INTER-RATER RELIABILITY AND VALIDITY OF SCORING MEN'S INDIVIDUAL TRAMPOLINE ROUTINES AT EUROPEAN CHAMPIONSHIPS 2014

Bojan Leskošek[1], Ivan Čuk[1] & César J.D. Peixoto[2]

[1]Faculty of sport, University of Ljubljana, Slovenia
[2]Faculty of Human Kinetics, University of Lisboa, Portugal

Original article

Abstract

Execution scores of men's individual trampoline routines at the European Championships (EC) 2014 in Guimarães, Portugal were analysed. In total, 66 men competed in the qualifying round. The old, classic format of scoring, by which the execution score is the sum of the scores of individual judges (discarding the lowest and highest scores), was compared with the new format, by which only the median scores of each skill are tripled and then summed for the final score. Execution was found to be the most significant component of the total score, surpassing degree of difficulty and time of flight in both routines. Intra-class correlation (ICC) coefficients and Kendall's coefficient of concordance W were computed. The bias of judging was small with only one judge found who scored significantly higher than the other judges did. Inter-rater reliability was found good for single skills (ICC around .9 and Kendall W around .7), while for the sum of all ten skills it was excellent (all ICC coefficients above .99 and Kendal W above .97) for both routines. Although the correlation coefficients between old and new format scores were high (r=.965 and r=.997 for first and second routine, respectively), there were some substantial differences in rankings of competitors between old and new scoring format (Spearman rank correlation rho=.94 and rho=.96 for first and second routines, respectively). Despite the reliability and validity of judging trampoline routines were high, some possible means of improvement are suggested. Regarding the differences between old and new formats, no clear (dis)advantages of one or another were found.

Keywords: *trampoline, judging, accuracy, objectivity.*

INTRODUCTION

Trampolining is a well-known sport, especially individual trampoline, which was accepted into the 2000 Summer Olympic Games as one of several gymnastic disciplines. The competition usually consists of two qualification routines and one final (voluntary) routine, each consisting of ten different skills (jumps).

The performance of each routine is the sum of three components: degree of difficulty (DD, also called *tariff*), execution (form), and time of flight (TOF). TOF is

objectively measured with a time measurement device, while the other two components are evaluated by judges. Evaluating the DD of a routine is usually less problematic, as competitors must announce the difficulty of their routines in advance (usually 2 hours before the competition starts) and the D-judges may check the official video recording of a routine in the case any deviation between scores of D-judges and the supervisor of the Technical Committee.

The most difficult part of evaluating performance is evaluating execution, as judges (usually five of them) generally disagree in their deductions (in the [.0, .5] range) which they give for the mistakes (e.g., poor form, incomplete moves, and moving too far from the trampoline's centre mark) in each skill. To resolve this disagreement, two formats are possible: the score of each skill is the sum of the middle (eliminating the highest and the lowest) three judges' scores or *tripled median score* (eliminating the two highest and two lowest scores). The tripled median score was introduced at the 2014 European Trampoline Championships, which also introduced the summing of each single skill score instead of the sum of 10 skills' score of (middle three) judges, which was the usual format in previous competitions.

Several studies were carried out to evaluate judges' performance in different gymnastics disciplines since the 1950s. These studies are rare in trampolining, but are more common in some other gymnastics disciplines, especially in artistic gymnastics. The studies mostly deal with bias and the reliability of judging. In respect to bias, different types of bias were detected. Several authors (Ansorge & Scheer, 1988; Leskošek, Čuk, Pajek, Forbes, & Bučar-Pajek, 2012; Scheer & Ansorge, 1975) found (inter)national bias, i.e., higher scoring of gymnasts from judges' countries and lower scoring of all others or just the closest competitors. A similar type, home advantage bias, was also proven for the 1896-1996 Olympic games (Balmer, Nevill, & Williams, 2003). Others (Bučar, Čuk, Pajek, Karacsony, & Leskošek, 2012; Leskošek, Čuk, Karácsony, Pajek, & Bučar, 2010; Leskošek et al., 2012) found substantial overall judge's bias, i.e., systematic under- or over-scoring of judges. Another bias was found based on the position of judge in accordance with the apparatus (Plessner & Schallies, 2005).

