006, 2008, 2013), or the performance of the preceding gymnast (Damisch et al., 2006; Kramer, 2017).

Knowledge, experience, and psychological factors such as attention, emotion recognition, and possible interventions may reduce judges' biases or stress, helping to avoid scoring mistakes (Flessas et al., 2015; Ste-Marie, 2000; Van Bokhorst et al., 2016). These factors can influence the outcome in sports where scoring and ranking depend on subjective evaluations. Gymnastics judges' performance can vary widely, but who judges the judges? Typically, other judges form a superior jury, whose evaluations occur post-competition. Some research has examined judges' overall performance in terms of reliability or concordance after events (Bučar et al., 2012; León-Prados & Jemni, 2022; Leskošek et al., 2018; Mercier & Heiniger, 2018; Premelč et al., 2019); however, none has focused on judges' work during competitions on a day-to-day basis.

Such evaluation requires careful monitoring of judges' accuracy and concordance, which demands significant time and effort at the end of each competition day. In real events, this can be challenging, as judges are often fatigued, and statistical and mathematical expertise is needed for these evaluations.

The FIG, in collaboration with Longines and the Université de Neuchâtel, designed and implemented the Judge Evaluation Program (JEP) for five gymnastics disciplines: Artistic, Acrobatic, Aerobic, Rhythmic Gymnastics, and Trampoline. This program analyzed the marks given by execution judges at international competitions during the 2013–2016 Olympic cycle (Heiniger & Mercier, 2021; Heiniger & Mercier, 2018; Mercier & Heiniger, 2018; Mercier & Klahn, 2017). The authors claimed that the JEP helps to ensure judges' objectivity during gymnastics competitions, allowing for post-competition analysis and an overall evaluation of judges by the respective Technical Committees (TCs). This post-competition control can be applied in competitions where the use of IRCOS (Instant Replay & Control System) is mandatory (FIG, 2020). Judges' scores must demonstrate accuracy, precision, consistency, and the absence of bias. The JEP evaluates gymnastics judges' performance compared to their peers, distinguishing between erratic and precise judges and detecting potential cheating or unintentional misjudging.

Since its inception in 2006, the JEP has evolved iteratively, although earlier versions were criticized for using unsound and inaccurate mathematical tools that didn't always evaluate what was intended. However, a new core statistical engine introduced in the 2013–2016 Olympic cycle provided more reliable feedback to judges and executive committees (Mercier & Heiniger, 2018). The FIG typically derives control scores using external judging panels and post-competition video reviews. This post-competition control establishes expert scores (considered "true scores") against which judges' scores are compared, ensuring evaluation on the fairest possible basis. Expert scores are provided by TC members, who individually assess each exercise (FIG, 2015).

However, previous studies have not clarified whether judges received their individual results after competitions or if they were given specific feedback on their performance during each session within competition days. The FIG has encouraged continental committees and national federations to adopt a similar system for their own events, which inspired our research (FIG, 2020).

How could we obtain this type of information about judges' performance in Acrobatic Gymnastics (ACRO) during real competitions? Could rapid daily feedback improve judges' performance and lead to fairer, more precise judging? Currently, in the absence of more objective feedback, judges rely on the only available in-competition feedback—the final trimmed mean execution and artistic score displayed on the scoreboard. The scores from the in-competition control panel remain unknown to the judges, even after the competition ends.

This study developed and implemented the Automatic Acrobatic Gymnastics Judges Individual Report Tool (ACROAJIR®) (Leon-Prados & Rosales, 2019) as a pedagogical tool to evaluate ACRO judges' performance in real-time, providing objective feedback on their work and potential judging impacts.

METHODS

This study had two goals: a) The ACROAJIR® design; and b) its practical application with judges in real events.

The ACROAJIR® Design

All the official Execution (E) and Artistic (A) scores were collected confidentially and were provided by SmartScoring, the European Gymnastics exclusive results service provider (Bakú, Azerbaijan).

Control scores validity. Looking for the true score

In practice, true performance level is unknown and we must work with approximations. In our study, we assumed that the highest category judges in the Superior Jury who provide the Technical or Artistic Control Scores (E/A C-Score), represented the “truer score” when they judged a competitive routine, compared to lower-level judges.

We proposed a model with two key considerations: 1) the Superior Jury’s scores are considered more representative of the performance, and individual judges’ deviations from the overall judging panel define their performance level; and 2) the model is based on the pre-defined tolerances established by the FIG for judges’ reference (FIG, 2017). If the scores for a routine fall outside this pre-defined deviation among the control judges, they must re-judge the routine using video recordings. This process could yield scores closer to the "true score" and provide better feedback on judges’ performance. Control scores can only be adjusted if the deviation between scores exceeds the allowed tolerance.

The "true score" is determined by the E/A C-Score, averaging three E/A C-scores: two expert judges’ scores from the Superior Jury, plus the Chair Judge’s score from the judging panel. All three expert scores for each E- and A-C score must fall within the allowed deviation. To ensure this, we used the coefficient of variation (CV), where CV = (Standard Deviation / Mean) * 100. The CV takes into account the weighting variable, as judges are generally more accurate when assessing higher-quality performances than lower-quality ones (Mercier & Heiniger, 2018). Since judging variation increases as scores decrease, the allowed inter-judge deviation thresholds increase with the number of deductions.

This means that the same absolute deviation between judges results in a higher dispersion when lower penalties are applied. In our model, when the average total deductions are less than 1 (resulting in a score of 9 or higher), a higher CV doesn’t necessarily indicate high variability, and a more accurate measure of score variability can be obtained from the classification rate, based on total deductions from a maximum of 10 points. We established different acceptable CVs for each 0.5 deduction from 10 points, all within the allowed deviation for each score range (Table 1).

