ACCURACY IN JUDGMENT THE DIFFICULTY SCORE IN ELITE RHYTHMIC GYMNASTICS INDIVIDUAL ROUTINES

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

[1] Faculty of Psychology, Education and Sport, University Lusófona of Porto, Porto, Portugal
[2] Faculty of Sport Science and Physical Education, University of Coruña, Coruña, Spain

Original article

Abstract

The main goal of this study was to analyse the accuracy in judging the Difficulty score in the Rhythmic Gymnastics Kiev World Championship 2013. The accuracy was determined analysing the judges' agreement on the evaluation of the routines difficulty elements. 1152 difficulty forms concerning 288 individual routines were analysed - 4 forms per routine, 1 per judge. To allow the comparison between gymnasts with different levels the individual routines were clustered into 3 subgroups according to their final ranking competition. Body difficulty elements were organized, according to the composition requirements stated in the RG Code of Points (FIG, 2012). Non-parametric tests - Cochran's Q and Chi-Square Tests were applied to determine whether there were significant differences between groups. As main results we can point out that in general the judges did not agree on difficulty evaluation in 40% of the elements. The level of accuracy was lower in the second part of the ranking, and in the Mastery and DER difficulty elements.

Keywods: *Evaluation, accuracy, judge, rhythmic gymnastics.*

INTRODUCTION

Rhythmic Gymnastics (RG) is characterized by the high level of difficulty of the body elements and apparatus handling, combining esthetical and artistic components. This complexity increases the difficulty of the judgment and its accuracy mainly in high level performances. The requirements are quantitative (amount and variety of body and apparatus movements) and qualitative (degree of difficulty and quality in performance) and they are evaluated by the judges according to the rules and evaluation criteria stated in RG Code of points (Bobo, 2002).

The Body and apparatus movements are grouped according to the type of skills, the level of difficulty and the complexity of the movements (Lebre, 2011). The main groups considered in the routines evaluation are: Jumps, Balances and Rotations, Mixed difficulties, additional criteria for the body movements - waves and pre-acrobatics,

Dance Steps, Mastery (special apparatus handling) and Dynamic Elements with Rotation and throw (DER).

In competition the performance is evaluated by 2 panels of judges: the difficulty (D) jury that judges the routines content and the execution (E) jury to evaluate the quality of the routines. The gymnasts present in each competition a difficulty form with all difficulties listed. Each judge must confirm the difficulty elements performed by the gymnast and cross out those that are not correctly performed or not performed at all (FIG, 2012). The final D score is the average of two intermediate scores. When the score become published on the screens, the judges can compare the final score to their own scores. Therefore, the judges score independently although there's still some feedback (Bucar, Cuk, Pajek, Kovac, & Leskosek, 2013).

In previous studies was noted that judging is not only a matter of identifying the sports performance. There are also various facts, identified in the literature, having an influence on the several stages of processing information in gymnastics judgment (Leandro, 2009).

Findlay and Ste-Marie (2004) found out that the were the judges tend to judge better the gymnast higher qualified in previous competitions, concluding that the reputation of the gymnasts have influence on the judging. The judge's experience has been also described as influencing the quality of judgment. Leandro, Ávila-Carvalho, and Lebre (2010) and Ste-Marie, Valiquette, and Taylor (2001) found that the more experienced judges had better perception and anticipation of the elements and there for, were better evaluators. Other factors, as the memorizing capacity (Ste-Marie, Valiquette, and Taylor, 2001), and the tendency to adapt their scores to those given by the judges of the same panel (Boen, Karen, Yves, Jos, and Tim, 2008) were also described. The observation angle (Plessner and Schallies, 2005) and the judges with experience as gymnasts (Heinen, Vinken & Velentzas, 2012) were also described as factors that can influence in the judges accuracy.

Besides these factors, is also relevant to know whether the factors related to the sport specificity as the structure/organization of the Code of Points, the evaluation criteria defined by the sports authorities has an influence (positive or/and negative) on the judge's performance and consequently on the gymnasts final scores.

Rhythmic gymnastics has been experiencing a constant and outstanding evolution in its' technic for the last few years because of the evolution of the Code of Points (Palomero, 1996). The evaluation of the gymnasts is made by a collective observation of judges that should be objective. However, this evaluation is not yet exact, probably due to huge amount of evaluation criteria defined for each difficulty element. This can be verified by the differences registered between the judges of the same panel when the evaluate the same routine. This fact is wellknown in the sport but not yet studied. The majority of studies available deal with the analysis of the technical content of exercises or with the final scores given at the end of each exercise. We could not find any study dealing with the analysis of the difficulty evaluation, element per element, trying to see if the final score of each judge are the product of the validation of the same difficulty elements.