Several authors reported sequential order bias (Ansorge, Scheer, Laub, & Howard, 1978; Damisch, Mussweiler, & Plessner, 2006; Morgan & Rotthoff, 2014; Plessner, 1999) and open feedback / conformity bias (Boen, Van Hoye, Vanden Auweele, Feys, & Smits, 2008). Conformity bias was also found in one of the rare studies specifically dedicated to officiating in trampolining (Johns & James, 2013), in which it was found that the differences between scores in real-time competition and the scores given in post-event video analysis could be high and were causing several and large differences in rankings of competitors, even in medal positions. Authors attributed those differences to social conformity, as well as poor arithmetic skills (when calculating results in real-time under time pressure) and suggested a remedy in using technical equipment (computers and video) to calculate scores and to check for possible deductions both in real competitions and in training courses for the judges.

In addition to bias as a systematic source of errors many studies have addressed unreliability, a random source of errors in judging. Most of these studies are focused on *inter*-rater reliability, i.e., differences in scores between several judges consisting judges' panel, with each member of the panel giving a score to the group of same competitors. This kind of reliability is usually measured by intra-class correlation coefficients (ICC), which may evaluate performance of only one (i.e., "typical") judge (so-called single ICC or single measure ICC) or the whole panel of judges (average ICC or simply ICC). The reliability of the panel of judges may also be evaluated non-parametrically by Kendall coefficient of concordance (W), which is computed on the ranks of competitors (not on the original

scores). There are no known studies of *inter*-rater reliability in trampolining, although regarding intra-rater reliability, i.e., consistent scoring of routines given by the *same judge* at different times, one study (Johns & James, 2013) found excellent reliability. Recent studies (Bučar et al., 2012; Leskošek et al., 2010) in artistic gymnastics reveal good inter-rater reliability with ICC around .95 in qualifying rounds of competitions; however, in apparatus finals several ICCs were much lower, going as low as .72 (in women's vault finals). The reason for this was probably low variability (small differences) in scores between competitors in the final round, which may be much lower than variability in the qualifying round, and the well-documented fact (Shrout, 1998) that low levels of between-subject variability causing depression of the ICC coefficients, even if the differences between judges' scores across the same competitor are small.

The aims of this study were to analyse the men's qualifying round of European Championships (EC) 2014 in Guimarães, Portugal with respect to: (a) importance of routine execution in relation to other components of total score (DD, TOF); (b) quality of judging, especially inter-rater reliability and validity (bias); (c) comparison of execution scores and ranks of competitors given within old (middle three skill deductions count) and new (only median score counts) format of scoring.

METHODS

The initial sample consists of all 66 men competing in qualifying round of European Championships 2014 in Guimarães, Portugal. The competition was organised according to FIG Code of Points 2013-2016. Scores of all competitors, i.e., including those 4 and 15 competitors who did not complete all 10 skills in their first and second routines, respectively, were considered for the data analysis.

Official result sheets from the European Union of Gymnastics (UEG) were collected. New format scores (i.e., sum of tripled median score of each skill) and old format scores (sum of middle 3 judges' sum of scores for all 10 skills) were computed. In addition to deductions made by each of the 5 judges for the execution of each of the 10 skills, DR (reception) deductions were also analysed, while DA (additional) deductions, which are given only by the chair of the judges' panel, were excluded from the analysis, as it was not possible to establish intra-rater reliability in this case.

In the preliminary study, the importance of different components of total scores (i.e., execution, degree of difficulty and time of flight) were established by multiple linear regression.

To access intra-rater reliability, Kendall's coefficient of concordance *W* and intra-class correlation (*ICC*) coefficients were computed. ICC coefficients were evaluated under the two-way random model, both for consistency (ICC_C) and agreement (ICC_A). If not otherwise noted, only ICC_A coefficients were reported and interpreted. Under the agreement model, standard error of measurement (*SEM*) was computed as $SD \times (1 - ICC_A)^{1/2}$ and *minimal differences needed to be considered real* (*MD*) as $MD = SEM \times 1.96 \times 2^{1/2}$ (Weir, 2005).

The bias of judging was evaluated parametrically using the repeated measures ANOVA (RANOVA) *F*-test and non-parametrically using the Friedman test. Effect sizes in these two tests were evaluated by partial eta-squared and Kendall's W coefficient, respectively.

