Journal of Sports Science 4 (2016) 232-240
doi: 10.17265/2332-7839/2016.04.007

The Evaluation Rules in the View of the Rhythmic Gymnastics Judges

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

1. Faculty of Psychology, Education and Sport, University Lusófona of Porto, Porto 4000-098, Porto, Portugal

2. Faculty of Sports, University of Porto, Porto 4200-450, Porto, Portugal

3. Faculty of sport Science and Physical Education, University of Coruña, Coruña 15179, Oleiros, Spain

Abstract: The Code of Points (CoP) is the evaluation tool of Rhythmic Gymnastics (RG) and contributes to its evolution as sport. This tool is applied in competition by the judges who have a crucial role in the performance evaluation. The aim of this study was to characterize the International judges, their perception of the objectivity of their judgment in competitions and their opinions about the content and the reliability of the Code of Points. 162 international RG judges answered a questionnaire specially designed for this study. For the data analysis, non-parametric tests were used. According to the judges, the evaluation of the difficulty component in the individual routines has more subjectivity in the items of Mastery (58.6%) and Dance Steps (55.6%) and is less subjective in the body difficulties of Balance (72.2%). Concerning the execution, the judges consider that the evaluation of the artistic faults is the most subjective in this domain (64.8%). Within the artistic faults the items unity of the composition and relation between the Music and the Movements, were those which registered higher significant results for the subjectivity in the evaluation (47.5% and 37.0%, respectively).

Key words: Evaluation, objectivity, judge, rhythmic gymnastics.

1. Introduction

Performance the evaluation in RG depends on a judging process done by specialized judges that apply a specific set of rules and procedures established in the official FIG Code of Poins (CoP) [1]. The CoP is used as an evaluation instrument by the judges, who have their own technical, human and social experiences. All these aspects take part in the judging process and from it depends, in part, the improvement of quality of the sports practice, the safeness of the physical and moral integrity of the athletes and the reenforcement of ethical values in sport.

Thus, the judge's background is a factor of sportive quality. The performance of international level judges overcomes the performance of less experienced judges, once they use of other cognitive strategies, increasing their global efficiency of error spotting [2]. Also, "If we seek sports excellence, it is not possible not to understand the job of judgement and refereeing in sport. It is not enough to train coaches with scholar degrees, masters and PhDs, it's necessary to think with vision, because it is definitely incongruent and even not logic for the one who evaluates the process and final result of the coach work, not to hold an integral formation that amounts to the level in which the process is occurring" [3]. Academical, professional and social formation of the referee is a characteristic acknowledged by them as being a guarantee of their performances quality [4].

On the other side, the CoP which rules and orientates all the judges' actions, works as an evaluation tool that depending on its structure, its content, and its reliability. All this factors may have a better or worse impact on the judges' performance as evaluators. It is concluded that the most valued skills are those related to the sport's technical parameters and the ability to adapt to any level of competition with self-confidence and self-assuredness [5]. Considering that the judge and the

Corresponding author: Catarina Paula Leandro Sousa Silva, Ph.D., professor, research fields: evaluation in sport.

CoP hold an inseparable dialectic, it is necessary to analyze them together. So the aims of this study are: to define the population of the "International RG Judges" according to personal information, education, professional experience and experience as judges; and to identify their opinion about the CoP 2012 related to its content, structure, clarity and validity of the rules to be applied, as well as possible changes to be introduced to contribute to and improvement of the judging and therefore the correct evolution of the sport.

2. Methods

162 international RG judges answered a specific questionnaire specially developed for this study. It was composed by 15 questions grouped in 2 categories: (1) personal information, education, professional experience and experience as a judge, and (2) objectivity of evaluation in RG and proposals to change the Code of Points [1].

This study was approved by the International Gymnastics Federation (FIG). All the international RG judges, from all over the world, 287 judges were invited by FIG to answer the questionnaire available at Google Drive.