Under this subject, the most relevant studies we found are Palomero (1996) and Bobo (2002), in which both the authors present a new proposal for the scoring, based in performance indicators. Čuk, Fink, & Leskošek (2012) studied the way the different type of final score calculation can change the gymnasts final ranking. Gambarelli, Laquinta & Piazza (2012) developed a formula to avoid pre-agreements between judges. They proposed that the score from the judge of the same country of the gymnast should not enter in the calculation for the gymnast final score. Furthermore, they consider that this would be a factor of guarantee of higher reliability of the final score.

Some of the studies demonstrate that the structure of the Code of Points itself holds decisive influence in scoring gymnasts. In this way is very important to suggest alternative evaluation tools that respect the principles of evaluation (objectivity, validity, reliability, discriminating power and practical utility) and allow a balanced appreciation of the different dimensions of the sport, in either aspects of quality or quantity in the performance of gymnasts (Bobo, 2002).

On the other side, the permanent changes in the Code of Points may cause a lack of understanding of the rules, which lead to a need of evaluation of judging instrument itself (Kirkpatrick & Hawk, 2006). Mark & Shotland (1987) remarked, any evaluation model has to be based on a group of principles, axioms and postulates that must be feasible. To have a Code of Points with an extremely complex model of evaluation that does not work when it has to be used, must be avoid.

According to Bartolomeis (1999) it is not possible to see everything at the same time. The essential point is that the evaluation instrument evaluates what it is supposed to evaluate. For Tamir (1998) the evaluation criteria used should be tested in both validity (precision) and reliability (internal consistency).

We could not find any study based on the analysis of the judges' activity based on the using of the difficulty forms during the competition, making this study a pioneer in this field.

Thus, before suggesting future changes, it is important to understand how it works in the present, finding out what should be changed and what should be kept. According the pyramidal structure of the evaluation process (Figure 1) we established the goal of the study.

The goal of this study was to analyse the accuracy in judging Difficulty in the Kiev World Championship 2013, trying to learn if the 4 difficulty judges evaluate in the same way the difficulty elements on the D form (agreement between the 4 judges). This accuracy was studied for each element declared in the difficulty form trying to understand if the perception of the validation criteria for each elements is similar for all judges. The final difficulty score given by each judge to the same gymnast were very similar, but, with this study, we will analyse if the judges arrived to the final score validating the same elements or validating different elements.

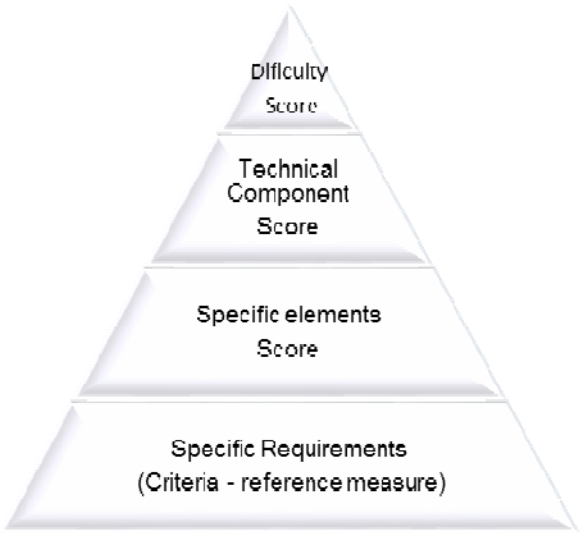

Figure 1. Pyramidal structure for analysis of the evaluation process.

After analysing the date in a global way, we will study the level of agreement between the judges concerning the validation of the difficulty elements according to: (1) the position of the gymnast on the final ranking (1st part, 2nd part and 3rd part), (2) the routine apparatus (hoop, ball, clubs and ribbon), and (3) the type of difficulty element.

METHODS

Subjects and design

1152 difficulty forms concerning 288 individual routines were analysed (4 forms per routine, 1 per judge). The routines were performed by gymnasts from 45 different countries competing at Rhythmic Gymnastics World Championship in Kiev, Ukraine in 2013.

This study was done with the permission of the International Gymnastics Federation. Full blinding of the judges involved was undertaken.

All difficulty elements reported in the difficulty forms provided by the gymnasts at the competition were analysed. Each element was considered validate or not according the notes done by the judge on the form. For each element, we studied the cases of agreement when all 4 judges validate or not the difficulty element and the disagreement when at least one of the judge did not validate and the others consider the element correctly done.