In the ACROAJIR® Excel tool, a "control scores validity macro" was implemented to process all mathematical calculations and quickly detect significant differences between control judges’ scores. When the difference between control judges exceeds a specific threshold, a video review becomes compulsory to redefine the true score accurately.

Table 1.

Examples of cases of control scores allowed (case A) and not allowed (case B), with the least differences between them, to check the acceptability of the control score as the "true score" for each range of scores. The grey boxes provide an example of a non-allowed control judge score, according to the allowed deviations for each range of scores. The same criteria could be applied to artistic scores.

Routine range scores	**10.0 to 9.5**		**9.499 to 9.0**		**8.999 to 8.5**		**8.499 to 8.0**		**7.999 to 7.5**		**7.499 to 6.5**	
Maximum inter-score deviation allowed/range score	0.1		0.2		0.3		0.4		0.5		0.6	
Case examples	Case A	Case B	Case A	Case B	Case A	Case B	Case A	Case B	Case A	Case B	Case A	Case B
SJ-E1 score	9.8	9.8	9.3	9.3	8.9	8.9	8.5	8.5	7.9	7.9	7.4	7.4
SJ-E2 score	9.8	9.8	9.3	9.3	8.9	8.9	8.5	8.5	7.9	7.9	7.4	7.4
CJP-E3 score	9.7	9.6	9.1	9.0	8.6	8.5	8.1	8.0	7.4	7.3	6.9	6.8
Average E control score	9.767	9.733	9.233	9.200	8.800	8.767	8.367	8.333	7.733	7.700	7.233	7.200
Maximum inter-score deviation	0.1	0.2	0.2	0.3	0.3	0.4	0.4	0.5	0.5	0.6	0.5	0.6
Total SJ-E1penalties	0.2	0.2	0.7	0.7	1.1	1.1	1.5	1.5	2.1	2.1	2.6	2.6
Total SJ-E2 penalties	0.2	0.2	0.7	0.7	1.1	1.1	1.5	1.5	2.1	2.1	2.6	2.6
Total CJP-E3 penalties	0.3	0.4	0.9	1.0	1.4	1.5	1.9	2	2.6	2.7	3.1	3.2
Average control E penalty	0.233	0.267	0.767	0.800	1.200	1.233	1.633	1.667	2.267	2.300	2.767	2.800
relative change with only 0.1 point differences according to previous Case A (see shaded scores)		12.5%		4.2%		2.7%		2.0%		1.4%		1.2%
Penalties CV (%)	24.7	43.3	15.1	21.7	14.4	18.7	14.1	17.3	12.7	15.1	10.4	12.4
Maximum CV allowed for the Inter-judges' deductions for each range (%)	25		16.5		15		14.5		13.5		12.5	
Action required with regard to scores		Check		Check		Check		Check		Check		Check

Each routine was judged on its execution (E) and artistic merit (A), evaluated by a randomized pool of judges. Accuracy was measured as the deviation of a judge's E- and A-scores from the respective E- and A-control scores. Bias (integrity) was assessed by examining the rankings assigned by a judge for the exercises in a single round and across the entire competition. Consistency was evaluated by identifying unusual changes in the standard of marks given for the exercises (FIG, 2020). Paired panel and control scores were used to assess score accuracy (quantitative) and association concordance (ranking), for quantitative and qualitative evaluation, respectively.Lin's Concordance Correlation Coefficient (LCCC) was used to measure the accuracy or concordance between each judge's score (Y) and the "true score" provided by the Control score (X) to quantify the agreement between these two measures) for the same gymnastic routine (Akoglu, 2018; Lin, 1989; McBride, 2005). The LCCC formula was as follows:

$$\rho_c = \frac{2s_{xy}}{S_x^2 + S_y^2 + (\bar{x} - \bar{y})^2}$$

where S_{xy} is the covariance, S^2 is the variance and $\bar{x}$ and $\bar{y}$ are the means for x and y raters,
$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})$

Strength-of-agreement criteria for Lin's concordance correlations coefficient were proposed as follows: <0.99 Almost Perfect, 0.95 to 0.99 Substantial, 0.90 to 0.95 Moderate and <0.90 Poor (McBride, 2005). However, for real competition, an acceptable range of deviation between judges' scores is defined in Table 2.

This difference varies depending on the level of the competitive routine and is determined by the number of penalties awarded for technical and artistic errors. We assumed that the interpretation of correlation coefficients varies significantly across research areas. In gymnastics evaluation, due to potential inter-judge variability, particularly when penalties are greater, we proposed an interpretation closer to Altman's, suggesting that the strength-of-agreement criteria for Lin's concordance should be aligned with other correlation coefficients, such as Pearson's, where < 0.2 is considered poor and > 0.8 is excellent (Akoglu, 2018).

For ACROAJIR®'s assessment of acrobatic gymnastics judges' performance, we defined Lin's concordance qualitative ranking criteria as: < 0.95 Excellent; 0.8 to 0.9499 Very Good; 0.7 to 0.7999 Good; 0.6 to 0.6999 Satisfactory; 0.5 to 0.5999 Poor; and less than 0.5 Very Poor. Additionally, we needed to measure the extent to which judges rank gymnastics routines in the correct order. Concordance and accuracy are crucial, and while small inter-score differences may be acceptable, the most important factor is ensuring that the final ranking is fair. To calculate judges' integrity, we used the strength of association between the judge and control rankings for each routine, applying the Kendall Concordance Coefficient (W).

