BIAS OF JUDGING IN MEN'S ARTISTIC GYMNASTICS AT THE EUROPEAN CHAMPIONSHIP 2011

Accepted
for publication
05.02.2012

AUTHORS: Leskošek B.[1], Čuk I.[1], Pajek J.[2], Forbes W.[3], Bučar-Pajek M.[1]

[1] Faculty of sport, University of Ljubljana, Slovenia
[2] University Medical Center of Ljubljana, Slovenia
[3] Australian Institute of Sport, Australia

Reprint request to:
Bojan Leskošek
Fakulteta za šport, Gortanova 22
1000 Ljubljana, Slovenia
E-mail: bojan.leskosek@fsp.uni-lj.si

ABSTRACT: The purpose of this study was to establish the validity (unbiasedness) and reliability of E-panel judges officiating execution of exercises in men's artistic gymnastics at the European Championship 2011 (EC 2011) in Berlin. Overall bias was established in terms of average over-scoring or under-scoring of each judge compared to the final E score of a judges' E panel. National bias was expressed as average over-scoring of gymnasts of the same nationality as the judge's. Both types of bias were mostly small (within the +/- 0.1 point range), but statistically significant and also substantial (over 0.2 point) in some cases. Compared to other competitions, it seems that bias is becoming smaller over time and is also smaller in competitions of higher importance. Analysis of possible consequences of bias showed that overall bias may influence both scores and ranks of competitors, while national bias may be especially problematic in the qualification round, where it may prevent some competitors from qualifying for apparatus finals.

KEY WORDS: sports, gymnastics, evaluation, bias, reliability, nationalism

INTRODUCTION

As one of the most widespread and popular sports, artistic gymnastics has been part of the official programme of the modern Olympic games since its beginning in 1896. As such it has attracted much attention in science over the years and recently a new journal (Science of Gymnastics Journal) emerged and several scientific papers were published elsewhere (e.g. [10]).

Artistic gymnastics is one of the sports (along with diving, figure skating and synchronized swimming) in which competition results (scoring and ranking of athlete's performance) heavily depend on the judges' evaluation. This is in contrast to some other sports, e.g. athletics, where results are recorded by precise technical instruments, or sports like basketball, where scoring is formally confirmed by the judge, but usually is not perceived as problematic by experts or spectators.

At first, in gymnastics competitions only one judge evaluated a gymnast, while nowadays at most important competitions there are more (up to six) judges evaluating execution (and artistry) and two judges evaluating difficulty, composition requirements and connection value. Difficulty judges determine their so-called D score in agreement, while each member of the execution jury gives its score independently of each other. At the European Championship 2011 the execution (E) score of every routine was calculated as the average of the middle four judges' scores (i.e. not counting the lowest and highest score). The final score was simply the sum of E and D scores.

While evaluating the D score is usually not seen as problematic, E score derivation has been many times criticized by spectators, commentators, officials, coaches, gymnasts and also by researchers. Although the Code of Points [9] exactly prescribes the deductions (i.e. 1 point for falling off the apparatus, 0.1, 0.3 and 0.5 for small, medium and large errors, respectively), E-panel judges usually give different scores for the same exercises. It is not possible to univocally decide if a particular judge is "right" or "wrong" as there is no golden standard to compare with. However, in many cases it is agreed that a large departure of a judge's score from the final E score is an error. The reasons for these errors should be divided into two categories: random and systematic. Random error is usually expressed as random variation of E-panel judges' scores around E (or mean) scores; the smaller this variation is, the higher the reliability.

Perhaps even more problematic are the systematic errors, i.e. frequent or consistent (under- or over-grading) departures of scores. They are usually expressed as the ratio between under- and over-grades (excluding ties) or as the average departure of a judge's score from the E score. The judge with the ratio closest to one or the smallest mean departure from the E score is considered the most unbiased.