To protect the judges' anonymity, the answers were received anonymously on google drive, so the full blinding of the judges involved was undertaken. The data was collected between July and September 2014.

2.1 Statistical Analysis

For the statistical analysis, we used the Statistical Package for the Social Sciences-Version 21.0 (SPSS 21.0, Chicago, USA) and Microsoft Office Excel 2010.

For the data analysis, non-parametric tests were used (Friedman test and Sign test between groups) to determine whether there were significant differences between groups. Significance level was set at $\alpha = 0.05$ (corresponding to a confidence level of 95%). The frequencies and percentages of the prognostic variables were calculated through the descriptive statistics.

3. Results

The characterization of the Judges is resumed in Table 1.

Table 1 Descriptive statistic of the judges characterization.

Characterization of Judges (N = 162)					
				Freq.	%
Personal information	Sex	Female		160	98.8
		Male		2	1.2
	Age	Mean	43.4		
		Minimum	22		
		Maximum	68		
	Country			59	
Education	High School			16	9.9
	University			80	49.3
	Master			51	31.4
	PHD			15	9.4
Work Experience (RG Coach)	Yes			141	87
	No			21	13
RG International Judge	Brevet I			6	3.7
	Brevet II			29	17.9
	Brevet III			66	40.7
	Brevet IV			61	37.7
RG International Judge Experience	Less than 1 Olympic cycles			42	25.9
	1-2 Olympic cycles			51	31.5
	More than 2 Olympic cycles			69	42.6

The judges were 43.4 years old, 49.3% have a university degree, 87% are also coaches, 40.7% are judges brevet III and 42.6% have been judges for more than 2 Olympic cycles.

We can see in Table 2 the summary of the collected data about the judges' opinion about the objectivity of Difficulty evaluation. The *Mastery* is considered to be the item with less objectivity in evaluation with the answer "less objective to evaluate" collecting 58.6% of the answers. Following we have, in increasing order for the objectivity in judgement, the following groups: *Dance Steps*, *DER (base)* and *Rotation*s in which the answer "more or less objective to evaluate" is the more frequent one 55.6%, 45.7% and 51.2%, respectively. In the groups *DER (criteria)* and *Jumps* the answer "objective to evaluate" is the more frequent with 40.7% and 46.3% respectively. Yet, it's in *Rotations (Basis)* and *Balances* that we see higher values of objectivity in the evaluation with the answer "objective to evaluate" getting 59.9% and 72.2% of the answers, respectively.

Fig. 1 presented the average indicator of objectivity in the evaluation of the different difficulty groups according to the judges' opinion (groups were written in increasing order within the indicator). Globally, we can state that there's a statistically relevant difference (Friedman test, $P = 0.000$) in the objectivity for the different difficulty groups.

Comparing groups, we can see in Table 3 that the *Mastery* show a significant difference ($P = 0.000$ from all of the others groups, being the Mastery the group where the evaluation is seen by the judges as less objective. Also, between *Rotations (base)* and *Jumps*, there are statistically significant differences ($P = 0.002$), being the evaluation less objective in *Jumps* rather than in *Rotations (base)*. The same happens between the groups of *Balance* and *Rotations (base)* ($P = 0.010$), where the evaluation is less objective in *Rotations (base)* than in *Balance*.

When comparing in pairs the groups *DER (base)*, *Rotation (Rot Add)*, *DER (Criteria)* and *Jumps* we can see that there are no statistically significant differences on the degree of objectivity in evaluation. We can even say that the objectivity in evaluation as seen by the judges holds a similar distribution in the four groups ($P = 0.117$).

We can see in Table 4, the summary of the collected data from the judges' opinion about the objectivity in evaluating Execution in Technical and Artistic Faults.