The analyse was done considering the all sample, and the sample clustered into 3 subgroups according to gymnasts final ranking as follows: the first part of the ranking - the top 24 gymnasts, the second part of the ranking - 24 middle gymnasts and third part of the ranking – the 24 lower placed gymnasts on the ranking, to allow the comparison the agreement level of the judges when they evaluate gymnasts with different levels. Then, we studied the sample according to the apparatus used to perform the routine (hoop, ball, clubs and ribbon), and the type of difficulty element performed listed according to the composition requirements of the Code of Points (FIG, 2012), (Figure 2).

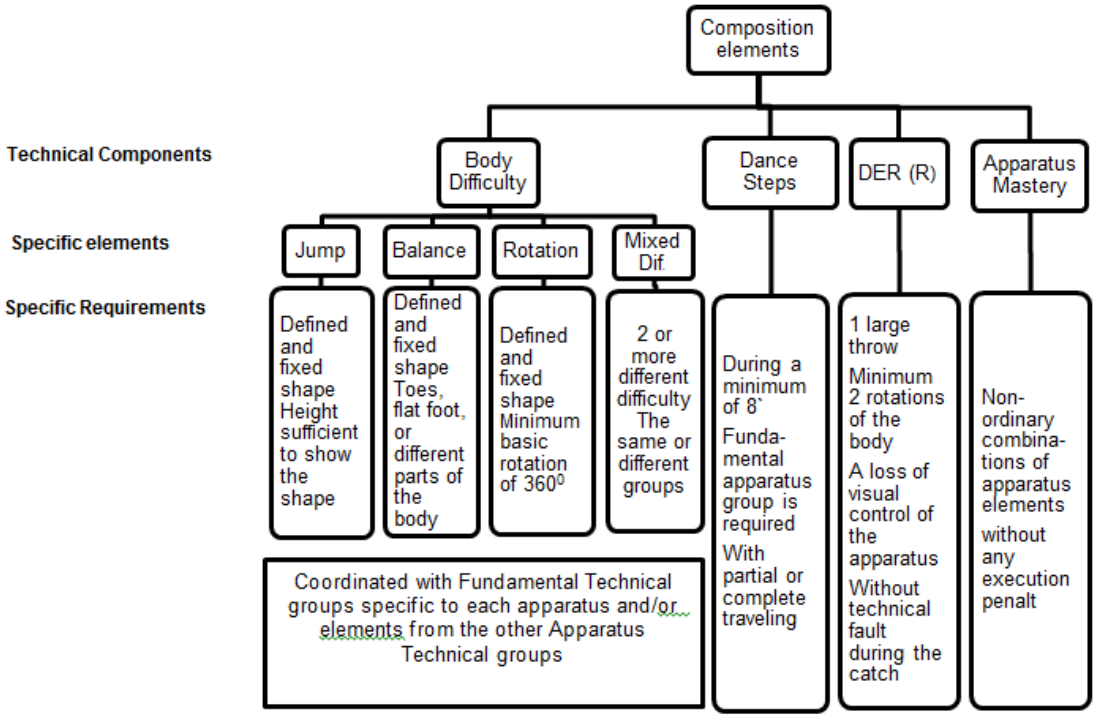

Figure 2. Technical Content of Rhythmic Gymnastics of Individual Gymnasts Routines (COP 2012/2016)

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.

Non-parametric tests (Cochran's Q and Chi-Square Tests) were applied to determine if there were significant differences between groups. We use the Chi-square Tests for two independent samples to study the differences between two groups for each variable and the Cochran`s Q test to analyse when a set of K differs significantly. Significance level was set at $\alpha = 0.05$ (corresponding to a confidence level of 95%).

RESULTS

The forms were analysed first in a global way. For each difficulty element presented on the forms, the percentage of agreement between the 4 difficulty judges concerning the evaluation of the elements was determined. Then, the level of agreement on the elements evaluation was also calculated with the sample dived in 3 groups according to the final ranking of the gymnasts (Table 1).

The judges agreed on the evaluation of 60.0% of the difficulty elements presented on the difficulty forms. When we observe the results according to ranking of the gymnasts, is visible that higher the gymnast is placed in the ranking, higher is the agreement of the judges on the difficulty elements evaluation: 68.8% on the first part of the ranking, 56.1% on the 2nd part and 54.6% on the 3rd part. According to the results of the Chi-Square test, the differences between the cases of agreement and disagreement on the evaluation of the difficulty elements were statistically significant in all cases.

Table 1.
Level of agreement on the evaluation of the difficulty elements presented on the D Forms for all sample, and for the 3 groups according to the final ranking of the gymnasts.