Kendall's W, which includes the presence of ties, was calculated as follows (Kendall & Babington-Smith, 1939; Wallis, 1939):

$$\mathrm{W} = \frac{12 \cdot \mathrm{S}}{\mathrm{m}^2(\mathrm{n}^3 - \mathrm{n}) - \mathrm{m} \cdot \sum_1^{\mathrm{m}} \mathrm{T_j}}$$

where m = number of raters, n = number of evaluated routines,

$$S = \sum_{i=1}^{n}(R_i - \bar{R})^2,$$

being R_i the sum of the ranges of the scores given by m evaluators to the ith subject and $\bar{R}$ is the arithmetic mean of the Ri, i = 1,…, n.

$$T_j = \sum_{i=1}^{g_j}(t_i^3 - t_i),$$

assigns the average of the rankings to the tied observation, where t_i is the number of tied values in the i-th grouping of ties, and gj is the number of tie groups in the j-th set of hierarchies, j = 1, ..., m.

Kendall's W values lie between 0 and 1, where 0 indicates the absence of agreement, and 1 represents total agreement. A high Kendall's W indicates that judges are likely to apply the same standards when evaluating the same competitive routines. As all switched ranking positions don't have the same relevance, the ranking swap costs vary. We proposed different Kendall's W reduction coefficients, according to the relevance of Judge and Control ranking positions being switched. When the relevance of the changed position increases, the coefficient that multiplies the value of Kendall's W decreases, and thus decreases the degree of agreement between judge and control. The different Kendall's W reduction coefficients are defined as follows: 0.7, 0.6, 0.8, 0.65, 0.82 and 0.75 when switching the 1vs3, 1vs4, 2vs3, 2vs4, 3vs4 and 3vs5 or more ranked positions, respectively.

To evaluate the qualitative ranking criteria for judges' performance, the final Kendall's W values were classified as follows: <0.95 excellent; 0.9 to 0.9499 very good; 0.8 to 0.8999 good; 0.7 to 0.7999 satisfactory; and 0.6 or less very poor.

These formulas were integrated into the ACROAJIR® spreadsheet, and a second macro function called "AJIR-macro" was developed. This macro used all the previously collected official E- and A- individual judges' scores, along with the revised control scores, to automatically and individually check all the predefined statistical and mathematical criteria. It was implemented to automatically analyze, generate, and export all the information presented in each individual report.

For each competitor and competition session, the report provides information about the execution or artistic score, the ranking assigned by each judge, and its relationship to the Control-and-Panel score and ranking, presented both numerically and graphically. If a judge's score deviation for a particular country exceeds the limit allowed by the FIG, a yellow alert is automatically displayed under the affected country in the score graph. If this difference impacts the rankings according to the criteria outlined in Table 1 and Table 2, the same yellow alert principle is applied.

The bias score compares a judge's score for their own country with the equivalent control score. If the judge-vs-control score or ranking deviation is more favorable to the judge's country than the defined allowable deviation, the score or ranking bias box will display a red alert. A quantitative and qualitative individual score and ranking evaluation was also included, using LCCC and Kendall's W values, to provide quick and understandable feedback on judges' performance.

Finally, the report presents a summary of all four E/A judges' panel evaluations. The ACROAJIR® "AJIR-macro" processes all the statistical and mathematical data to create and name each individual report. All data analysis was performed using an Excel spreadsheet (Microsoft, version 365-2019, US).

We collected data from 76 experienced international acrobatic gymnastics judges, who officiated at four official events during the 2017–2022 Olympic cycle: the 10th and 11th European Age Group Acrobatic Gymnastics Competitions (EAGC) and the 29th and 30th European Acrobatic Gymnastics Championships (ECh) held in 2019 and 2021, respectively.

To evaluate whether daily feedback on each judge's results improved their subsequent performance (in terms of accuracy and agreement with control scores) as the event progressed, each competition was divided into two parts. The first part was completed when either all judges had judged at least once or half of the competitive session had been finished. The second part encompassed all remaining competitive sessions. Judges evaluated routines approximately 3.25 ± 0.7 times in the EAGC and 4.05 ± 0.8 times in the ECh during each part. Each judge was evaluated at least once in each part.

Table 2.
Score and Ranking evaluation criteria defined between judges and control rankings. Higher differences generate a red or yellow highlighted alert.

Score evaluation criteria

Control score between		Allowed deviation
min	**max**	**judge vs Control**
9.5	10.00	0.1
8.7	9.499	0.2
8.0	8.699	0.3
7.0	7.999	0,4
6.0	6.999	0.5
5.0	5.999	0.7
0.0	4.999	1.0

Ranking evaluation criteria

Ranking positions intervals	Ranking differences between control and judge's rRanking
1st and 2nd	0 or 1 If the control score between 1st and 2nd place is greater than or equal to 0.1 point, then the difference in ranking with the control scores can be 1 place.
3rd and 4th	1
5 to 8th	2
9 to 12th	3
12th or more	4

Only competitive sessions with 6 or more competitors were used to assess judges to avoid small differences in scores causing large disparities in rankings and potentially resulting in unfair evaluations. With 6 or more competitors, the validity of the judges' evaluations improves. Since the final competition in the second part of EAGC events could only be assessed by higher-category judges, which might act as a confounding variable, we only included the qualification routines for EAGC. For the ECh event, both qualification and final competitive routines were included.

