Journal of Human Kinetics volume 60/2017, 159-165 DOI: 10.1515/hukin-2017-0099
Section III – Sports Training

DE GRUYTER OPEN

Judging in Rhythmic Gymnastics at Different Levels of Performance

by
Catarina Leandro[1,3], Lurdes Ávila-Carvalho[2], Elena Sierra-Palmeiro[3], Marta Bobo-Arce[3]

This study aimed to analyse the quality of difficulty judging in rhythmic gymnastics, at different levels of performance. The sample consisted of 1152 difficulty scores concerning 288 individual routines, performed in the World Championships in 2013. The data were analysed using the mean absolute judge deviation from the final difficulty score, a Cronbach's alpha coefficient and intra-class correlations, for consistency and reliability assessment. For validity assessment, mean deviations of judges' difficulty scores, the Kendall´s coefficient of concordance W and ANOVA eta-squared values were calculated. Overall, the results in terms of consistency (Cronbach's alpha mostly above 0.90) and reliability (intra-class correlations for single and average measures above 0.70 and 0.90, respectively) were satisfactory, in the first and third parts of the ranking on all apparatus. The medium level gymnasts, those in the second part of the ranking, had inferior reliability indices and highest score dispersion. In this part, the minimum of corrected item-total correlation of individual judges was 0.55, with most values well below, and the matrix for between-judge correlations identified remarkable inferior correlations. These findings suggest that the quality of difficulty judging in rhythmic gymnastics may be compromised at certain levels of performance. In future, special attention should be paid to the judging analysis of the medium level gymnasts, as well as the Code of Points applicability at this level.

Key words: *rhythmic gymnastics, evaluation, bias, validity, reliability.*

Introduction

In artistic sports like rhythmic gymnastics (RG), the performance in competition is evaluated by judges that apply a tool (Code of Points) and give a score that determines the value of the routine and the position of the gymnast in the final ranking. Since the performance does not come out from an objective measure, but from a complex judging process, quite often RG is considered to be a subjective sport (Gateva, 2014).

Recent research has paid attention mainly to the experience and the capacity of the judges to use cognitive and perceptual strategies to interpret and register gymnast's performance in competition (Dallas and Kirialanis, 2010; Heinen et al., 2012; Plessner and Schallies, 2005; St. Marie et al., 2001). Furthermore, research has also emphasized the judges' need for developing a set of skills that contributes to an effective assessment process (Fernandez-Villarino et al., 2013), and to the overall error detection efficiency (Flessas et al., 2015). In a RG competition, the performance is evaluated by two panels of judges: the difficulty (D) jury which judges the routines' content (what the gymnast performs) and the execution (E) jury which evaluates the quality of the routines (how the gymnast performs). The present Code of Points states that minimum four judges are required on the D jury, as well as on the E jury.

[1] - *Faculty of Psychology, Education and Sport, University Lusófona of Porto, Portugal.*
[2] - *Sports Faculty, University of Porto, Portugal.*
[3] - *Faculty of sport Science and Physical Education, University of Coruña, Spain.*

Authors submitted their contribution to the article to the editorial board.
Accepted for printing in the Journal of Human Kinetics vol. 60/2017 in December 2017.

For both, the final score is determined calculating the average of two intermediate scores (FIG, 2012).

The judging process for difficulty and execution evaluation is different. Difficulty judges have to check the content of the routines that is stated and signed by the coaches in the specific forms. Their task is to validate the difficulty elements declared while the gymnast performs her routine. These difficulty elements may range from 0.1 to 1.5 points or more, up to a total maximum of 10 points. Preciseness in the judgement is needed since differences between the judges may cause great deviations in the final D score, and this score has a great influence on the gymnast's final position in the ranking (Cuk et al., 2012; Leskosek et al., 2015).

In higher level competitions, the more experienced judges are assigned to evaluate the difficulty component of the routine. However, quite often the differences between the athletes' performances are so small, that little and consistent mistakes made by the judges may interfere in the final classification of the gymnast (Bucar et al., 2011, 2013). Consequently, to verify the quality of judging in rhythmic gymnastics, it is necessary to identify the extent to which the scoring system is objective. Therefore, reliability and validity of the scores must be verified.