	All Sample		1st part of the Ranking		2nd part of the Ranking		3rd part of the ranking	
	n	%	n	%	n	%	n	%
Not Agree	4871	40.0	1300	31.2	1836	43.9	1735	45.4
Agree	7294	60.0	2865	68.8	2343	56.1	2086	54.6

Chi-Square Test (Asymp.Sig.(2sided)) .000 * (*P<0.05)

Table 2.
Level of agreement on the evaluation of the difficulty elements presented on the D Forms according to the routine apparatus.

	Hoop		Ball		Ribbon		Clubs	
	n	%	n	%	n	%	n	%
Not Agree	1370	41.2	1129	37.3	1191	41.0	1244	40.6
Agree	1867	58.8	1894	62.7	1715	59.0	1818	59.4

Chi-Square Test (Asymp.Sig. (2sided)) .000 * (*P<0.05)

Table 3.
Results of the Cochran`s Q test comparing the results the agreement level for Hoop, Ball, Clubs and Ribbon routines.

	Hoop	Ball	Clubs	Ribbon
N	3174	3023	3062	2906
Cochran's Q	9.960	6.512	25.174	6.232
Sig.	**.018***	.090	**.000***	.099

(*P<0.05)

Table 4.
Results of the Cochran`s Q (C Q) test comparing the results the agreement level for Hoop, Ball, Clubs and Ribbon routines according to the final ranking of the gymnasts.

	Hoop			Ball			Clubs			Ribbon		
	1st part	2nd part	3rd part	1st part	2nd part	3rd part	1st part	2nd part	3rd part	1st part	2nd part	3rd part
N	1069	1078	1027	1044	1036	943	1050	1061	951	1002	1004	900
C Q	5.167	22.273	2.385	10.793	6.660	6.281	7.482	16.485	4.821	18.351	10.042	5.405
Sig.	.173	**.000***	.499	**.013***	.083	.095	.061	**.001***	.185	**.000***	**.019***	.145

(*P<0.05)

Table 5
Level of agreement on the evaluation of the difficulty elements presented on the D Forms according to the different type of elements.

	Not Agree		Agree	
	N	%	N	%
Mastery	726	62.5	436	37.5
Dance Steps	220	28.7	546	71.3
DER	1871	40.6	2735	59.4
Jumps	270	35.6	489	64.4
Balance	302	43.1	398	56.9
Rotations	1065	32.0	2263	68.0
Mixed Difficulties	93	38.3	150	61.7
Criteria assoc. to diff.	324	53.9	277	46.1

Chi-Square Test (Asymp.Sig.(2sided)) .000 * (*P<0.05)

Table 6
Results of the Cochran`s Q test comparing the results the agreement level for different groups of elements according to the final ranking of the gymnasts.

	1st part			2nd part			3rd part		
	N	Cochran's Q	Sig.	N	Cochran's Q	Sig.	N	Cochran's Q	Sig.
Jumps	257	1.227	.817	244	4.483	.208	258	2.92	.401
Balances	207	1.224	.785	238	5.89	.121	255	6.084	.106
Mastery	361	116.05	**.000***	394	46.744	**.000***	407	32.992	**.000***
DER	1607	62.548	**.000***	1567	8.492	**.036***	1432	17.251	**.001***
Dance Steps	260	8.12	**.047***	244	14.709	**.002***	262	2.121	.551
Rotations	1168	56.937	**.000***	1185	1.625	.652	975	4.288	.224
Mix. Diff.	108	10.553	**.015***	81	10.881	**.012***	54	8.937	**.030***
Criteria	197	12.425	**.005***	226	5.158	.164	178	3.774	.282

(*P<0.05)

Table 7.

Results of the Cochran`s Q test comparing the results the agreement level for different groups of rotations elements according to the final ranking of the gymnasts.

	1st part			2nd part			3rd part		
	N	Cochran's Q	Sig.	N	Cochran's Q	Sig.	N	Cochran's Q	Sig.
RPIV Base	195	2,769	.586	167	2,780	.448	188	7,554	.051
RPIV Rotations	431	37,748	**.000***	346	1,213	.763	333	0,283	.969
RFF Base	99	2,314	.594	96	4,116	.295	65	7,627	**.050***
RFF Rotations	273	11.634	**.008***	198	3,915	.283	106	7,382	.060
RF	206	13,481	**.004***	378	7,774	**.050***	283	2,928	.419

Studying the difficulty forms according to the routine apparatus (Table 2) we observed that the range between the disagreement values for the elements evaluation in the 4 apparatus is not very wide (from 37.3% in ball to 41.2% in hoop). However, when we observed the results of the Chi-Square test we could verify that for all apparatus there were significant differences between the values of the agreement and the disagreement on the evaluation of the difficulty elements.