The intervention was designed to minimize significant inconsistencies in judging from one day or group to the next. Such inconsistencies were largely reduced, except for individual finals at ECh (balance or dynamic exercises). In these cases, it would require that the same judge be selected for the same role after a random draw. It is impossible for a judge to act in the same role for the same routine they had judged in qualifications at the EAGC, and it is limited to a pool of a few high-category judges at the ECh.

The independent variable was the performance in two parts of each competition event, while the dependent variables were changes in score accuracy and ranking concordance. Individual reports were sent after the completion of the 10th EAGC and 29th ECh events, without daily feedback conditions (NFBC). In contrast, for the 11th EAGC and 30th ECh events, individual reports were provided daily, within a maximum of 12 hours after the end of each competition day and before the next day’s session began, under daily feedback conditions (FBC). We compared a total of

953 reports: 272 from the EAGC and 680 from the ECh competitions.

Daily, after each competition, the control jury received all scores and validated their own accuracy in judging. The control scores validity macro quickly identified any significant differences between individual control judges' scores for all evaluated sessions. If significant differences were detected, the affected competitive routine was re-judged using video recordings at the end of each day’s last competitive session to provide a more reliable true score within the defined deviation.

Once all control judges' scores were finalized, paired judge-and-control scores were obtained for accuracy (scores) and concordance (ranking) using the AJIR macro, which generated an individual report for each judge. A total of 1280 reports were created and sent confidentially. The computer used for this analysis was a Microsoft Surface Pro 7, 12.3" (Intel Core i5-1035G4, 8GB RAM, 256GB SSD, Microsoft, Redmond, USA). To analyze the effects of judging performance, we compared judges’ performances between the first and second parts of each event, noting that daily evaluation reports were provided only for the 11th EAGC and 30th ECh events.

Standard statistical methods were used to calculate means and confidence intervals for accuracy and consistency, as previously defined. The Kolmogorov-Smirnov and Levene tests assessed normality and homogeneity of sample distributions. Data were analyzed using parametric or non-parametric tests based on these results.

Since each judging panel was drawn randomly, an unpaired t-test was used to evaluate the effects of prospective judging quality between the first and second parts of each event. Significance was set at $P \leq 0.05 P \leq 0.05$. All analyses were conducted using SPSS software version 23.0 (SPSS, Chicago, IL).

RESULTS

Figure 1 illustrates the effects of prospective judging performance between the first and second parts of each event. Inter-judge performance was significantly higher in the 2021 (FBC) compared to the 2019 (NFBC) European ACRO events, with score accuracy improving from 0.75 ± 0.14 to 0.78 ± 0.15 ($p = 0.044$) and ranking concordance improving from 0.80 ± 0.13 to 0.82 ± 0.14 ($p = 0.007$). Judges' ranking concordance significantly improved when daily evaluations were provided (0.82 ± 0.13 vs 0.77 ± 0.19; $p = 0.000$), while score accuracy improved but not significantly (0.75 ± 0.16 vs 0.76 ± 0.17; $p = 0.305$).

Within events, judges' overall accuracy was significantly better in qualification competitions at the 11th EAGC compared to the 10th EAGC (0.76 ± 0.11 vs 0.80 ± 0.11; $p = 0.007$), with 160 vs 192 AJIRs, respectively. Judges' performance in ranking concordance significantly improved in the 30th ECh compared to the 29th ECh (0.76 ± 0.21 vs 0.82 ± 0.14; $p = 0.007$), with 368 and 320 AJIRs, respectively.

Comparing judging of execution and artistic performance in the first and second parts of competition events, the 10th EAGC (NFBC) showed a significant reduction in score accuracy differences for execution in the second part (0.80 ± 0.073 vs 0.72 ± 0.18; $p = 0.013$). Although judges’ ranking concordance was lower in the second part (0.84 ± 0.10 vs 0.82 ± 0.13; $p = 0.446$), the difference was not significant. For artistic performance, there was no significant reduction in score accuracy differences in the second part (0.74 ± 0.10 vs 0.70 ± 0.15; $p = 0.221$), but there was a significant reduction in ranking concordance differences (0.77 ± 0.13 vs 0.70 ± 0.06; $p = 0.039$).

For the 29th ECh (NFBC), no significant differences in judges' accuracy for execution and artistic performance were found between the first and second parts of the competition. However, there was a significant reduction in judges’ ranking concordance in the second part for execution (0.83 ± 0.15 vs 0.75 ± 0.27; p = 0.027) and artistic performance (0.76 ± 0.12 vs 0.71 ± 0.25; p = 0.043).

In the FBC, at the 11th EAGC, judges' concordance for both artistic and execution scores improved in the second part of the competition for accuracy and ranking, with significant differences observed only for execution accuracy (0.83 ± 0.13 vs 0.70 ± 0.07; p = 0.017). At the 30th ECh (FBC), the only significant increase in judges' score accuracy was for artistic performance (0.63 ± 0.23 vs 0.73 ± 0.18; p = 0.005).

A total of 1280 individual reports were created and analyzed at the end of each competition day, but only 640 were sent confidentially for the 2021 events. For easier understanding by the judges, the accuracy and concordance values in the individual reports were multiplied by 100 (Figure 2).