Many types and reasons for bias of officiating have been hypothesized and empirically substantiated. Several authors [1,18] have found (inter)national bias, i.e. higher scoring of gymnasts from the judges' own country and lower scoring of all others or just the closest competitors. A similar type of bias, home advantage bias, was also proven for the 1896-1996 Olympic Games [4]. Others [6,11,12]

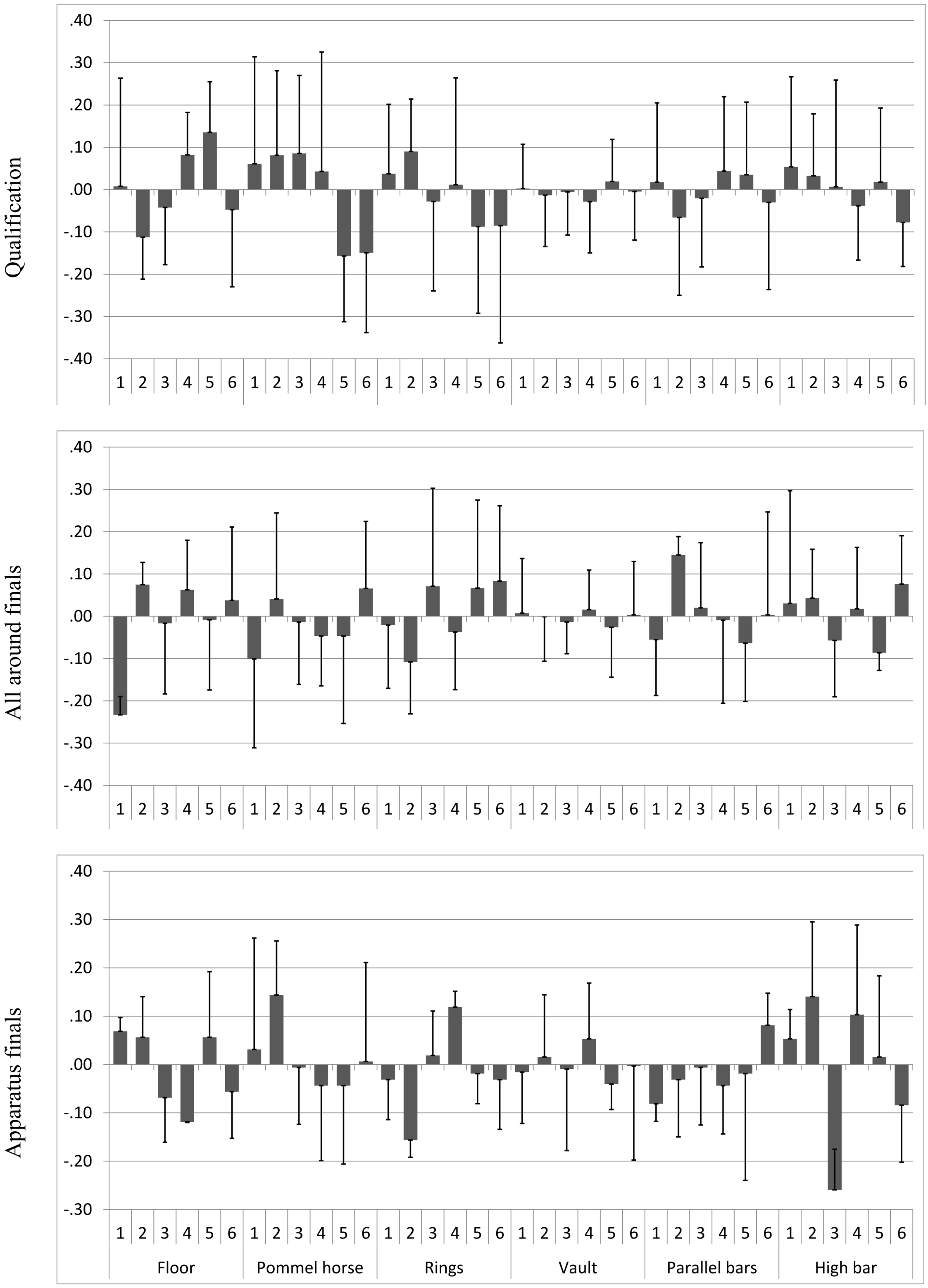

FIG. 1. AVERAGE AND STANDARD DEVIATION OF DIFFERENCE BETWEEN JUDGE'S E SCORE AND FINAL E SCORE FOR INDIVIDUAL JUDGES BY SESSION AND APPARATUS

have found substantial overall judges' bias, i.e. systematic under- or over-scoring of judges. Several authors have reported sequential order bias [2,7,14,15,17] and open feedback / conformity bias [5]. Another bias was found based on the position of the judge in relation to the apparatus [16].