Comparing the data between apparatus through the Cochran`s Q test (Table 3) we could find that there is a significant difference between the values registered for Hoop and Clubs (p value 0.018 and 0.000 respectively), what showed that there was differences in judges agreement level on the elements evaluation for the different apparatus.

Continuing the analysis in each apparatus, we studied the lack of agreement between judges regarding the final ranking of the gymnasts.

The results of the Cochran`s Q test (Table 4) revealed that in Hoop, and Clubs the judges disagreed significantly only on evaluation the difficulty elements of the gymnasts ranked in the 2nd part of the final ranking; in Ball they disagree significantly on the gymnasts in the 1st part of the final ranking; and finally for Ribbon they disagree significantly on the 1st and 2nd part of the final ranking.

We studied the level of judges agreement on the difficulty elements considering the different group of elements described in the Code of Points (Table 5).

In the most part of the groups of elements the agreement percentage between the judges was higher than the disagreement percentage. Only for the evaluation of the Mastery group and the criteria associated to the difficulties (waves and acrobatic skills) the percentage of disagreement between the judges was higher than the agreement - 62.5% and 53.9% respectively for the agreement against 37.5% and 46.1% for the disagreement. Despite this remark, the results of the Chi-Square test the differences between the cases of agreement and disagreement on the evaluation of the difficulty elements were statistically significant in all cases.

The level of agreement between the judges evaluating the different groups of elements was, then, studied regarding the final ranking of the gymnasts (Table 6).

Observing the results we can see that for Jumps and Balances was not remarked a significant disagreement between judges on the evaluation of the elements performed by the gymnasts independently of their placement in the final ranking. For the Dance Steps, there was only a significant disagreement between the judges for the gymnasts placed in the first and second parts of the ranking. Regarding the Rotations and the Criteria associated to the difficulties the significant disagreement was registered only for the gymnasts placed on the first part of the ranking. When we observe the Table 6, we can see that there

are statistically significant differences for the Mastery elements, the DER elements and Mixed Difficulties in the 3 parts of the ranking, once the p value are null or extremely low, what shows clearly the disagreement between the judges.

For the analysis of the rotations we divided them in 3 sub-groups (RPIV - *relevé* rotations (pivot), RFF - rotations on the flat foot or on other part of the body and RF - *fouetté* rotations), because of their different characteristics that means different evaluation requirements (COP, 2012). In each sub-group of RPIV and RFF rotations, we analysed separately the basis of the rotation and the number of rotations associated to the basis.

The level of agreement of the judges evaluating the different type of rotations elements was, then, studied regarding the final ranking of the gymnasts (Table 7).

On the Table 7 we can see that for the basis of RPIV and RFF, there is no statistically significant difference between the evaluation done by the judges in the first and second parts of the ranking. We can see, also, that the values for significance drop substantially in the third part of the ranking. When we analyse the rotations associated to the basis of RPIV and RFF, we can see that in the first part of the ranking that the p value shows clearly the disagreement between the judges in evaluating such part of the difficulty.

Concerning the *fouetté* rotations, there is no agreement between the judges in the first and second parts of the ranking.

DISCUSSION

The goal of this study was determine the accuracy of the judges on the evaluation of each difficulty element presented in the difficulty forms.

Studying the forms in a global way we found that the percentage of elements where the 4 judges of panel agreed on the elements evaluation was higher than the disagreement cases. Nevertheless, we could observe that the judges agreed only in 60% of the elements, what is not enough for an evaluation that is supposed to be exact and accurate.

When we divided the gymnasts in 3 groups according to their place in the final ranking we found out that the judges showed a higher percentage of agreement for the gymnasts placed in the first part of the ranking and lower when we went down through the ranking. These results may suggest that it is more difficult for the judges to evaluate with precision the average and low level gymnasts. This evidence might be related to some criteria to validate the elements that, probably are not enough specific, what can cause some pliability in the evaluation. To solve this problem Simões (2000) suggests that all evaluation systems should hold precise criteria to allow judging correctly the performance. When the gymnast performs perfectly or almost perfectly the element, as usually happens with the top ranked gymnasts, is easier to the judges to recognize the difficulty, applying the evaluation criteria clearly, and tend to agree on its the evaluation. According to Bartolomeis (1999), the evaluation criteria are defined based on a successful criteria, which can facilitate the agreement of judges when the gymnast perform the elements with success, which is the case for the top ranked gymnasts. For the average and low level gymnasts is clearly more difficult to determine the "drop off" point to validate the difficulty elements because these gymnasts are doing the elements with some technical faults which leads the judges to struggle in applying the evaluation criteria stated in Code of Points (FIG, 2012).