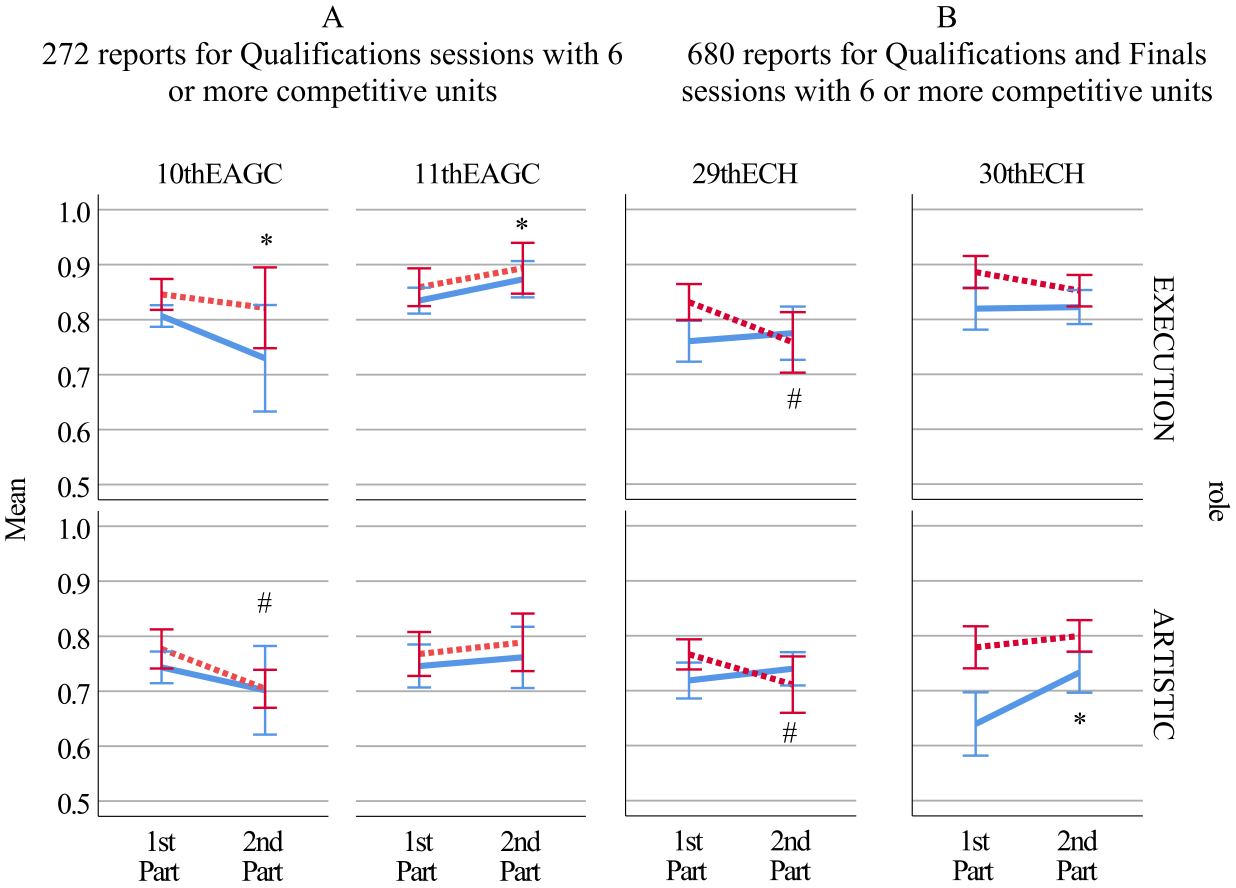
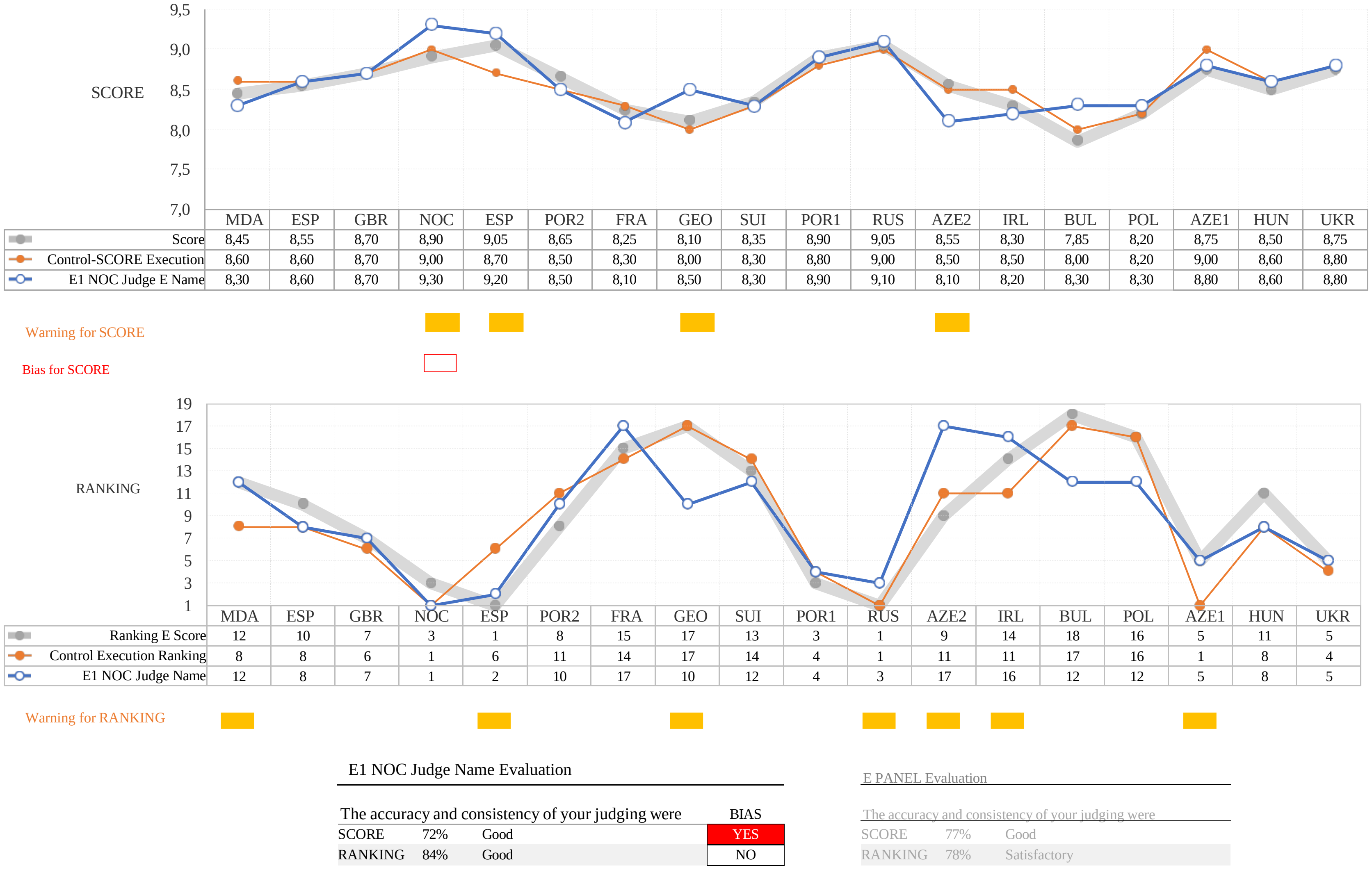