The aim of this study was to evaluate overall and national biases in E-panel judges at one of the most important competitions in gymnastics, the 4th European Artistic Gymnastics Individual Championships for seniors, held during 6–10 April, 2011 in Berlin, Germany. As bias may be sometimes fictitious, i.e. it may actually express unreliability instead of validity of judging, especially when the number of competitors is small (e.g. in apparatus finals), reliability of officiating was also evaluated. An additional goal of the study was to compare the results of this study to official criteria of selecting and evaluating performance of judging, described in the 2009 FIG General Judges' Rules [9], and to propose possible modification and refinement of those rules.

MATERIALS AND METHODS

E scores for all male gymnasts competing in any session (qualification, all around and apparatus finals) were obtained from the official book of results.

For each set of analysis we calculated statistics for the E score, item (individual judge) and scale (all judges together) scores. The following reliability and validity measures were calculated: intraclass correlation, Cronbach alpha, Kendall coefficient of concordance W, and Armor's theta coefficient, which is based on the first (largest) eigenvalue from the principal component analysis of the correlations between judges' scores.

Nationality bias was calculated both as number of scores lower/higher than the final E score and also as the average difference of the judge's E score from the final E score. As there were no judges of the same nationality as competitors in the apparatus final, nationality bias was not evaluated for that session.

Full blinding of the judges involved was undertaken. To protect the judges' and countries' anonymity, we randomly changed their position in the analysis from the book of results. All data were analysed with PASW Statistics 18.0.3 whenever possible, otherwise in Microsoft Excel.

RESULTS

Overall judges' bias. The bias of each judge was expressed as the (average) difference between the judge's E score and the final E score (Figure 1). Most of the biases are within the +/- 0.1 range. In the qualification round, the largest positive bias (overestimation) was found for judge #5 (M=+0.14) on floor, while the largest negative bias (underestimation) existed in judges #5 and #6 on pommel horse (M=–0.16 and M=–0.15, respectively). In those two judges, also the standard deviation of the E score difference was high (over s=0.3), meaning that those judges not only systematically underestimated athletes' performance, but also that their judging was unreliable. Unreliable judgment was also found in some other judges with less biased judgment, e.g. judge #6 on rings and #4 on pommel horse.

In all-around finals by far the largest bias was shown by judge #1 on floor, whose absolute value of bias (M=–0.23) even exceeds the value of the standard deviation (s=0.19). A similar, but even more extreme case was judge #3 in high bar apparatus finals

TABLE 1. VALIDITY AND RELIABILITY MEASURES OF JUDGING BY SESSION AND APPARATUS

Session	apparatus	n	Kendall W*	Cronbach Alpha	ICC$_{average}$	ICC$_{single}$	Armor's theta
qualification	Floor	94	0.11	0.98	0.87	0.97	0.98
	P. horse	94	0.10	0.98	0.87	0.98	0.98
	Rings	93	0.04	0.96	0.79	0.96	0.96
	Vault	113	0.01	0.98	0.89	0.98	0.98
	Par. bars	92	0.02	0.97	0.84	0.97	0.97
	High bar	89	0.06	0.97	0.86	0.97	0.98
all around finals	Floor	24	0.20	0.95	0.72	0.94	0.96
	P. horse	24	0.06	0.96	0.79	0.96	0.96
	Rings	24	0.07	0.87	0.52	0.87	0.89
	Vault	24	0.01	0.98	0.91	0.98	0.99
	Par. bars	24	0.10	0.97	0.82	0.96	0.97
	High bar	24	0.08	0.96	0.77	0.95	0.96
apparatus finals	Floor	8	0.24	0.98	0.87	0.98	0.98
	P. horse	8	0.04	0.99	0.93	0.99	0.99
	Rings	8	0.22	0.95	0.73	0.94	0.96
	Vault	16	0.07	0.98	0.90	0.98	0.98
	Par. bars	8	0.20	0.97	0.83	0.97	0.97
	High bar	8	0.24	0.97	0.78	0.96	0.97