Agreement between old and new format in final scores was accessed using Pearson (product-moment) correlation coefficients *r*, while agreement between competitors' final rankings was accessed by Spearman rank correlation ρ_S.

All analyses were separated for first (i.e., includes special requirements as required by Code of Points) and second (voluntary) routines.

All analyses were carried out with *IBM SPSS Statistics Version 23* software package and R library IRR (Gamer, Lemon, Fellows, & Singh, 2012) and BlandAltmanLeh

(https://CRAN.R-project.org/package=BlandAltmanLeh).

RESULTS

Execution was found to be the most important part of the total score, followed by the time of flight and degree of difficulty (Figure 1). Although this order is valid both in first and second routines, the differences were much more expressed in the first routine, in which the difficulty matters only in the last two of ten skills.

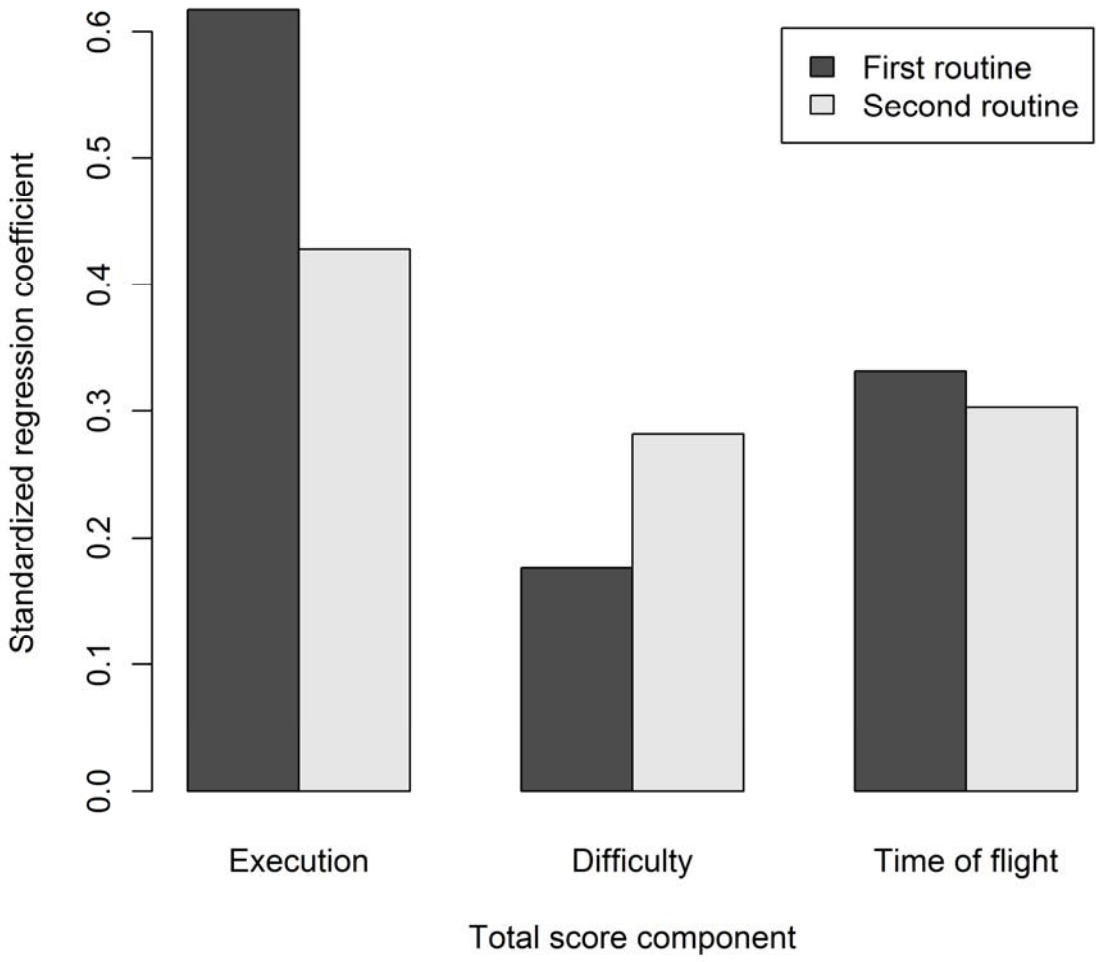

Figure 1. Relative importance of different components of trampoline total score in compulsory and voluntary routines in male individual qualifying round at the 2014 Trampoline European Championships.