Figure 1. Mean and 95% confidence intervals for score accuracy (bold line) and ranking (dashed line) judges' evaluations for artistic performance or execution in the first and second parts of EAGC (A) and ECh (B) events (* p<0.05 score significant differences; # p<0.05 ranking significant differences).



	MDA	ESP	GBR	NOC	ESP	POR2	FRA	GEO	SUI	POR1	RUS	AZE2	IRL	BUL	POL	AZE1	HUN	UKR
Score	8,45	8,55	8,70	8,90	9,05	8,65	8,25	8,10	8,35	8,90	9,05	8,55	8,30	7,85	8,20	8,75	8,50	8,75
Control-SCORE Execution	8,60	8,60	8,70	9,00	8,70	8,50	8,30	8,00	8,30	8,80	9,00	8,50	8,50	8,00	8,20	9,00	8,60	8,80
E1 NOC Judge E Name	8,30	8,60	8,70	9,30	9,20	8,50	8,10	8,50	8,30	8,90	9,10	8,10	8,20	8,30	8,30	8,80	8,60	8,80
Warning for SCORE				🟨	🟨			🟨				🟨						
Bias for SCORE				□														

	MDA	ESP	GBR	NOC	ESP	POR2	FRA	GEO	SUI	POR1	RUS	AZE2	IRL	BUL	POL	AZE1	HUN	UKR
Ranking E Score	12	10	7	3	1	8	15	17	13	3	1	9	14	18	16	5	11	5
Control Execution Ranking	8	8	6	1	6	11	14	17	14	4	1	11	11	17	16	1	8	4
E1 NOC Judge Name	12	8	7	1	2	10	17	10	12	4	3	17	16	12	12	5	8	5
Warning for RANKING	🟨				🟨			🟨			🟨	🟨	🟨			🟨		

E1 NOC Judge Name Evaluation

The accuracy and consistency of your judging were			BIAS
SCORE	72%	Good	YES
RANKING	84%	Good	NO

E PANEL Evaluation

The accuracy and consistency of your judging were		
SCORE	77%	Good
RANKING	78%	Satisfactory

Figure 2. An example of individual reports in one competitive session. Session and judge data are hidden to protect confidentiality. (Wide bright line: Panel score; Thin line with filled circles: Control score; Bold line with empty circles: Judge score

DISCUSSION

The aim of this study was to design and apply the ACROAJIR® tool to control and evaluate ACRO judges' performance and prospective judging effects at major competitive events. The tool had practical applications in three domains: a) a specific Technical Committee (TC) control task; b) feedback for individual judges; and c) assessment of prospective judging effects.

Fulfilling a specific TC control task, the ACROAJIR® results provided the Technical Committee (TC) with an overview of results and specific, objective information about judges' performance in each competitive session. This information facilitated easy, quick, and accurate identification of individual judges' mistakes. It offered strong evidence for managing these mistakes, supporting correct decision-making, and alerting judges to potential future issues. TC members received objective data on the accuracy, concordance, and bias of judges' scores.

Performance feedback is commonly used to influence behavior, and providing information about past performance is a widely adopted strategy in competitions. However, the effects of daily performance feedback have not been previously evaluated. This quantitative study analyzed how daily feedback, supported by the AJIRs-based formative assessment process, affected judges' accuracy and concordance. Each judge received a simple daily report on their performance in terms of accuracy and concordance at the 2021 competition events, and those with the best scores were congratulated. Knowledge of the control score enhanced judges' self-confidence and consistency by providing an objective assessment of their accuracy, consistency, and concordance relative to the control score. None of the judges disagreed with the reports received, and all appreciated the daily feedback effort. No prior studies with similar designs were found.

Knowing that they would be evaluated daily appeared to motivate judges to consistently perform their best. The effects were differentiated based on whether feedback was provided daily or not, particularly for characteristics such as score accuracy and concordance.