We can also speculate that there could be an influence from what is expected, once the judges might expect better gymnasts to perform the difficulty elements correctly, as Findlay & Ste-Marie (2004) found, in a study with figure skating performances, that the judges gave higher scores to the better known skaters, comparing to the less known ones.

Other point that should be added to this discussion is the fact that the evaluation criteria for some difficulty elements include,

according to the Code of Points (FIG, 2012) some points concerning the quality of execution that may contribute to a higher variability on the validation of the elements. The interference of these execution quality criteria may create some variability in the work of the difficulty judge, creating some "grey zones" in the evaluation of difficulty elements. According to Askew (2002), the evaluator should direct all his attention for a specific profile and ignore the interference of any other information from a different profile.

The analysis of the results by apparatus revealed that the percentage agreement had not big differences for the routines performed with different apparatus. The results showed that behavior does not change from one apparatus to another; on the contrary we could remark that there was a consistency on the lake of accuracy in the difficulty elements evaluation. This consistency is due to the fact that the difficulty elements used in the different apparatus are basically the same and therefore, the requirements to validate the apparatus are the same (FIG, 2012).

Observing the results obtained for the judgment accuracy when we studied it for each apparatus and according to final ranking of the gymnasts we found out that the lower values of accuracy in the judgment were registered mainly in the gymnasts of the second part of the ranking. Besides what was already discussed about the lack of precision in defining the evaluation requirements, we are still able to speculate about the short amount of time that each judge has to consider a great amount of requirements defined for every single element in the routine composition, which may cause high variability between judges scores (Čuk & Karacsony, 2004). This is a problem for the average gymnasts because in opposite to higher level gymnasts where is easy to identify the difficulty elements correctly done and to lower level gymnast where visible when they do not perform the difficulty elements correctly, the average gymnasts often present an unclear version of the difficulty element making the decision to validate an element even more difficult than usual.

The results obtained when we analysed the level of agreement of the judges according to the type of difficulty element evaluated showed that the judges could not agree on the evaluation of the Mastery elements, and the Criteria (waves and pre-acrobatic elements) associated to the difficulty elements. These two groups showed levels of disagreement higher than the agreements, clearly in opposition to what happened with the other groups. The results suggests that definition of the evaluation requirements may have not an enough clear statement in the Code of Points (FIG, 2012), which can lead the judges in troubles to decide when the elements should be validate or not. According to the technical requirements to validate a Mastery element, it should be "a combination of extraordinary apparatus elements performed without technical faults". The definition of "extraordinary apparatus elements" is too vague to allow the judge to evaluate the elements with accuracy and could be also influenced by the international experience of the judge: after judging a certain number of international competitions the level of expectation for an "extraordinary element" can be raised. Knowing that in the World Championships the judges (one for each country participating) has different background experiences, we can understand that they cannot evaluate this technical requirement with same level of accuracy. In this way we strongly recommend that the Code of Points should include much more precise definitions of the technical requirements, because, according to Simões (2000) the evaluation criteria should be understood in equal manner by the various evaluators, in a way that the effect of the evaluation done may be valid and reliable.

After a more detailed analysis of each group of difficulty elements according to the gymnasts ranking, we could see that for the Jumps and Balances the level of agreement between judges was similar in the 3 parts of the ranking, showing that in these elements

the judges apply the same evaluation criteria. The evaluation criteria are understood and applied in the same way by the evaluators, once they produce the same result. This result allows us to speculate that visual image of the element allows a quicker and more reliable understanding, once the stated difficulties are presented. Boen, Karen, Yves, Jos, and Tim (2008) reach the conclusion that the possibility of feedback creates agreement between gymnastic judges. We know (unpublished study), that jumps and balances are repeated frequently in exercises, by the gymnasts in different apparatus routines, what facilitates the visual experience of the judge and therefore more precision in the application of evaluation criteria. According to Ste-Marie, Valiquette, & Taylor (2001), the visual image that is kept in the memory can influence the judge's performance. The agreement may be higher in the elements that appear often in exercises and because of that the judges have a clearer visual image and therefore a more precise evaluation.

In opposition, we can see that there are statistically significant differences between the 3 parts of the ranking in Mastery elements and DER elements, what clearly reveals the disagreement between the 4 judges on the validation of these difficulties. Besides what was already discussed above about the validation of Mastery elements, it is still relevant underline these elements are not listed and therefore the higher number of possible combination of handling contribute to make the evaluation of these type of elements even more difficult. We understand here that the absence of a list of Mastery elements would bring high improvements in routines creativity, although this could also bring the possibility for mixing originality concepts that should and must be evaluated in the originality item stated in COP (FIG, 2012). According to Balcells, Martín & Anguera (2009) it is possible to evaluate the originality and creativity with validity and reliability defining evaluation criteria that can be seen by the evaluators.