Table 1
Statistics related to bias of judging: average deduction by judge, repeated measures ANOVA and Friedman test.

		Average deduction (points) – Judge No.					Repeated measures ANOVA			Friedman test		
Routine	n	1	2	3	4	5	*F*	*p*	$\eta^2_{part.}$	χ^2	*p*	*W*
1	66	1.35	1.34	1.32	1.34	1.38	4.88	.001	.070	15.01	.005	.06
2	66	2.22	2.22	2.23	2.22	2.25	1.56	.196	.023	18.72	.001	.07

Legend. *n*=number of competitors; $\eta^2_{part.}$=partial eta-squared; *W*=Kendall's coefficient (as a measure of effect size in Friedman test).

Table 2
Intra-rater reliability statistics for single skills.

				Intra-class correlation coefficient (*ICC*), value and 95% CI							
				Consistency model				Agreement model			
Routine	Skill	*n*	*W*	Single		Average		Single		Average	
1	1	66	.72	.68	[.59, .77]	.92	[.88, .94]	.67	[.57, .76]	.91	[.87, .94]
1	2	66	.66	.65	[.55, .74]	.90	[.86, .94]	.62	[.51, .72]	.89	[.84, .93]
1	3	66	.68	.68	[.59, .77]	.92	[.88, .94]	.66	[.56, .75]	.91	[.86, .94]
1	4	66	.66	.61	[.51, .71]	.89	[.84, .93]	.60	[.50, .70]	.88	[.83, .92]
1	5	66	.64	.62	[.52, .72]	.89	[.85, .93]	.62	[.51, .72]	.89	[.84, .93]
1	6	66	.66	.59	[.49, .70]	.88	[.83, .92]	.57	[.46, .68]	.87	[.81, .91]
1	7	65	.73	.67	[.57, .76]	.91	[.87, .94]	.66	[.56, .75]	.91	[.86, .94]
1	8	65	.69	.61	[.51, .71]	.89	[.84, .92]	.59	[.47, .69]	.88	[.82, .92]
1	9	63	.68	.59	[.48, .69]	.88	[.82, .92]	.58	[.47, .68]	.87	[.81, .92]
1	10	62	.71	.62	[.52, .72]	.89	[.84, .93]	.59	[.48, .70]	.88	[.82, .92]
1	DR†	66	.77	.96	[.94, .97]	.99	[.99, .99]	.96	[.94, .97]	.99	[.99, .99]
2	1	66	.68	.61	[.51, .71]	.89	[.84, .92]	.60	[.50, .70]	.88	[.83, .92]
2	2	63	.72	.64	[.54, .74]	.90	[.86, .93]	.63	[.52, .73]	.89	[.85, .93]
2	3	61	.74	.68	[.58, .77]	.91	[.88, .94]	.68	[.58, .77]	.91	[.87, .94]
2	4	56	.74	.69	[.59, .78]	.92	[.88, .95]	.67	[.57, .77]	.91	[.87, .94]
2	5	56	.74	.66	[.55, .76]	.91	[.86, .94]	.63	[.51, .74]	.89	[.84, .93]
2	6	55	.76	.67	[.56, .76]	.91	[.86, .94]	.64	[.53, .75]	.90	[.85, .94]
2	7	52	.67	.60	[.49, .72]	.88	[.83, .93]	.60	[.48, .71]	.88	[.82, .92]
2	8	52	.67	.63	[.52, .74]	.90	[.84, .93]	.63	[.51, .74]	.89	[.84, .93]
2	9	52	.70	.61	[.49, .72]	.89	[.83, .93]	.59	[.47, .71]	.88	[.81, .92]
2	10	51	.76	.71	[.61, .80]	.93	[.89, .95]	.69	[.58, .79]	.92	[.88, .95]
2	DR†	66	.72	.96	[.94, .97]	.99	[.99, .99]	.96	[.94, .97]	.99	[.99, .99]

Legend.. *n* = number of competitors; *W* = Kendall's coefficient of concordance; CI = confidence interval.
† DR reception deductions

Table 3
Intra-rater reliability statistics for total execution scores (sum of scores for all skills).