Feedback included comparisons with benchmarks beyond the in-competition trimmed mean. Overall, judges' performance significantly improved when they were aware of daily evaluations. Specifically, judges' score accuracy was significantly better with daily reports at the 11th EAGC, and ranking concordance was better at the 30th ECh. While judges knew they would be evaluated at the start of each competition, they did not anticipate receiving daily reports.

Overall, judges' performance was significantly worse in the second part of competitions where no feedback was given, compared to when they received daily feedback, which either maintained or improved performance. Significant declines were observed in execution accuracy scores and artistic ranking concordance at the 10th EAGC (NFBC) and in artistic ranking concordance at the 29th ECh (NFBC) during the second parts of these events.

In contrast, with daily feedback conditions (FBC), both accuracy and ranking concordance improved in the second part of the competition for both artistic performance and execution at the 11th EAGC. At the 30th ECh, judges' score accuracy for artistic performance was significantly higher.

Previous studies examining judges in gymnastics, judo, rope climbing, and synchronized swimming found that when judges received open feedback (i.e., the ability to hear or see their colleagues' scores after each performance), the variation between scores was significantly lower. This suggests that conformity was influenced by informational factors (Auweele et al., 2004; Boen et al., 2006, 2008, 2013), which supports our findings.

However, reference panel scores can sometimes be incorrect, potentially influencing a judge's decisions. This normative conformity bias can be dangerous and lead to unfair results. Even if a judge's score is accurate, consistently aligning with the panel's score when deviations occur can compromise judgment, leading to normative conformity bias. Daily feedback on control scores mitigates this risk by boosting judges' confidence in their own judgments, thereby motivating better performance in future sessions.

In summary, judges' performance either remained stable or improved when they were consistently updated on their performance. Overall, higher score accuracy was associated with greater ranking concordance. However, when routines were at a similar level, small changes in score accuracy led to significant changes in ranking concordance. This assessment proved valuable for detecting instances where judges might exploit small but permissible scoring gaps to favor their own countries. It also provided feedback on judges' scoring patterns, which could be useful for training and accrediting judges. Updates and feedback can help propose corrective measures for judges who perform below expectations (Mercier & Heiniger, 2018).

Judges aim to perform at their best, and the results demonstrated a high level of quality overall. Although a consistently high performance might limit improvements as the event progresses, knowledge of daily evaluations during the 2021 events led to significantly more accurate scores as the competition continued. This article introduces a novel approach to evaluating judges' performance during live competitions. To our knowledge, providing individual written feedback reports during competitions has not been previously implemented. This method suggests new active methodologies and formative evaluations for future use.

The current study had several limitations. First, the use of expert superior jury scores as 'true' scores introduces potential issues, as these expert scores might also be inaccurate or not align with the judging panel. This could affect the evaluation of judges during live competitions. The ranking swap costs defined in this study might be better represented by more sophisticated regression equations to explain all relevant ranking swap cases. Additionally, refining the definitions for the first and second periods and using the tool solely for pedagogical purposes, without sanctions for biased or incorrect judgments, could impact the number of significant differences observed in judges' performance as the events progressed. Although post-feedback improvements in accuracy were noted, understanding the process behind this alignment would provide insights into the cause of discrepancies. Future research should include more examples to validate the findings of this study. With more comprehensive evidence, further actions can be taken to enhance the rating system for the discipline (Anderlucci et al., 2020).

CONCLUSION

The ACROAJIR® tool offered timely, valuable, and personalized feedback on accuracy and concordance scores for acrobatic gymnastics judges during competitions. It demonstrates that such feedback can be effectively delivered during, rather than only after, competition events. The tool facilitates specific TC scoring control tasks, provides judges with evidence-based feedback, and suggests targeted improvements for prospective judging.

REFERENCES

Akoglu, H. (2018). User's guide to correlation coefficients. Turkish Journal of Emergency Medicine, 18(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Anderlucci, L., Lubisco, A., & Mignani, S. (2020). Investigating the Judges Performance in a National Competition of Sport Dance. Social Indicators Research.

https://doi.org/10.1007/s11205-019-02256-z

Auweele, Y. Vanden, Boen, F., De Geest, A., & Feys, J. (2004). Judging bias in synchronized swimming: Open feedback leads to nonperformance-based conformity. Journal of Sport and Exercise Psychology, 26(4), 561–571. https://doi.org/10.1123/jsep.26.4.561

Boen, F., Ginis, P., & Smits, T. (2013). Judges in judo conform to the referee because of the reactive feedback system. European Journal of Sport Science, 13(6), 599–604. https://doi.org/10.1080/17461391.2012.756070

Boen, F., van Hoye, K., Auweele, Y. Vanden, Feys, J., & Smits, T. (2008). Open feedback in gymnastic judging causes conformity bias based on informational influencing. Journal of Sports Sciences, 26(6), 621–628. https://doi.org/10.1080/02640410701670393

Boen, F., Vanden Auweele, Y., Claes, E., Feys, J., & De Cuyper, B. (2006). The impact of open feedback on conformity among judges in rope skipping. Psychology of Sport and Exercise, 7(6), 577–590. https://doi.org/10.1016/j.psychsport.2005.12.001

Bučar, M., Čuk, I., Pajek, J., Karacsony, I., & Leskošek, B. (2012). Reliability and validity of judging in women's artistic gymnastics at University Games 2009. European Journal of Sport Science, 12(3), 207–215. https://doi.org/10.1080/17461391.2010.551416

Dallas, G., Mavidis, A., & Chairopoulou, C. (2011). Influence of angle of view on judges' evaluations of inverted cross in men's rings. Perceptual and Motor Skills, 112(1), 109–121. https://doi.org/10.2466/05.22.24.27.PMS.112.1.109-121

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. Journal of Experimental Psychology: Applied, 12(3), 166–178. https://doi.org/10.1037/1076-898X.12.3.166

De Bruin, W. B. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. Acta psychologica, 118(3), 245-260. https://doi.org/10.1016/j.actpsy.2004.08.005

Dosseville, F., Laborde, S., & Garncarzyk, C. (2014). Current research in sports officiating and decision-making. In C. Mohiyeddini (Ed.), Contemporary topics and trends in the psychology of sports (pp.13–38). Nova Publishers.