The aim of this study was to analyze the reliability and validity of the difficulty scores of individual routines in RG at different levels of performance and with different apparatus. It was hypothesized that the level of performance of the gymnast as well as the type of apparatus used (hoop, ball, clubs and ribbon) may affect the reliability and validity of the scores in competition.

Methods

Participants

The sample consisted of 1152 difficulty scores corresponding to 288 exercises performed at the Kiev World Championships in 2013, clustered according to the position of the gymnast in the final ranking (1st part, 2nd part and 3rd part) and to the apparatus (hoop, ball, clubs and ribbon). The scores were obtained from the official book of results of the qualification competition. The study was ethically approved by the International Gymnastics Federation. Full blinding of the judges involved was undertaken. To protect the judges' anonymity, we randomly changed their position in the analysis from the book of results.

Procedures

The sample was divided into three groups according to the gymnast's final ranking considering each apparatus: the first part of the ranking (top 24 gymnasts), the second part of the ranking (medium 24 gymnasts) and the third part of the ranking (last 24 gymnasts), to allow the comparison of the reliability and validity values at different performance levels with all 4 apparatuses. For each of the groups, four judges' D scores were considered.

Statistical analysis

For each group (top, medium and last gymnasts), descriptive statistics for the D score were calculated, as well the distributional statistics (mean and standard deviation) for an individual judge's D score and mean deviation from the final D score. This mean deviation is a measure of bias (systematic under- or over-estimation) and provides information related to the validity of scoring. When examining validity, the ideal test of validity would have to implement a comparison of concrete judging with the gold standard of judging performance (Bučar et al., 2011); however, no such a gold standard currently exists. It is possible, however, to focus on a special case of validity, which deals with the presence of systematic over- or underrating of scoring of competitors, what is also called bias.

Additionally, two analyses of between-judges differences were performed including the Kendall's concordance coefficient and repeated measures ANOVA to identify possible systematic bias.

The correlation between individual judge's scores and total scores was also calculated.

The consistency and reliability assessment of the evaluation were measured using the Cronbach's alpha coefficient for each group of judges on each apparatus. Two types of intra-class correlation (ICC) were calculated: the single measure ICC and the average measures ICC.

Data were analyzed using the Statistical Package for the Social Sciences - Version 21.0 (SPSS 21.0, Chicago, USA) and Microsoft Office Excel 2010.

Results

The variability of D scores (dispersion) was in general larger for the 2nd part of the ranking and relatively smaller in the 1st part of the ranking, in the hoop, ball, clubs and ribbon (Table 1). The average value of the D score for different apparatuses did not show great variability in each part of the ranking. The worst individual deviations in judging for each part of the ranking and apparatus (all remaining individual judge values were better) were presented. Besides the worst deviations, also the smallest values for item-total correlation were indicated as well as the Cronbach's alpha coefficient for each apparatus. It can be observed that maximal individual judge mean deviations from the final D score were overall relatively small, all of them below 0.2. Only in the 3rd part of the ranking in the hoop and clubs, we found maximum deviations with values of 0.38 and 0.33, respectively. In terms of measures of common performance for the 2nd part of the ranking in all apparatuses, we obtained the poorest values of Cronbach's alpha and the smallest values of minimum item-total correlation. However, most of the values were still above 0.9 in the 1st and 3rd part of the ranking. We did not observe significant differences between the different apparatuses.

When testing the inter-judge differences with repeated measures ANOVA, we can observe eta-squared values. These values represented the bias effect size (Figure 1), and were quite concordant with Kendall`s results. Kendall's W was statistically significant in the 1st part of the ranking for the hoop and clubs, in the 2nd part of the ranking for the ball and in the 3rd part of the ranking for the hoop.

Regarding the analysis of between-judge correlations, the Pearson's correlation coefficients are shown in Table 2. It can be seen that most of the correlation coefficients are above 0.7 in the 1st and 3rd parts of the ranking, while in the 2nd part of the ranking a high number of correlations are below 0.5.