				Intra-class correlation coefficient (*ICC*), value and 95% CI							
				Consistency model				Agreement model			
Complete?	Routine	*n*	*W*	Single		Average		Single		Average	
Yes[†]	1	62	.977	.981	[.973, .988]	.996	[.994, .998]	.980	[.971, .987]	.996	[.994, .997]
Yes	2	51	.984	.986	[.978, .991]	.997	[.996, .998]	.983	[.973, .990]	.997	[.995, .998]
No	1	66	.977	.982	[.974, .988]	.996	[.995, .998]	.981	[.972, .987]	.996	[.994, .997]
No	2	66	.991	.990	[.986, .993]	.998	[.997, .999]	.990	[.985, .993]	.998	[.997, .999]

Legend. *n* = number of competitors; *W* = Kendall's coefficient of concordance; CI = confidence interval.
[†] 'Yes' means, that only those competitors, who finished all ten skills, were included in the computation.

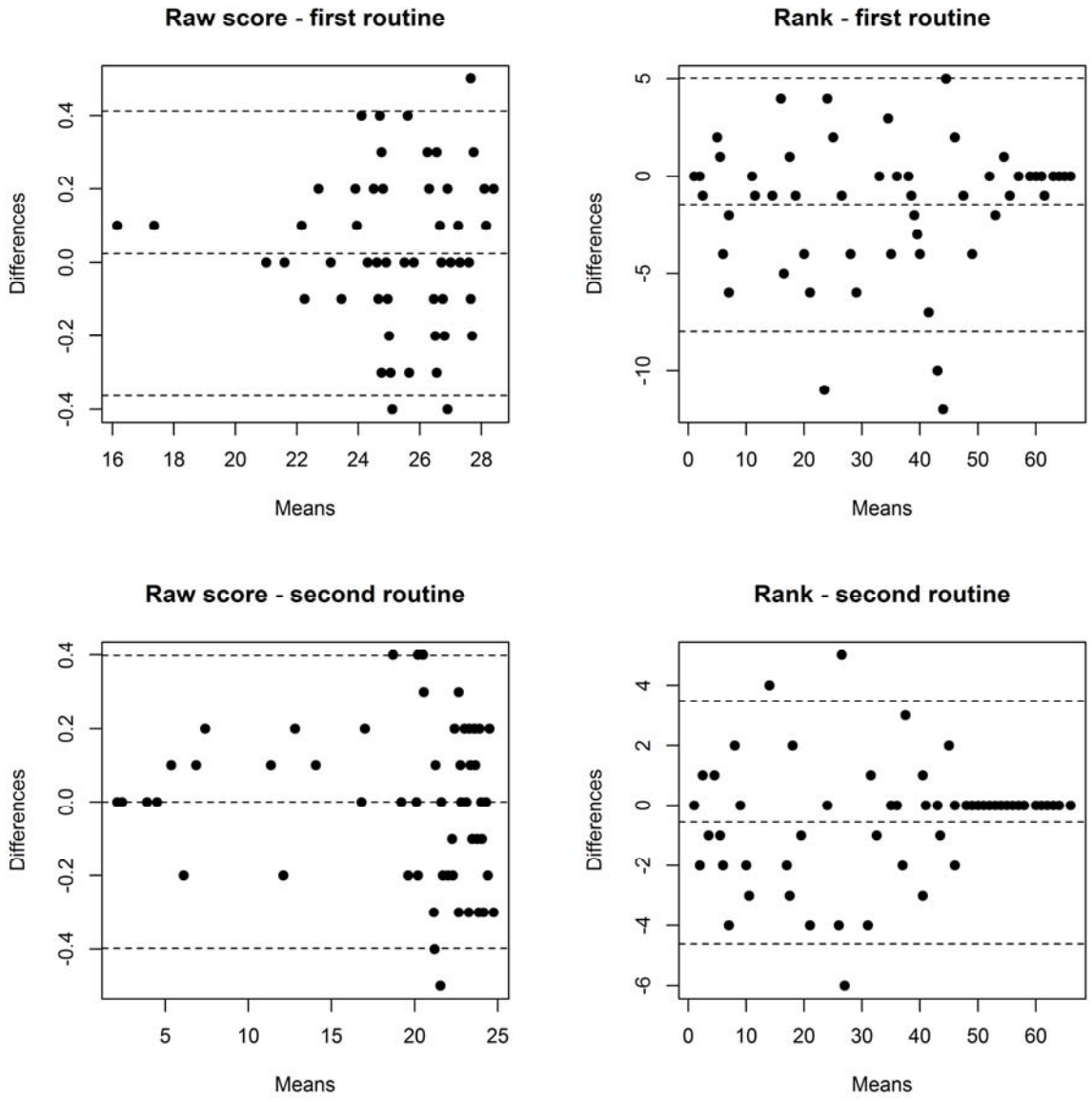

Figure 2. Bland-Altman plot for the differences in raw scores and ranks of execution scores in first and second routines.