Notes: * underlined values of Kendall W are significant at alpha=0.05 level

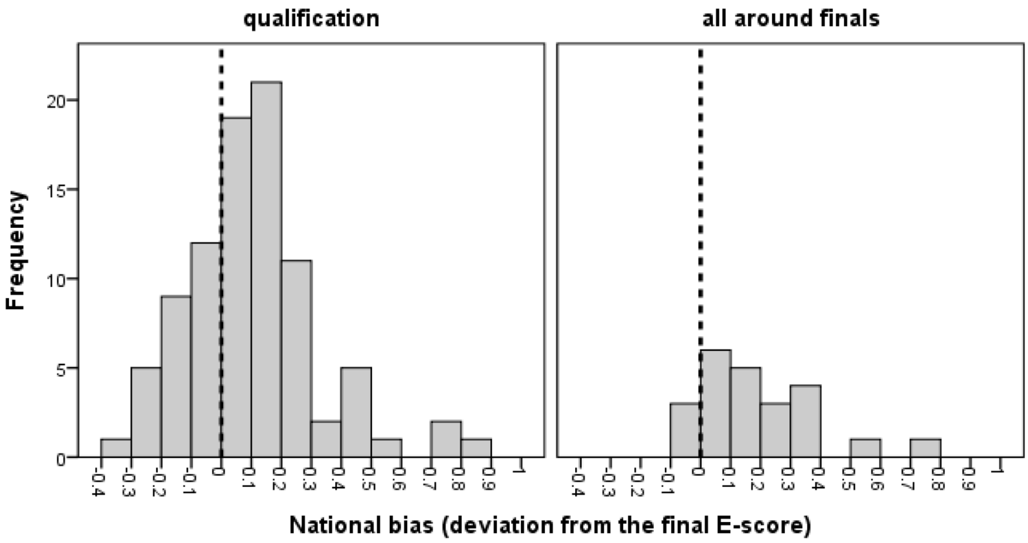

FIG. 2. DISTRIBUTION OF DEVIATION OF E SCORES FOR ATHLETES OF THE SAME NATIONALITY AS THE JUDGE'S FROM THE FINAL E SCORE BY SESSION. VERTICAL DASHED LINES REPRESENT POINTS OF UNBIASED SCORES