Table 1

Statistics of D scores and the performance of individual judges

	Apparatus	Mean ± SD	Dev. max.	Ab. dev. max.	R min.	Cα
1st part of the ranking	Hoop	8.25 ± 0.53	-0.15	0.29	0.71	0.91
	Ball	8.34 ± 0.49	-0.03	0.20	0.76	0.91
	Ribbon	7.98 ± 0.60	-0.12	0.26	0.75	0.92
	Clubs	8.21 ± 0.55	-0.20	0.29	0.76	0.91
2nd part of the ranking	Hoop	6.61 ± 0.45	-0.13	0.41	0.16	0.65
	Ball	6.85 ± 0.60	0.26	0.41	0.47	0.79
	Ribbon	6.52 ± 0.63	-0.08	0.38	0.55	0.77
	Clubs	6.68 ± 0.48	-0.16	0.42	0.24	0.59
3rd part of the ranking	Hoop	4.58 ± 1.31	0.38	0.56	0.82	0.94
	Ball	4.64 ± 1.33	-0.22	0.36	0.89	0.96
	Ribbon	4.36 ± 1.38	-0.10	0.32	0.92	0.97
	Clubs	4.56 ± 1.39	0.33	0.44	0.88	0.95

Minimum (min) and maximum (max) values, mean and standard deviation (SD), Dev. max: maximal judge average deviation from D score, Ab. Dev. Max.: maximum of average absolute deviation from D score; R min: minimum of corrected item-total correlation of individual judges; Cα: Cronbach's alpha coefficient.

Table 2

Correlation Matrix for between - judges correlation

		1st part of the ranking			2nd part of the ranking			3rd part of the ranking		
	Judge	2	3	4	2	3	4	2	3	4
Hoop	1	0.72**	0.76**	0.85**	0.42*	0.15	0.73**	0.86**	0.87**	0.75**
	2		0.48*	0.80*		0.33	0.30		0.85**	0.81**
	3			0.74**			-0.14			0.78**
Ball	1	0.71**	0.80**	0.68**	0.71**	0.49*	0.47*	0.86**	0.83**	0.89**
	2		0.74**	0.79**		0.34	0.55**		0.89**	0.79**
	3			0.61**			0.38			0.86**
Ribbon	1	0.64**	0.79**	0.69**	0.38**	0.44*	0.40*	0.90**	0.91**	0.93**
	2		0.74**	0.82**		0.45*	0.32		0.87**	0.93**
	3			0.87**			0.61**			0.92**
Clubs	1	0.74**	0.83**	0.69**	0.23	0.19	0.12	0.83**	0.83**	0.86**
	2		0.70**	0.71**		0.34	0.36		0.88**	0.83**
	3			0.69**			0.34			0.84**

*** correlation is significant at the 0.01 level (2 - tailed);*
** correlation is significant at the 0.05 level (2 - tailed)*

Table 3

Overall measures of inter-judge reliability

	Apparatus	ICC Single	ICC Average	Kendall´s W	P (W)
1st part of the ranking	Hoop	0.70	0.90	0.133	0.023*
	Ball	0.73	0.91	0.013	0.821
	Ribbon	0.76	0.92	0.053	0.283
	Clubs	0.88	0.89	0.177	0.005*
2nd part of the ranking	Hoop	**0.33**	0.66	0.021	0.671
	Ball	**0.46**	0.77	0.109	0.049*
	Ribbon	**0.47**	0.78	0.002	0.982
	Clubs	**0.26**	0.58	0.071	0.162
3rd part of the ranking	Hoop	0.80	0.94	0.112	0.045*
	Ball	0.84	0.95	0.088	0.096
	Ribbon	0.91	0.97	0.028	0.564
	Clubs	0.83	0.95	0.064	0.203

ICC single (average): intra-class correlation
for single (average) scores; p(w): p value of Kendall´s W