Fasold, F., Memmert, D., & Unkelbach, C. (2012). Extreme judgments depend on the expectation of following judgments: A calibration analysis. Psychology of Sport and Exercise, 13(2), 197-200. https://doi.org/10.1016/j.psychsport.2011.11.004

FIG. (2015). Regulations for the judges' evaluation programme (JEP) "former fairbrother system" and its' application. Federation International de Gymnastique.

FIG. (2020). 2017- 2020 Fig Judges' Rules Specific Rules for Acrobatic. In Specific Rules for Acrobatic Gymnastics. Federation International de Gymnastique.

FIG. (2017). Appendix to the Codes of Points (COP). Federation International de Gymnastique.

Flessas, K., Mylonas, D., Panagiotaropoulou, G., Tsopani, D., Korda, A., Siettos, C., Di Cagno, A., Evdokimidis, I., & Smyrnis, N. (2015). Judging the judges' performance in rhythmic gymnastics. Medicine and Science in Sports and Exercise, 47(3), 640–648. https://doi.org/10.1249/MSS.0000000000000425

Gambarelli, G., Iaquinta, G., & Piazza, M. (2012). Anti-collusion indices and averages for the evaluation of performances and judges. Journal of Sports Sciences, 30(4), 411–417. https://doi.org/10.1080/02640414.2011.651153

Heiniger, S., & Mercier, H. (2018). National Bias of International Gymnastics Judges during the 2013-2016 Olympic Cycle. http://arxiv.org/abs/1807.10033

Heiniger, S., & Mercier, H. (2021). Judging the judges: evaluating the accuracy and national bias of international gymnastics judges. Journal of Quantitative Analysis in Sports, 17(4), 289-305. https://doi.org/10.1515/jqas-2019-0113

Kramer RSS. (2017) Sequential effects in Olympic synchronized diving scores. Royal Society Open science. 4: 160812. http://dx.doi.org/10.1098/rsos.160812

Kendall, M. G., & Babington-Smith, B. (1939). The Problem of m Rankings. The Annals of Mathematical Statistics, 10(3), 275–287. https://doi.org/http://dx.doi.org/10.1214/aoms/1177732186

León-Prados, J. A., & Jemni, M. (2022). Reliability and agreement in technical and artistic scores during real-time judging in two European acrobatic gymnastic events. International Journal of Performance Analysis in Sport, 22(1), 132–148. https://doi.org/10.1080/24748668.2021.1996913

León-Prados, J. A., & Rosales, A. (2019). ACROAJIR®; Automatic ACRO Judges Individual Report Tool. Universidad Pablo de Olavide.

Leskošek, B., Čuk, I., & Peixoto, C. J. D. (2018). Inter-rater reliability and validity of scoring men's individual trampoline routines at European championships 2014. Science of Gymnastics Journal, 10(1), 69–79.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45(1), 255–265.

McBride, G. (2005). A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient. NIWA Client Report, HAM2005-062 https://www.medcalc.org/download/pdf/McBride2005.pdf

Mercier, H., & Heiniger, S. (2018). Judging the Judges: Evaluating the Performance of International Gymnastics Judges. http://arxiv.org/abs/1807.10021

Mercier, H., & Klahn, C. (2017). Judging the judges : Evaluating the performance of international gymnastics judges. MIT Sloan Sports Analytics. https://arxiv.org/pdf/1807.10021.pdf

Plessner, H., & Schallies, E. (2005). Judging the cross on rings: A matter of achieving shape constancy. Applied Cognitive Psychology, 19(9), 1145–1156. https://doi.org/10.1002/acp.1136

Premelč, J., Vučković, G., James, N., & Leskošek, B. (2019). Reliability of judging in DanceSport. Frontiers in Psychology, 10. 1001. https://doi.org/10.3389/fpsyg.2019.01001

Ste-Marie, D. M. (2000). Expertise in women's gymnastics judging: an observational approach. Perceptual and Motor Skills, 90(2), 543–546. https://doi.org/10.2466/pms.2000.90.2.543

Van Bokhorst, L. G., Knapová, L., Majoranc, K., Szebeni, Z. K., Táborský, A., Tomić, D., & Cañadas, E. (2016). "It's always the judge's fault": Attention, emotion recognition, and expertise in rhythmic gymnastics assessment. Frontiers in Psychology, 7(JUL). https://doi.org/10.3389/fpsyg.2016.01008

Wallis, W. A. (1939). The Correlation Ratio for Ranked Data. Journal of the American Statistical Association, 34(207), 533–538. https://doi.org/10.1080/01621459.1939.10503552

Corresponding author:

Juan Antonio León-Prados,
Physical Performance & Sport Research Center, Pablo de Olavide University,
Pablo de Olavide' University, Carretera de Utrera km 1
41013, Seville, Spain,
phone: (+34) 606701338,
e-mail: jaleopra@upo.es

Article received: 3.12.2023
Article accepted: 2.4.2024