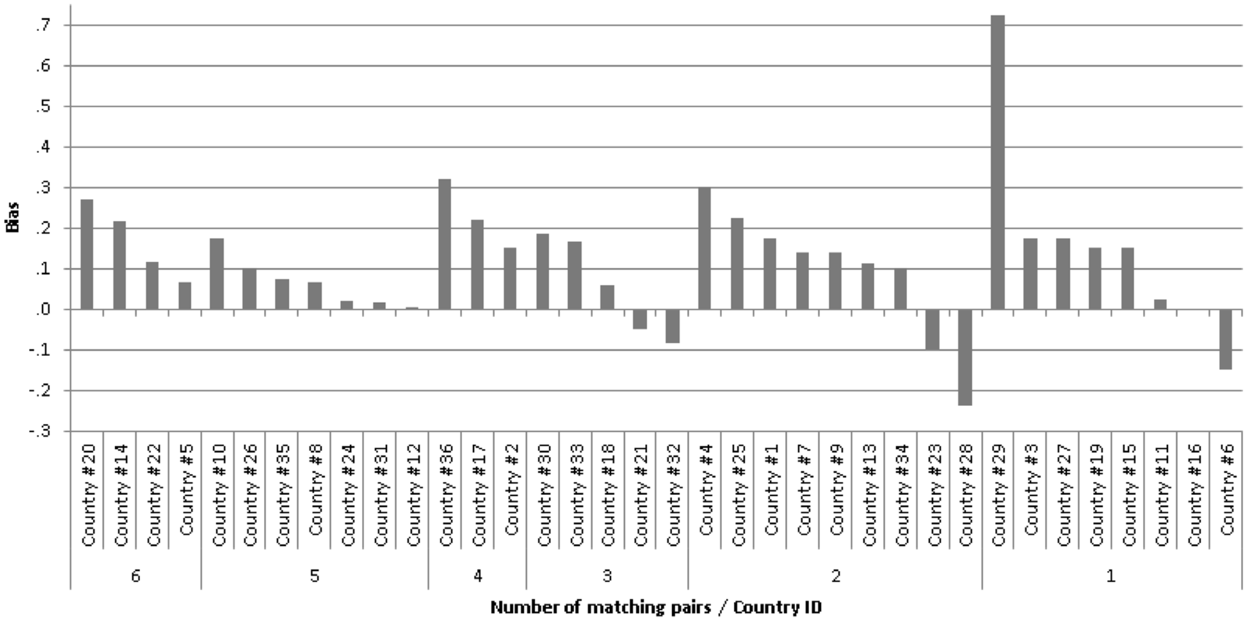

FIG. 3. BIAS (AVERAGE DEVIATION FROM FINAL E SCORE) FOR JUDGES MATCHING THE NATIONALITY OF COMPETITOR BROKEN BY COUNTRY AND NUMBER OF COMPETITOR/JUDGE MATCHING PAIRS

(M=0.26, s=0.18). Those two judges were the only ones in all three sessions for whom the (absolute) value of bias exceeded 0.2, while some other judges were found both in all-around and apparatus finals whose bias exceeded 0.1. The smallest bias in all three sessions was found on vault (Figure 1).

In order to evaluate the validity and reliability of judging, several measures were computed (Table 1). Regarding validity (unbiasedness), in apparatus finals Kendall W exceeded 0.2 in all apparatuses except pommel horse and vault, but none of these values were statistically significant at the 5% level, as the number of cases (athletes) in this round is the smallest (i.e. eight, compared to 24 in all-around finals and around 90 in qualification). Regarding reliability, most measures, especially in qualification, are high. The notable exception is rings in all-around finals, where Cronbach alpha and Armor's theta fall below 0.9, while in all other apparatuses/sessions they are over 0.95. Similarly, intraclass correlation coefficients on rings in all-around finals are much lower ($ICC_{average}$=0.52, ICC_{single}=0.87) than in any other apparatus/session.

National bias

112 out of 775 (14.5%) execution scores were given by a judge whose nationality (National Olympic Committee) matched the nationality of an athlete. According to the rules, only one judge from each country is allowed to participate in each apparatus jury. At EC 2011 the number of athlete-judge pairs matching on nationality varied from 10 (on pommel horse) to 22 (on vault) in qualifications and

from 3 (on floor) to 6 (on pommel horse) in all-around finals. There were no matching pairs in all-around finals on high bar. Also, no matching pairs in apparatus finals exist, as they are forbidden by the rules.

A small, but evident nationality bias exists when analysing all 112 "national" scores (Figure 2). In 77 (69%) cases a judge of the same nationality as the competitor gave a higher execution mark than the final E score. In 6 cases the marks were the same, while only in 29 cases was the "national" judge E score lower than the final E score.

On average, scores of "national" judges exceed the final E score by M=0.119 points (Me=0.1, s=0.219), much more so in all-around finals (M=0.173, n=23) than in qualifications (M=0.099, n=89). Among the apparatuses, all biases were positive, both in qualification and in all-around sessions. The highest bias was found in rings, both in qualification (M=0.17, n=16) and in all-around finals (M=0.34, n=5), whereas the smallest one was found in qualification on pommel horse (M=0.05, n=10) and in all-around finals on vault (M=0.07, n=4).