The Technical Faults item is considered the one with higher objectivity in evaluation with the answer "objective to evaluate" getting 80.2% of the answers and the Artistic Faults item is considered the one with less objectivity in evaluation with the answer "Less objective to evaluate" holding 64.8% of the answers. The difference in objectivity in the evaluation of Artistic Faults and Technical Faults is statistically significant (Friedman test, $P = 0.000$).

Regarding the Artistic Faults sub-items (Table 5), the Unity of Composition item is considered the less objective one in evaluation with the answer "less objective to evaluate" collecting 47.5% of the answers. Next, in ascending order of objectivity, the items Music/Movement, and Body Expression were considered "more or less objective to evaluate" (45.7% and 56.6%, respectively). The higher values of objectivity in evaluation are seen in the item Use of Space with the answer "Objective to evaluate" getting 54.3%. Globally, there are statistically significant differences (Friedman, $P = 0.00$) in the objectivity of the evaluation of the different items within the Artistic Faults group, being the degree of objectivity higher in some items than others.

The Artistic Faults item is considered the one with less objectivity in evaluation with the answer "Less objective to evaluate" holding 64.8% of the answers. Regarding the Artistic Faults sub-items, the Unity of Composition item is considered the less objective one in evaluation with the answer "less objective to evaluate" collecting 47.5% of the answers. Next, in ascending order of objectivity, the items Music/Movement, and Body Expression were considered

Table 2 Descriptive statistics of the judges' opinion about the objectivity in difficulty evaluation.

Objectivity of Evaluation	Mastery	Dance Step	DER (Base)	DER (Criteria)	Jumps	Balance	Rotation (Base)	Rotation (Added)
	Frequency Tables (%)							
Less objective	58.6	22.8	21.0	21.6	11.1	3.1	7.4	16.0
More or less objective	34.0	55.6	45.7	37.7	42.6	24.7	32.7	51.2
Objective	7.4	21.6	33.3	40.7	46.3	72.2	59.9	32.7

Friedman test $P = 0.000^{*}$.

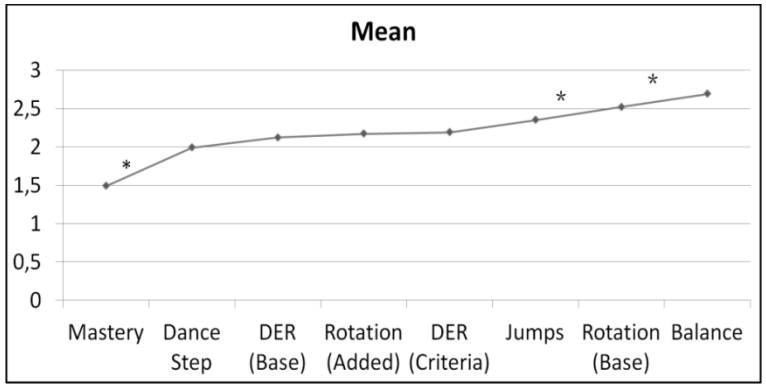

Fig. 1 Average indicator of objectivity in evaluation in the different difficulty groups.
Friedman test $P = 0.000^{*}$.

Table 3 Sign test between difficulty groups by the objectivity in the evaluation.

Sign Test							
	Dance Step-Mastery	DER (base)-Dance Step	Rot (Added)-DER (Base)	DER (crit.)-Rot. (Added)	Jump-DER (crit.)	Rot. (Base)-Jump	Balance-Rot. (Base)
Z	-6.904	-1.724	-0.625	-0.256	-0.891	-3.086	-2.585
Asymp.sig	0.000^{*}	0.085	0.532	0.798	0.373	0.002^{*}	0.010^{*}

"more or less objective to evaluate" (45.7% and 56.6%, respectively). The higher values of objectivity in evaluation are seen in the item Use of Space with the answer "Objective to evaluate" getting 54.3%. Globally, there are statistically significant differences (Friedman, P = 0.00) in the objectivity of the evaluation of the different items within the Artistic Faults group, being the degree of objectivity higher in some items than others.