In the case of DER elements, the results lead us to the high number of criteria to bear in mind for the judge during the observation. According the Code of Points (FIG, 2012), the DER has an unlimited value and may contain till 19 different criteria that can be repeated. The judge has to memorize the criteria done to have the possibility to cross out on the difficulty form those what were not performed correctly or not done at all. Ste-Marie and Lee (1991) and Ste-Marie, Valiquette, & Taylor (2001) showed that the objectivity of a judge can be compromised by biases of memory. Also, the high number of criteria performed in such short may be responsible for this lack of agreement between the judges. We can speculate that the small amount of time that the judge has to observe and make all the possible deductions on the Difficulty form could be other source of variability between judges which may cause the evaluation of this group more vulnerable. Bucar, Čuk, Pajek, Kovac, & Leskosek (2013) and Čuk & Karacsony (2004) identified this same problem in the evaluation of the Vault execution in female artistic gymnastics, once this is also done in few seconds with 21 possible deductions.

The data concerning the Dance Steps showed also a significant disagreement of the judges in the validation. Dance Steps has, as criteria to be validate, the duration of at least 8 seconds, which can cause high variability in the evaluation, since this evaluation is done without a stopwatch or other device, but through the sensibility of the judge, and can be serious influenced by the *tempo* of the music.

The evaluation of the Mixed difficulties and Criteria associated to the difficulty elements (acrobatic elements and waves) reveals a significant disagreement between the judges, which could be due to the statement on the Code of Points concerning the link between the wave or acrobatic element and difficulty element itself. According to COP (FIG, 2012) the link must be immediately before or after but it is not clearly specified if it should be in continuity of the difficulty element or if it

could be a composition of two elements. According to Plessner (2005), the non-stated rules which can be considered as social norms, may influence the judge's decisions. It's important that they have great knowledge of the rules, to avoid wrong decisions.

Concerning the rotations, we can see that when evaluated the base of RPIV and RFF, there's no significant difference in the evaluation, in the first and second parts of the ranking. However, we can see that the values of a significant decrease in the third part of the ranking. Normally it is on the third part of the ranking where we find the lower level gymnasts and therefore with poor execution technique straight from the base of the rotation. According to the COP, the judge has to see the form, the degrees (360º) of the first turn and the technical faults that cancel the difficulty. The junction of all this factors (which are more present in the lower level gymnasts) belonging to two different profiles (difficulty and execution), may be explain the results of variability between judges found in the evaluation of this part of the difficulty.

Concerning the number of rotations associated to the base of RPIV and RFF, we can see that in the first part of the ranking there is clearly disagreement between the judges in evaluating these difficulty elements. About *fouetté* rotations, we found that in the first and second parts of the ranking there is no agreement between the four judges.

It is in the first and second parts of the ranking that the rotations performed done have a higher number of turns. By the evaluation criteria stated in COP, the judge has to count the number of full turns performed that is sustained fixed, without technical faults. Then, the difficulty in counting a high number of turns performed (that can go upper than 10 turns, mainly in *fouettés*) at high speed in few seconds, identifying the technical faults that implies the cancellation of the difficulty, may be in the origin of this variability for this kind of elements, in the first part of the ranking. Once again, we highlight here the interference of some criteria concerning execution, when judges are judging difficulty. According to Plessner (2005), positive and negative effects of prior knowledge on referee decisions and observation of a high amount of demand in such a short amount of time, may cause the loss of important information.

CONCLUSIONS

The four judges of difficulty panel did not agree in their evaluation in 40% of the difficulty elements presented in the difficulty forms. Regarding the final ranking of the gymnasts the agreement level is higher in the high and low level gymnasts. The level of accuracy was lower in the second part of the ranking, and in the difficulty elements which validation criteria depends not only from difficulty criteria but also from execution criteria.

The analysis by type of difficulty elements showed that for the Jumps and Balances the judges agreed on the evaluation of the elements which means an acceptable accuracy of judgement, but for the other types of elements the level of disagreement between the judges was significantly high to be an accurate judgement, where we highlight the Mastery and DER difficulty elements. This study provides updated information about the precision of difficulty judging in rhythmic gymnastics, to be considered in the possible alteration of the present code of points, in particular in the definition of the evaluation criteria of the elements where we see the highest disagreement between judges.

REFERENCES

Askew, S. (2002). Feedback for learning. *Journal of eduction for teaching*, *28*, 83-90.