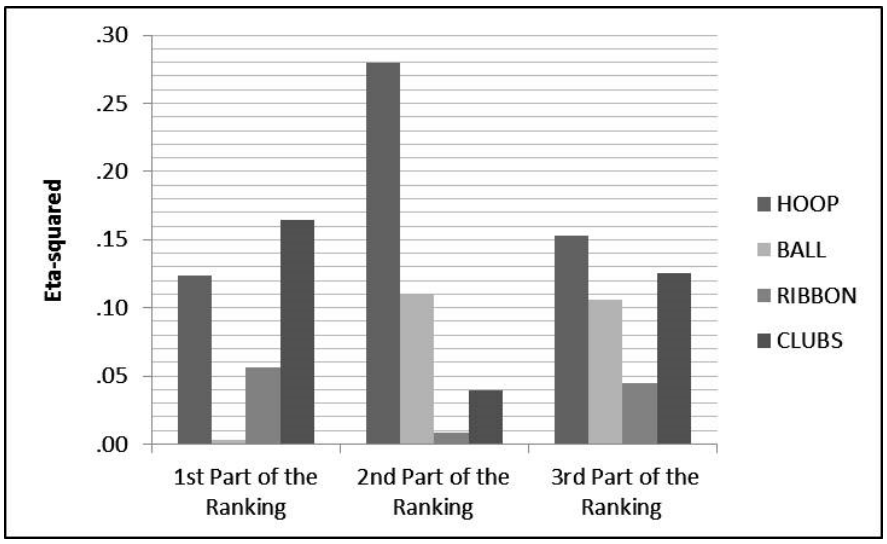

Figure 1

The eta-squared (η^2) values of repeated measures ANOVA of D-scores in all apparatuses clustered according to the position of the gymnast in the final ranking (1st part, 2nd part and 3rd part)

Overall measures of inter-judge reliability are shown in Table 3. The poor concordance of judges on the 2nd part of the ranking (as evident from Cronbach's alpha coefficients) can be also inferred from the calculated ICC of single values, otherwise the observed ICC values are high, mostly above 0.7. The values of ICC for average values are quite close to Cronbach's alpha coefficient values. In the second part of the ranking, the ICC for single values confirmed the highest sensitivity for the deviations in inter-judge agreement and reliability when compared to other measures (Cronbach's alpha and ICC for average measures).

Discussion

The aim of the current study was to analyse the quality of difficulty judging in rhythmic gymnastics at different levels of performance. To the best of our knowledge, this is the first study that has analysed the reliability and validity of difficulty judging, perhaps, because only in this Olympic cycle, the final score was determined by calculating the average of two intermediate scores (of 4 judges) and not for consensus, a joint score of two judges.

Overall, the results suggest that the reliability of the judgment in RG is satisfactory in the first and third parts of the ranking, as the Cronbach's alpha was above 0.90, minima of item total correlations and the ICC of average scores were above 0.80. For the World Championships analyzed, regarding the final ranking of the gymnasts, the indices of consistency were satisfactory in both high and low level gymnasts.

However, the level of consistency indices

was lower in the 2nd part of the ranking. When trying to explain the inferior reliability results for medium level gymnasts, it is valuable to inspect the between-judge correlation matrix, as many of the reliability measures of judges' performances are based on Pearson's correlations. We could identify several judges (without highlighting any over the others) whose correlation coefficients were below 0.5 in all apparatuses.

The validity in our analysis was assessed through systematic bias in judging, considered as repetitive under- or over-estimation of particular judges. When looking at the results as a whole, systematic bias in individual judge's scores and judges' panels was modest or poor in the 2nd part of the ranking. Popovic (2000) also detected international bias in judging rhythmic gymnastics at the Sydney 2000 Olympic Games. It is obvious that the quality of judging differs when evaluating different levels of gymnasts' performances. There are numerous objective and subjective factors for those differences. According to Ferreirinha and Carvalho (2012), besides these external factors that may lead judges to commit mistakes that go further away from the dimension of conscience and therefore, are not intentional, there are other factors related directly to the evaluation rules (Code of Points) that may be in the origin of these deviations. Also Bucar et al. (2014) reached similar results when analysing the evaluation of the artistic component in female gymnastics. Fernandez-Villarino et al. (2013) claim that the specific situation in which the judges must evaluate gymnasts of different ages and different levels during the same competition may create problems in the ability to discriminate performances.

Our results show that the bias in the judgment of rhythmic gymnastics competition routines is not so much due to the performance of specific judges, but more to the differences in the level of performance of the gymnasts at the same competition.

To further clarify the factors contributing to the observed phenomenon, we can speculate that these differences are perhaps a source of additional variability in the judge's scores and that part of the problem may originate in the judging rules (Code of Points) that are not well defined to evaluate the gymnasts as they lack clearness and precision. This situation, according to Debien et al. (2014), may be a source of variability between the judges caused by stress which appears because of the acknowledgment of something which is not expectable. The apparatus used by the gymnast does not seem to be the reason for variability in judging since we found equivalent values of mean deviations calculated from final D scores for all apparatuses in each part of the ranking

In conclusion, the comparison of reliability and validity indices brought our attention to medium level gymnasts, as at this particular level, these indices seemed more sensitive to deviations compared to high reliability indices found in other level gymnasts (1st and 3rd parts of the ranking). Further work is necessary to explain the inferior results at medium level gymnastics and test the solutions for improvement. This study provides updated information about the individual routines judgment in rhythmic gymnastics, to be considered for possible modifications of the present Code of Points, in particular with regard to the definition of evaluation criteria in order to reach higher levels of reliability and validity in judgment.