According to the judges' opinion, we can see in Fig. 2 the average indicator of objectivity in the evaluation of Artistic and Technical Faults (groups are in ascending order within the indicator).

When we try to analyze if there are significant differences between the items distributions, we see in Table 5 that the item Unity Composition differs from the item Music Movement (P = 0.002), with also a significant difference from all the others items, once it's in the item Unity Composition that the evaluation is seen by the judges as less objective. Also between the items Body Expression and Music Movement, there are statistically significant differences (P = 0.000), being the evaluation less objective in Music Movement than in Body Expression. The same happens between the items Use of Space and Body Expression (P = 0.017), where the evaluation is less objective for Body Expression than for Use of Space.

Concerning the Difficulty, in the opinion of the judges, the evaluation criteria for Mastery should be changed, holding 69.8% of the answers (Fig. 3). The

Table 4 Descriptive statistics of the Execution faults by the objectivity in the evaluation.

Execution Faults		
	Technic Faults	Artistic Faults
Objectivity of Evaluation:	*Frequency (%)*	
Less objective	1.9	64.8
More or less objective	17.9	29.6
Objective	80.2	5.6

Friedman Test P = 0.000.

Tabela 5 Descriptive statistics of the Artistic Faults by the objectivity in the evaluation.

Artistic Faults				
	Unity Composition	Music/Movement	Body Expression	Use Space
Objectivity of Evaluation:	*Frequency (%)*			
Less objective	47.5	37.0	16.7	23.5
More or less objective	42.0	45.7	56.8	22.2
Objective	10.5	17.3	26.5	54.3

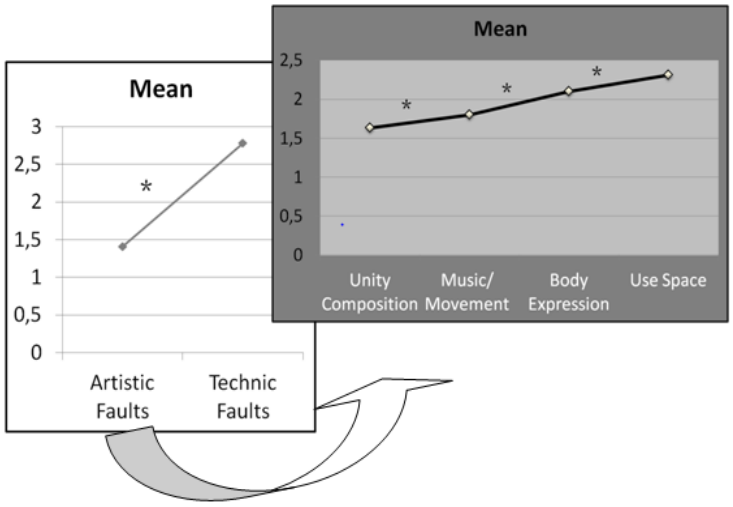

Fig. 2 Average indicator of objectivity in the evaluation of Artistic and Technical Faults.
Friedman test $P = 0.000^{*}$.

Table 5 Sign test between Artistic Components by the objectivity in the evaluation.

	Sign Test		
	Music Movement-Unity Composition	Body Expression-Music Movement	Use Space-Body Expression
Z	-3.125	-4.533	-2.388
Asymp.sig	0.002*	0.000*	0.017*