Balcells, M., Martín, C., & Anguera, M. (2009). Instrumentos de observación ad hoc para el análisis de las acciones motrices en Danza Contemporánea, Expresión Corporal y Danza Contact-Improvisatio. Apunts educación física y deportes.

Ciencias aplicadas a la actividad física y el deporte, 14-23.

Bartolomeis, F. (1999). *Avaliação e Orientação: Objectivos, instrumentos e métodos.* Lisboa: Livros Horizonte.

Bobo, M. (2002). *El juicio deportivo en Gimnasia Rítmica. Una propuesta de evaluación basada en indicadores de rendimiento. (PHD Thesis*), Universidad da Coruña, Instituto Nacional de Educación Física de Galicia, Coruña.

Boen, F. B., Karen, H., Yves, V. A., Jos, F., & Tim, S. (2008). Open feedback in gymnastic judging causes conformity bias based on informational influencing. *Journal of Sports Sciences*, *26*(6), 621–628.

Bucar, P. M., Cuk, I., Pajek, J., Kovac, M., & Leskosek, B. (2013). Is the Quality of Judging in Women Artistic Gymnastics Equvalent at Major Competitions of Different Levels? *Journal of Human Kinetics*, *37*, 173-181.

Čuk, I., Fink, H., & Leskošek, B. (2012). Modeling The final score in Artistic Gymnastics by different weights of difficulty and execution. *Science of Gymnastics Journal*, *4*, 73 – 82.

Čuk, I., & Karacsony. (2004). *Vault: methods, ideas, curiosities, history*. Ljubljana: STD. Sangvinck.

FIG. (2012). *Code of Points for Rhythmic Gymnastics Competitions.* Available at: http://www.fig-gymnastics.com/site/page/view?id=472

FIG. (2013). *Gymnastics Results*. Available at: http/www.gymnasticsresults.com

Findlay, C., & Ste-Marie, M. (2004). A Reputation Bias in Figure Skating Judging. *Journal of Sport & Exercise Psychology*, 26 (1), 154-166.

Gambarelli, G., Laquinta, G., & Piazza, M. (2012). Anti-collusion indices and averages for the evaluation of performances and judges. *Journal of Sports Sciences*, *30*(4), 411-417.

Heinen, T., Vinken, P., & Velentzas, K. (2012). Judging Performance In Gymnastics: A Matter Of Motor Or Visual Experience? . *Science of Gymnastics Journal, 4,* 63 – 72.

Kirkpatrick, J., & Hawk, L. (2006). *Curricula and evaluation: Maximizing results.* Measuring and Evaluating. Available from EBSCO .

Leandro, C. (2009). *Avaliação de Juízes de Ginástica Rítmica.* (Master Thesis), Porto University, Porto.

Leandro, C., Ávila-Carvalho, L., & Lebre, E. (2010). The avaluation of the performance of Rhythmic Gymnastics` Judges. Palestrica of the *Third Millennium Civilization & Sport, 11*(3), 202-206.

Lebre, E. (2011). *Technical principles for the new framework.* Crossroads to the Future [Press release].

Mark, M., & Shotland, R. (1987). *Multi Methods in Program Evaluation. New Directions For Program Evaluation* (Vol. 35). Londres: Jossey – Bass Inc.

Palomero, M. L. (1996). *Hacia una objetivación del Código*. Barcelona: Internacional de Gimnasia Ritmica Deportiva.

Plessner, H. (2005). Positive and negative effects of prior knowledge on referee decisions in sports. In T. Betsch, Haberstroh,S (Ed.) *The routines of decision making* (pp. 311–324).

Plessner, H., & Schallies, E. (2005). Judging the Cross on Rings: A Matter of Achieving Shape Constancy. *Applied Cognitive Psychology*, *19,* 1145-1145.

Simões, G. (2000). *A avaliação do desempenho Docente.* Lisboa: Texto Editora.

Ste-Marie, D., & Lee, T. D. (1991). Prior processing effect on gymnastic judging. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 126-136.

Ste-Marie, D. M., Valiquette, S. M., & Taylor, G. (2001). Memory Influenced Biases in Gymnastic Judging Occur Across Different Prior Processing Conditions. *Research Quarterly for Exercise and Sport*, *72*(4), 420-426.

Tamir, P. (1998). Assessment and Evaluation in Science Education . Opportunities to Learn and Outcomes *International Hand book of science*

Education (pp. 761-789): Dordrecht: Kluwer Academic Publishers.

Corresponding author:

Catarina Leandro
University Lusófona of Porto
Faculty of Psychology, Education and Sport
Rua Augusto Rosa, nº 24 (à Pç. da Batalha)
Porto 4000-098
Portugal
E-Mail: catarinaleandro@sapo.pt