Acknowledgements

We would like to thank the International Gymnastics Federation (FIG) and the Chairperson of the Technical Judges Committee.

References

Bučar M, Cuk I, Pajek J, Karacsony I, Leskosek B. Reliability and validity of judging in women's artistic gymnastics at University Games 2009. *Eur J Sport Science*, 2011; 12: 207-215

Bučar M, Cuk I, Pajek J, Kovac M, Leskosek B. Is the Quality of Judging in Women Artistic Gymnastics Equvalent at Major Competitions of Different Levels? *J Hum Kinet,* 2013; 37: 173-181

Bučar M, Kovač M, Pajek J, Leskošek B. The Judging of artistry components in female gymnastics: a cause for

concern? *Sci Gymnastics J*, 2014; 6(3): 5-12

Catteeuw P, Helsen W, Gilis B, Wagemans J. Decision-making skills, role specificity, and deliberate practice in association football refereeing. *J Sport Sci*, 2009; 27: 1125-1136

Čuk I, Fink H, Leskošek B. Modeling the final score in Artistic Gymnastics by different weights of difficulty and execution. *Sci Gymnastics J*, 2012; 4: 73-82

Dallas G, Kirialanis P. Judges' evaluation of routines in men's artistic gymnastics. *Sci Gymnastics J*, 2010; 2: 49-58

Debien P, Noce F, Debien J, Costa V. Stress in rhythmic gymnastics refereeing: a systematic review. *Revista da Educação Física/UEM*, 2014; 25 (3): 489-500

Ferreirinha J, Carvalho J. Tendencies and deviations of judging in gymnastics. *Revista ENGym*, 2012; 1: 2-3

Fernandez-Villarino M, Bobo-Arce M, Sierra-Palmeiro E. Practical Skills of Rhythmic Gymnastics Judges. *J Hum Kinet*, 2013; 39: 243-249

FIG-International Gymnastics Federation. Code of Points for Rhythmic Gymnastics Competitions, 2012; Available at: http://www.fig-gymnastics.com/site/page/view?id=472; accessed on 15.11.2012

FIG - International Gymnastics Federation. Gymnastics Results, 2013; Available at: http/www.gymnasticsresults.com; accessed on 01.10.2013

Flessas K, Mylonas D, Panagiotaropoulou G, Tsopani D, Korda A, Siettos C, Di Cagno A, Evdokimidis I, Smyrnis N. Judging the Judges' Performance in Rhythmic Gymnastics. *Med Sci Sports Exer*, 2015; 47(3): 640-648

Gateva M. Investigation of the effect of the training load on the athletes in rhythmic and aesthetic group gymnastics during the preparation period. *Research in Kinesiology*, 2014; 4: 40-44

Heinen T, Vinken P, Velentzas K. Judging Performance In Gymnastics: A matter of motor or visual experience? *Sci Gymnastics J*, 2012; 4: 63-72

Leskošek B, Čuk I, Bučar-Pajek M. Trends in E and D scores and their influence on final results of male gymnasts at European Championships 2005–2011. *Sci Gymnastics J*, 2015; 5(1): 29-38

Plessner H, Schallies E. Judging the cross on rings: a matter of achieving shape constancy. *Appl Cognitive Psych*, 2005; 1: 1145-1156

Popovic R. International bias detected in judging rhythmic gymnastics competition at Sydney 2000 Olympic Games. *Facta Universitatis*, 2000; 1: 1-13

Ste-Marie DM, Valiquette SM, Taylor G. Memory Influenced Biases in Gymnastic Judging Occur Across Different Prior Processing Conditions. *Res Q Exercise Sport*, 2001; 72(4): 420-426

Corresponding author:

Catarina Leandro
Affiliation: Faculty of Psychology, Education and Sport, University Lusófona of Porto, Portugal
Address: Rua Augusto Rosa, nº 24, 4000-098 Porto, Portugal
Phone/mobile: +351917640998
E-mail: catarinaleandro@sapo.pt