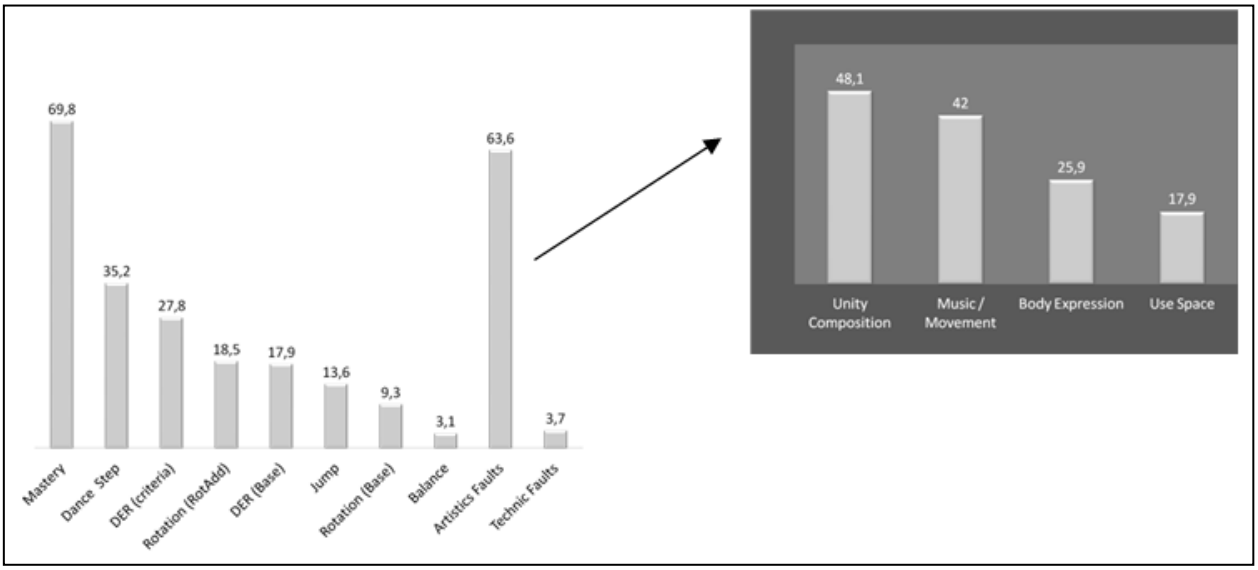

Fig. 3 Frequency table of the difficulty and execution groups by the judges opinion be changed.

Table 6 Frequency of the judges' opinion ("agree that there should be a limit of repetitions for the body difficulty in the different apparatus").

Frequency Table		
	Frequency	%
NOT	57	35.2
YES	105	64.8
Total	162	100.0

Balances group is the one that should not suffer any alteration in the evaluation criteria, holding 3.1% of the answers.

For the Execution, the Artistic Faults evaluation criteria should be changed (63.6%), in opposition to the Technical Faults, which got only 3.7% of the answers.

Studying the items of the Artistic Faults, the evaluation criteria for the Unity of Composition and Music/Movement should be changed, holding 48.1% and 42% of the answers, respectively. The item Use of Space is the one that should not suffer any alteration in the evaluation criteria, holding 17.9% of the answers.

We can see in Table 6 that the majority of the judges (64.8%) agree that there should be a limit of repetitions for the body difficulty in the different apparatus, to improve the variety and the composition of RG routines.

4. Discussion

The sportive judgement, as human, individual and collective action, holds intellectual, volitional and ethical components, which should be taken into consideration in an integrated global way, so the evaluation of sport performance is done in a responsible manner [6].

In this study, we can verify that the majority of international judges has a high academic level, works as a RG coach and has a large experience in judgement. This type of background offer efficiency conditions to the judges, since they register quality judging values, when compared to quality values of judgement got by judges with less experience, as well as lower academic level [7].

The criteria that can distinguish a "specialist in the matter" goes from the connection of the specialist with

the problem, the professional experience, the personal qualities or professional ability, to the guaranty of the quality of the answers and the skill of recognizing detailed information [8].

The most significative difficulties in judgement, come from the excessive amounts of information that the judge needs to quickly summarize in order to transform into a score; in a practical level the judges are able to solve their problems thanks to their experience and knowledge about gymnastics [9], which also allows us to think that the results support the idea that RG judgment is done by individuals who, besides their knowledge of the scoring code, have other resources such as experience and insight, which could be a *plus* in the judgement once the experiences and global vision of the sport lead the judges to deduce, from logic, some aspects that maybe hard to identify or differentiate by younger or inexperienced judges.