Statistics related to bias, i.e., *systematic* under- or overestimation of some judges (Table 1), reveal a small but statistically significant bias of judging, except for the *F*-test in RANOVA of the second routine (p = .196). The most notable bias was found for Judge No. 5, whose deductions in both routines are higher than in other judges.

Intra-rater reliability statistics for *single* skills (Table 2) show moderate and statistically significant (in all cases $p < .001$) agreement between judges. In both routines, Kendall's coefficients of concordance W were around .70. The ICC coefficients for single judges under the consistency model were similar to the W coefficients, while the average ICCs (for all 5 judges) were around .90. Under the agreement model, the ICC coefficients were only slightly lower than under the consistency model.

Intra-rater reliability statistics for the sum of execution scores of *all* skills (Table 3) show very high agreement between judges. This agreement is somewhat higher in the second routine, where deductions for execution are generally almost one full point higher than in the first routine (see first 5 columns of Table 1) and are also much more variable (coefficient of variation CV= 35.3% vs, CV=9% in the first routine). Under the agreement model, ICC coefficients were only slightly lower than under the consistency model.

The standard error of measurement (SEM) computed under agreement model was .037 and .043 in the first and second routines, respectively. Minimal differences needed to be considered real (MD) was .102 and .120 in the first and second routines, respectively.

Execution scores as computed by new and old format scoring are generally different both in raw scores (points), as well as ranks of these scores (Figure 2) In raw scores (left two plots in Figure 2), differences between new and old format scores in the first and second routines range from –.4 to +.5, and –.5 to +.4 points, respectively. In the ranks of execution scores (right two plots in Figure 2), differences between the new and old format ranks in first and second routines range from –12 to +5, and –6 and +5, respectively. Tied ranks are much more frequent with the new format, e.g., in the first routine there are nine competitors tied for 38th place and seven competitors tied for 9th place in the second routine, while in the old format the maximum number of tied competitors are 5 and 3 in the first and second routines, respectively. Ties in the new format are especially frequent for competitors with highest execution scores.

The highest differences between the old and new formats usually occur when there are several disagreements between groups of two and three judges, e.g., French competitor A. M. received in four skills of his first routine a .1 deduction from two judges and no deduction from other three judges, resulting (together with differences in some other skills) in a +.5-point higher score in the new format compared to the old one.

DISCUSSION

Execution scores of the 2014 Trampoline European Championships were analysed; only the qualifying round of male individual trampoline were included in the analysis, as this was the round and discipline with the most (66) competitors in the event; therefore, the most valid results can be expected in this case.

Execution was found to be a much more important component of total score than time of flight and degree of difficulty (Figure 1). While this might be expected in the first routine, where difficulty only matters in the last two of ten skills, it is very informative in the second routine, where execution is substantially more important than difficulty; all ten skills' difficulty count in the second routine. This finding may have two important consequences: first, the execution of routines should be in the primary focus of athletes when preparing for competitions and second, all trampoline federations from the local (regional) to the global level should ensure that the judging

at competitions is as fair as possible, both with regards to the reliability and the unbiasedness of judging.

Unbiasedness (objectivity, validity) of judging seems to be a minor problem in execution scoring in this discipline of the 2014 European Championships. Although three of four tests (one parametric and two nonparametric) showed statistically significant bias, the bias was low and mostly attributable only to one of five judges, who tended to make higher deductions than other the four judges in both the first and second routines. This type of bias, i.e., over- or underestimation in scoring, may not be a major problem that would jeopardise the fairness of the competition results if the bias of a single judge is persistent in all (or at least the majority) of competitors, at it seems was the case in the competition. Namely, extreme scores (four in each skill in new format and two extreme sums of judges' scores in old format) are excluded from the execution score. However, if deductions of one or even more judges are excluded from the execution score in the majority of cases, it may raise questions about the reliability of judging, as (according to the classical test theory and the Spearman-Brown formula (Weir, 2005)) a lower number of judges means lower reliability. Therefore, it is important for the bodies governing the quality of judging to ensure consistent, harmonised criteria of judging and implement mechanisms to educate, check, inform and penalise, if necessary, the judges who persistently deviate from other judges.