When considering the countries with more than three competitors matching the nationality of judges, the bias was above 0.2 (M=0.32, M=0.27, and M=0.22, respectively) in three out of 14 countries (Figure 3). Negative bias (underestimation) was found only in 5 countries, all of which had only three or fewer competitors of matching nationality.

DISCUSSION

In this study, we report the biases and reliability indices at one of the highest ranked male gymnastics competitions. Since different kinds of bias have been well known and documented for decades and the European Championship is one of the most important competitions, one would expect that nowadays bias in officiating would be minor or non-existent. Contrary to this expectation, both types of bias were found in our study.

Overall biases, expressed as the mean difference of judges' E score from the final E score, fell in most cases within the +/- 0.1 range and only in two cases exceeded 0.2 points. Even though these biases may seem small, they are relatively large when compared to the small variability of individual E scores for a single competitor. Besides this, since the differences in final E scores between competitors are generally small, even small biases may result in a change of competitor's rank in the result list, which is most problematic for the medal positions.

It seems that biases are somewhat higher in apparatus finals than in qualification and all-around finals. However, as there are only eight competitors in event finals (compared to almost 100 in qualification and 24 in all-around finals), it is hard to conclude that in general event finals are the most biased session. This fact is also expressed in the Kendall W coefficients, which are highest in event finals, although none of them is statistically significant, while 4 coefficients in qualification and two in all-around finals were significant at the 5% level.

Although overall bias of most judges is small, in some cases it is substantial. On floor in all-around finals one judge underestimated execution on average by 0.23 point; in 22 out of 24 cases his scores were lower than the final E score and in 14 cases they were the smallest among all the judges. However, only three of his scores were outside the deviation from the E score allowed by the 2009 FIG General Judges' Rules [9] (Table 2), so according to those rules this judge's officiating was marked as *good*, i.e. on the third level of a five-level scale, so no actions – not even a verbal warning – were taken against him. Similarly, average underestimation of M=0.26 was found for one judge in apparatus finals on high bar. Out of 8 scores, in only one case was his grade equal to the final E score, while in four cases his grade was the smallest among the six judges. Again, no actions were taken against this judge; on the contrary, as none of the scores of this judge were outside the allowed deviation from the final E score, his officiating was marked as *excellent.*

TABLE 2. FIG ALLOWED DEVIATION BY JUDGE ACCORDING TO FINAL E-JURY SCORE [9]

E-JURY DEDUCTION TOLERANCES

Gymnast's final deduction	Score	Allowed deviation by judge
0 – 0.40	9.60 - 10.00	0.10
>0.40 – 0.60	9.40 - 9.60	0.20
>0.60 – 1.00	9.00 - 9.40	0.30
>1.00 – 1.50	8.50 - 9.00	0.40
>1.50 – 2.00	8.00 - 8.50	0.50
>2.00 – 2.50	7.50 - 8.00	0.60
>2.5	< 7.50	0.70

Comparing the apparatuses, by far the smallest biases were found in vault. This is probably not a surprise as officiating on vault seems most simple because only one "element" needs to be evaluated. Scores for vault were found not only most unbiased, but also most reliable. The highest bias was found on floor. Again, it seems this finding is associated with the number of elements usually found in floor exercises, which is highest among all the apparatuses.

Overall bias was rarely computed in previous studies. In a study of the data from the University Games 2009, a competition ruled by the same Code of Points as the European Championship 2011, although there were only four judges in qualification and all-around finals, Leskošek et al. [12] found similar overall bias in qualification and all-around finals. In those sessions the average of absolute values of biases was between 0.05 and 0.06 in both competitions. In apparatus finals, in which six judges were officiating in both competitions, this average was much higher in Belgrade 2009 (M=0.087) than in Berlin 2011 (M=0.059) competitions. Differences between these two competitions also exist in biases on different apparatuses. While in Berlin 2011 biases on vault were the smallest between all the apparatuses and biases on the floor were among the highest, no such consistency was found in the Belgrade 2009 data;

e.g. in qualifications by far the highest biases were found on rings, followed by vault. However, also in Belgrade vault had the smallest biases in all-around finals and event finals. On the basis of these data and regarding the European Championship as a much more important competition than the University Games, we may speculate that on one hand overall bias is partly random (depending on the selection of a particular panel of judges) and partly apparatus dependent (with the smallest biases on vault and probably the highest on floor); it is worth noting that in doubtful situations judges have to decide what amount of error will be deducted, small (0.1):medium (0.3) or medium (0.3):big (0.5) [13], which could also be a basis of bias.