Besides de characterization of the judges, we tried to know what they think about the CoP, once they are the ones who use it as an evaluation tool in RG routines. When considered the indicator "objectivity in evaluation" in the different difficulty groups under analysis, the results showed that globally there are statistically significant differences in the objectivity of the evaluation, being the degree of objectivity superior in some groups compared to others. The group with less objectivity in evaluation is Mastery. Following in increasing order of objectivity are Dance Steps, DER (Basis), Rotations (Additional rotations) and DER (criteria).

These results may suggest that the judges find more difficult to evaluate with precision some elements performed by the gymnasts, since their opinion about the evaluation of these groups is less objective probably due to the way the evaluation criteria is described in the CoP, allowing different interpretations.

The evaluation of human performance for some sports is not possible to be done through mechanical ways [10], thus making the reference of the pattern criteria the way to assure validity and reliability in the result of the evaluation, when trying to evaluate the quality of a movement.

It is also important to state that the complexity in the evaluation of the referred difficulty groups (Mastery, Dance Steps and DER) may be also related to the fact that there is no pattern reference in the CoP, compared to what happens for other difficulty groups such as Jumps, Balances and Rotations.

The difficulty groups in which the evaluation is more objective are Jumps, Rotations (Basis) and Balances. These results indicate that in the opinion of the judges, the evaluation criteria of these groups described in CoP allow an objective evaluation, with easy application. It is important to state that the CoP holds a list with pattern images for these three groups, which in our opinion allow an immediate perception of the correctly performed difficulty and consequently the objectivity in evaluation. The criteria to determine the quality of the evaluation should refer to a pattern, model or arbitrary level of that same quality [11].

The study about the degree of agreement between the 4 judges in the evaluation of the different difficulty groups declared in the competition cards used in KIEV 2013 WC, supports the results obtained here [12], once this study highlights the same difficulty groups, where we see more disagreement in the evaluation done by the judges.

When considering the indicator "objectivity in evaluation" within the Execution items, the results showed that globally that there are significant differences in the evaluation objectivity of Technical Faults and Artistic Faults. The Artistic Faults group is considered the less objective in the evaluation. About this group, we found out that there are statistically significant differences in the objectivity of evaluation of the different items that integrate it, with higher objectivity for some items rather than others. The parameter Unity of Composition is considered the less objective in the evaluation. Following in increasing order of objectivity there are Music/Movement and

Body Expression. It's the Use of Space item where we get higher values of objectivity. In the same way, Bučar et al. [13] found results of low validity and reliability in the judgment of artistic components in Artistic Gymnastics, what allows us to consider that the current instruments of evaluation for gymnastics artistic components require monitoring for a possible reassesment and eventual restructuring.

We also tried to identify which evaluation criteria would the judges like to modify in order to potentiate their performances as evaluators. The results found show proposals of changes in the evaluation criteria for Mastery group and for the Artistic Faults group, in particular for "Unity Composition" and "Music/Movement" parameters. Finally, the results indicate that the majority of the judges (64.8%), consider that the CoP should limit the repetitions of body difficulties in the different apparatus, helping to enrich the compositions of RG routines and consequently the evolution of the sport. Liu et al. [14] analyzed the evolution of scoring codes in RG and found out that the evolutive tendency should contemplate variety and diversity allowing exploration of new skills.

5. Conclusions

The evaluation system to determine the final scores of a RG exercise is given by the CoP, being this an instrument elaborated by the FIG. Yet, the judges are the ones using it as an evaluation tool thus their opinions represent a reference value to be considered in its elaboration. They manifested different opinions about the objectivity in the evaluation of the Difficulty, Execution, as well as the different parameters of evaluation in artistic faults.