Bias at the 2014 European Championships was similar to that found at qualification round of the 2011 European Championships in men's artistic gymnastics (Leskošek et al., 2012), were Kendall's W coefficients were between .01 and .11, with four of six apparatus' coefficients (all but vault and parallel bars) being statistically significant.

Intra-rater reliability was found high for single skills and very high for the sum of all skills, both in the first and second routines and with all statistics used (Kendall coefficient of concordance W, ICC coefficients under consistency and agreement model). Compared to recent research findings in other gymnastics disciplines and artistic gymnastics (Bučar et al., 2012; Leskošek et al., 2010), it seems that reliability in trampolining with ICC coefficients above .99 is (much) better than in artistic gymnastics, where ICC coefficients rarely exceed .98 and may be lower than .95, even if there are similar numbers of judges (4 to 6, compared to 5 in trampolining). The factors that influence higher reliability in trampolining vs. artistic gymnastics may be: higher duration of each skill in trampolining (take-off, flight, and landing takes around two seconds) compared to artistic gymnastics (some elements may take just a fraction of a second); in trampolining, the athlete's body is in the air all the time, not obstructed by an apparatus and is well visible for all judges from the raised judges' platform; and, most skills are performed with several rotations in different planes, so even if the judges are on the different position on the platform, any lack of form of the athlete's body is more likely to be seen by all of the five execution judges.

High reliability coefficients do not necessarily mean there is no room for improvement. Especially in single skills, many disagreements (and, therefore, relatively low reliability) may be seen.

Somewhat lower reliability in the first than in the second routine may be expected, as in the first routine execution deductions were much lower than in the second routine (Table 1) and, therefore, also have lower variability (coefficients of variation were 9.0% and 35.3%, in the first and second routines, respectively), which in turn depresses reliability coefficients (Shrout, 1998). Similarly, lower reliability may be expected in the most important, decisive final round of competition (not analysed in this study), in which differences in scores between competitors are usually smaller than in the qualification round (Bučar et al., 2012).

Following high reliability, standard errors of measurement (SEM) were low, i.e., .037 and .043 points in the first and second routines, respectively. However, even with small SEM, minimal differences needed to be considered real (MD), and are higher than .1 points in both routines, which may cause some unfair rankings of competitors.

As expected, the total scores computed under the new format are different from those computed under the old format. Generally, these differences are small and never exceed +/- .5 point. However, even these small differences in scores may produce big differences in rankings (computed only for execution, while excluding difficulty and TOF). In both routines, differences in rankings were even higher than 10 ranks (places) in some cases, which may also, of course, produce differences in final rankings (including difficulty and TOF). As there is no golden standard for evaluating execution, there are no means to say which format is better or more accurate. However, there are two possible problems, which may speak against new format. The first one is that new format produces many more tied scores (and ranks); however, these ties may be split on the basis of difficulty or TOF. The second possible problem is that many different deductions of single judges produce the same deductions in execution scoring. For example, all the following deductions (in tenths of point) for the five judges (0, 0, 1, 1, 1), (1, 1, 1, 1, 1) and (1, 1, 1, 2, 2), which were quite frequent in that competition, resulted in the same .3 points deduction in the new format, but very different deductions in the old format, namely .2, .3 and .4 points, which may seem more realistic (fair), especially if several large differences between the old and new scoring in different skills of the same competitor exist.

CONCLUSIONS AND RECOMMENDATIONS

The execution score was found to be the most important component of success in trampolining, at least in the qualifying round of this competition, surpassing both degrees of difficulty and TOF in both the first and second routines. Therefore, fair evaluation (judging) of execution is of paramount importance for the fair ranking of competitors. In both aspects of quality of judging, i.e., validity and reliability, trampolining was found very good, even better than in artistic gymnastics, a gymnastic discipline with the longest tradition. However, even if the quality of judging was generally high, small flaws that were found in some cases may jeopardise fair scoring and rankings; therefore, maintaining the high quality of judging is vital. This may be accomplished by judges' education and selection, evaluation, as well as penalising, when necessary.