Available data do not allow us to estimate the consequences of overall judge's bias on competitor performance, i.e. the rank achieved on a result list. If a bias is a result of a consistent under- or over-scoring of all or the majority of competitors, for example as a result of overall strict (rigorous, severe) or mild (lenient, easy) judging, this should only affect competitors' scores but not their ranks. However, if a judge's bias is a result of e.g. over-scoring of a small group of competitors (of the same nationality, home competitors, stars etc.) on one hand and under-scoring of a large group of competitors (of different nationalities, not from the home country, rookies etc.) on the other, this would result in a change of both the score and the rank. It should also be noted that even if with a particular judge there is no overall bias found, it does not guarantee an absence of different specific biases, e.g. sequential order bias, which is defined by under-scoring at the start of competition and hidden by over-scoring later on.

The second type of bias, national bias, was greater than the overall judge's bias. In many cases average over-scoring of a competitor from the same country as the judge exceeds the final E score by more than 0.1 point and even 0.2 point. Although this result may be somewhat unreliable due to the small number of competitors, a consistently positive bias was found in all apparatuses and in both sessions (qualification, all-around finals) and also from judges from all countries with more than three competitors of the matching nationality. The results found may also be considered statistically significant, as the probability for 77 or more over-scores in 112 independent (unbiased) "trials" is close to zero ($p<0.001$).

The results of the study are concordant with previous studies of national bias, although it seems that the bias is becoming smaller. Even though (due to changes made to the Code of Points) the results may not be directly comparable to the Berlin 2011 event, in men's results at the Olympic Games 1984, Ansorage & Scheer [1] found 160 (81%) over-scores, 17 under-scores and 21 ties in final scores, compared to 77 (69%) over-scores, 29 under-scores and 6 ties at the European Championship 2011. Note that biases found in the 1984 Olympics might have been even greater if there had been no boycotts from most Eastern bloc countries.

In contrast to overall bias, there is no doubt about the consequences of national bias. In the qualification session, it is obvious that national bias may influence qualifying for apparatus finals. In all-around finals the consequences are not so dramatic, as all 36 judges were of different nationalities, and all the competitors except one were judged by one and only one judge of the same nationality. As the Code of Points [8] does not allow judges of the same nationality in apparatus finals, no national bias (as defined here) exists in that round. But again, similar kinds of bias (e.g. neighbouring countries, countries with the same or similar political, ethical or religious structure, etc.) may still exist in any session of the competition.

As bias may be inflated by unreliable officiating (especially in apparatus finals, where only eight competitors take part), reliability indices were also computed. However, it was found that reliability was high. With the exception of rings in all-around finals, all Cronbach alpha and intraclass correlation coefficients were higher than 0.94. These indices are in concordance with those found at the World Championship in London 2009 [3], where all Cronbach alpha coefficients were at or above 0.94.

CONCLUSIONS

The scores given to male gymnasts at the European Championship 2011 clearly reveal both types of biases examined, i.e. overall and national bias. Although those biases are generally small, they are consistent and statistically significant. Regardless of whether the biases are conscious and intentional or not, it is evident that the pledge of complete impartiality is not fully secured at this and probably at most other gymnastic events.

According to the current (2009) rules, officiating of judges is evaluated solely on the number of excessive deviations from the final E score. This study showed that high (good and excellent) grades were given even to judges with obviously poor officiating. Therefore, FIG should consider introducing more stringent criteria for evaluating officiating and probably also for obtaining a specific judge category (brevet). Additionally, in order to reveal overall, national and other types of bias, further criteria should be introduced into FIG judges' evaluation system beside the number of excessive deviations.