They highlighted the Mastery, Dance Steps and DER, in Difficulty and the Artistic Faults (mainly "Unity Composition" and "Music/Movement") in execution, as being the ones with most complexity in evaluation when considered the objectivity. They suggested changes in the evaluation criteria of these groups, in order to become more precise in the final evaluation.

Finally the judges stated that they would like to have in the CoP some rules to limit the repetition of difficulties in the different apparatus routines, in order to improve the diversity and variety in RG routines, promoting an enrichment of the sport.

Therefore, we conclude that the instrument of evaluation used right now is not yet ideal to absolutely assure the validity and reliability in RG judgment. These results may contribute for a reconstruction of the CoP and consequently help in the evolution of the sport.

We expect that new rules of artistry evaluation will bring improvement of reliability and consistency of judges and this should be verified through further research of future competitions.

Acknowledgments

We are especially grateful to International Gymnastics Federation (FIG). We also thank to international Rhythmic Gymnastics Judge.

References

[1] FIG-International Gymnastics Federation. 2012. "Code of Points for Rhythmic Gymnastics Competitions, 2012-2016." Accessed November 18, 2012. http://www.fig-gymnastics.com/site/page/view?id=472

[2] Flessas, K., and Mylonas, G. 2015. "Judging the Judges' Performance in Rhythmic Gymnastics." *Medicine & Science in Sports & Exercise. American College of Sports Medicine* 47 (3): 640-8.

[3] Guardo, M. 2004. "Toward a Theory of Arbitration for Sport." *Rev. Digital. Buenos Aires 68*. Accessed July 8, 2004. http:/www.efdeportes.com/ (in Spanish)

[4] Martin, S. 2006. "Desirable Characteristics for Arbitration in Sports Trial in Judo". *Latin American Journal on Exercise and Sport Psychology* 1: 27-40. (in Spanish)

[5] Fernandez-Villarino, M. 2013. "Practical Skills of Rhythmic Gymnastics Judge." *Journal of Human Kinetics* 39: 243-249.

[6] Palomero, M. L. 1996. "Towards an Objectification of the International Code of Points Rhythmic Gymnastics." Ph.D. thesis, The Barcelona University. (in Spanish).

[7] Leandro, C., Ávila-Carvalho, L., and Lebre, E. 2010. "The Evaluation of the Performance of Rhythmic Gymnastics'

Judges." *Palestrica of the Third Millennium Civilization & Sport* 11 (3): 202-6.

[8] Almenara, J. 2013. "The Use of Expert Judgment for the Evaluation of TIC: The Coefficient of Expert Competence." Ph.D. thesis, Sevilha University. (in Spanish).

[9] Plessner, H. 2005. "Positive and Negative Effects of Prior knowledge on Referee Decisions in Sports." In *The Routines of Decision Making*, edited by Betsch, T., and Haberstroh, S. 311-24.

[10] Morrow, J., Jackson, A., G-Disch, J., and Mood, D. 1995. *Measurement and Evaluation in Human Performance*. Library of Congress Mystat, USA.

[11] Simões, G. 2000. *The Evaluation of Teacher Performance*. Texto Editora, Lisboa, Portugal. (in Portuguese).

[12] Leandro, C., Ávila-Carvalho, L., Sierra-Palmeiro, E., and Bobo-Arce, M. 2015. "Accuracy in Judgment the Difficulty Score in Elite Rhythmic Gymnastics Individual Routines." *Science of Gymnastics Journal* 7 (3): 81-93.

[13] Bučar Pajek, M., Kovač, M., Pajek, J., and Leskošek, B. 2014. "The Judging of Artistry Components in Female Gymnastics: A Cause for Concern?" *Journal of Human Kinetics* 37: 173-81.

[14] Liu, X. X., and Kuang, L. 2001. "Review of Evolvement Course of International Evaluation Rules in Rhythmic Gymnastics and Its Effects on Technique Development." *Journal of Beijing Sport University* 3 (24): 412-5.