Differences in scoring in the new and old formats were generally small, so it may be expected that they only produce rare and small differences in (execution and total) rankings of competitors. Although it seems it does not matter much which format to use in the future, some subtle differences were addressed (fewer ties, probably fairer scores when the five judges disagree) that may be in favour of the old format. However, the scoring of each skill in the new format, which replaces sums of scores for all ten routines in the old format, may speak in favour of the new format, as it should reduce social conformity bias in the old format.

Although the overall quality of judging was good both in terms of reliability and validity, that does not mean there is no room for improvement. One opportunity for even better judging is the integration of video and computers into the real-time judging, as well as judges' education and monitoring. Currently, video is not used in real-time by E-judges. As several other studies have shown, video and scoring machines in different sport disciplines (e.g., trampolining (Johns & James, 2013), boxing (Di Felice & Marcora, 2013) and artistic gymnastics (Pajek, Forbes, Pajek, Leskošek, & Čuk, 2011)) may improve reliability and

reduce conformity bias and arithmetic errors in the scoring of athletes' performance.

REFERENCES

Ansorge, C. J., & Scheer, J. K. (1988). International bias detected in judging gymnastic competition at the 1984 Olympic Games. *Research quarterly for exercise and sport, 59*(2), 103-107.

Ansorge, C. J., Scheer, J. K., Laub, J., & Howard, J. (1978). Bias in judging women's gymnastics induced by expectations of within-team order. *Research Quarterly. American Alliance for Health, Physical Education and Recreation, 49*(4), 399-405.

Balmer, N. J., Nevill, A. M., & Williams, A. M. (2003). Modelling home advantage in the Summer Olympic Games. *Journal of Sports Sciences, 21*(6), 469-478.

Boen, F., Van Hoye, K., Vanden Auweele, Y., 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.

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.

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.

Di Felice, U., & Marcora, S. (2013). Errors in judging Olympic boxing performance: False negative or false positive? In Peters, D. M. & P. O'Donoghue (Eds.), *Performance Analysis of Sport IX* (pp. 190-195): Routledge.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84. *Internet resource: https://cran.r-project.org/web/packages/irr/index.html*, 2017

Johns, P., & James, B. (2013). The efficacy of judging within trampolining. In D. M. Peters & P. O'Donoghue (Eds.), *Performance Analysis of Sport IX* (pp. 214-221): Routledge.

Leskošek, B., Čuk, I., Karácsony, I., Pajek, J., & Bučar, M. (2010). Reliability and validity of judging in men's artistic gymnastics at the 2009 university games. *Science of Gymnastics Journal, 2*(1), 25-34.

Leskošek, B., Čuk, I., Pajek, J., Forbes, W., & Bučar-Pajek, M. (2012). Bias of judging in men's artistic gymnastics at the european championship 2011. *Biology of Sport, 29*(2), 107.

Morgan, H. N., & Rotthoff, K. W. (2014). The harder the task, the higher the score: Findings of a difficulty bias. *Economic Inquiry, 52*(3), 1014-1026.

Pajek, M. B., Forbes, W., Pajek, J., Leskošek, B., & Čuk, I. (2011). Reliability of real time judging system. *Science of Gymnastics Journal, 3*(2), 47-54.

Plessner, H. (1999). Expectation biases in gymnastics judging. *Journal of Sport and Exercise Psychology, 21*(2), 131-144.

Plessner, H., & Schallies, E. (2005). Judging the cross on rings: A matter of achieving shape constancy. *Applied Cognitive Psychology, 19*(9), 1145-1156.

Scheer, J. K., & Ansorge, C. J. (1975). Effects of naturally induced judges' expectations on the ratings of physical performances. *Research Quarterly. American Alliance for Health, Physical Education and Recreation, 46*(4), 463-470.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical methods in medical research, 7*(3), 301-317.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength and Conditioning Research, 19*(1), 231-240.

Corresponding author:

Bojan Leskošek
University of Ljubljana, Faculty of sport
Gortanova 22
1000 Ljubljana
Slovenia
tel. +386 1 520-77-00
fax +386 1 520-77-40
e-mail: bojan.leskosek@fsp.uni-lj.si