In this study we quantified and explained the causes and consequences of biases in officiating at one of the most important gymnastic events. But in some cases, available data and former studies do not allow us to fully evaluate all possible biases. Therefore further studies are needed for such evaluation, and rule changes are needed to minimize the impartiality in gymnastics officiating.

Acknowledgement

We are grateful to Gianfranco Marzolla, the President of the Men's Technical Committee of the European Gymnastics Federation, who supported testing of the Real Time Judging System (RTJS) during the European Championship in Berlin 2011. This article is the first step in RTJS research.

REFERENCES

1. Ansorge C.J., Scheer J.K. International bias detected in Judging Gymnastics Competition at the 1984 Olympic Games. Res. Q. Exerc. Sport 1988;59:103-107.
2. Ansorge C.J., Scheer J.K., Laub J., Howard J. Bias in Judging Women’s Gymnastics Induced by Expectations of Within-Team Order. Res. Q. 1978;49:399-405.
3. Atiković A., Delaš Kalinski S., Bijelić S., Avdibašić Vukadinović N. Analysis results judging world championships in men’s artistic gymnastics in the London 2009 year. SportLogia 2011;7:95-101.
4. Balmer N.J., Nevill A.M., Williams A.M. Modelling home advantage in the Summer Olympic Games. J. Sport Sci. 2003;21:469-478.
5. Boen F., van Hoye K., Vanden Auweele Y., Feys J., Smits T. Open feedback in gymnastic judging causes conformity bias based on informational influencing. J. Sport Sci. 2008;26:621-628.
6. Bučar M., Čuk I., Pajek J., Karacsony I., Leskošek B. Reliability and validity of judging in women’s artistic gymnastics at University Games 2009. Eur. J. Sport Sci. Available at: http://dx.doi.org/10.1080/17461391.2010.551416. Accessed 14.10.2011.
7. Damisch L., Mussweiler T., Plessner H.Olympic Medals as Fruits of Comparison? Assimilation and Contrast in Sequential Performance Judgments. J. Exp. Psychol.-Appl. 2006;12:166-178.
8. FIG. Code of Points for Men Artistic Gymnastics Competitions (2009 Edition). Available at: http://figdocs.lx2.sportcentric.com/external/serve.php?document=2921. Accessed 14.10.2011.
9. FIG. 2009 FIG General Judges’ Rules. Available at: http://figdocs.lx2.sportcentric.com/external/serve.php?document=2024. Accessed 14.10.2011.
10. Heinen T., Jeraj D., Thoeren M., Vinken P.M. Target-directed running in gymnastics: the role of the springboard position as an informational source to regulate handsprings on vault. Biol. Sport 2011;28:215-221.
11. Hraski Ž. Valorizacija suđenja u muškoj sportskoj gimnastici / Valorization of judging in artistic gymnastics (in Croatian). Kineziologija 1988;20:143-152.
12. Leskošek B., Čuk I., Karacsony I., Pajek J., Bučar M. Reliability and validity of judging in men’s artistic gymnastics at the 2009 University Games. Sci. Gymnastics J. 2010;2:25-34.
13. Marinšek M., Čuk I. Landing errors in the men’s floor exercise are caused by flight characteristics. Biol. Sport 2010;27:123-128.
14. Morgan H., Rotthoff K. Bias in sequential order judging: Primacy, recency, sequential bias, and difficulty bias (Unpublished manuscript, Fall 2010). Available at: http://pirate.shu.edu/~rotthoku/papers/Order%20Matters.pdf. Accessed 14.10.2011.
15. Plessner H. Expectation biases in gymnastics judging. J. Sport Exerc. Psychol. 1999;21:131-144.
16. Plessner H., Schallies E. Judging the cross on rings: a matter of achieving shape constancy. Appl. Cognitive Psychol. 2005;19:1145-1156.
17. Scheer J.K., Ansorge C.J. Effects of naturally induced judges’ expectations on the ratings of physical performances. Res. Q. 1975;46:463-470.
18. Ste-Marie, D.M. International bias in gymnastic judging: conscious or unconscious influences? Percept. Motor Skill 1996;83:963-975.