THE JUDGING OF ARTISTRY COMPONENTS IN FEMALE GYMNASTICS: A CAUSE FOR CONCERN?

Maja Bučar Pajek[1], Marjeta Kovač[1], Jernej Pajek[2] and Bojan Leskošek[1]

[1] Faculty of Sports, University of Ljubljana, Slovenia
[2]University Medical Center Ljubljana, Slovenia

Original article

Abstract

Due to its nature and relatively poor definitions in the Code of points, judging of artistry may suffer from serious flaws in reliability and validity. We have used the balance beam artistry evaluation forms given by 5 execution judges at World Championship in Tokyo 2011 to analyze reliability and validity. Data on 194 competitors was gathered. Deductions were received by a highly variable number of competitors from separate judges in the same components of artistry. The variability of average total artistry deduction was relatively large, ranging from 0.18 to 0.39 points. The average correlation coefficient in total artistry deductions between all judge pairs was 0.6±0.06 and average correlation coefficient in total deductions from execution score was 0.73±0.04, $p < 0.001$. Kendall's coefficient W revealed significant systematic over- or under-rating of judges in the components of artistry of presentation, sureness of performance and variation in rhythm, but also in total artistry deductions (W values ranged from 0.05 to 0.53, $p < 0.001$ for all W coefficients). We conclude that neither reliability nor validity of artistry judging was satisfactory in this analysis. Further thorough evaluations of judging performance in artistry are needed to guide accommodations and hopefully improvements in this field.

***Keywords**: Artistic Gymnastics, Evaluation, Panel Judging, Bias.*

INTRODUCTION

Judging in artistic gymnastics has long tradition and crucially influences outcome. The differences between gymnasts are often small, especially if the homogenous group such as the world class gymnasts competes at the higher level competitions as World Championships or Olympic Games (GymnasticsResultsCom, 2012).

Several aspects of judging performance were already described in the past (Aronson, 1970; Ansorge et al., 1978; Ansorge & Scheer, 1998; Boen, Van Hoye, Auweele, Feys & Smits 2008; Bučar Pajek et al., 2011; Bučar et al., 2012; Pajek et al., 2013; Dallas & Kirialanis, 2010; Leskošek et al., 2010; Plesner, 1999; Plessner & Schallies, 2005; Popović, 2000; Ste-Marie, Valiquette & Taylor; 2001)

The Code of Points for women 2009 defined 5 judges for evaluating exercise execution at World Championship in Tokyo 2011. This results in the E (execution) score. In addition, 2 judges evaluate exercise content and they provide the D (difficulty) score. E scores range from 10 points down in decrements of 0.1 and D scores go from 0 points rising in increments of 0.1 (FIG, 2009). According to the Code of points the judges giving execution (E) scores may penalize competitors for general mistakes, specific execution mistakes and artistic flaws (FIG, 2009).

In the recent years our group has performed several analyses of the judging performance at various competitions and several propositions for further improvements in this field were made (Bučar, Čuk, Pajek, Karacsony, & Leskošek, 2012; Bučar Pajek, Forbes, Pajek, Leskošek, & Čuk, 2011). It was our impression that evaluation of artistry components suffers from serious flaws in reliability and validity of judging. We also questioned the relevance and justification for deductions in some components of artistry, such as gesture and mimic, which may be highly variable between the judges and subject to personal and subjective opinions. Since the sum of all artistry deductions may rise up to 0.8 points, this may significantly impact the final result and we feel that such an impact should be justified by quantitative data.

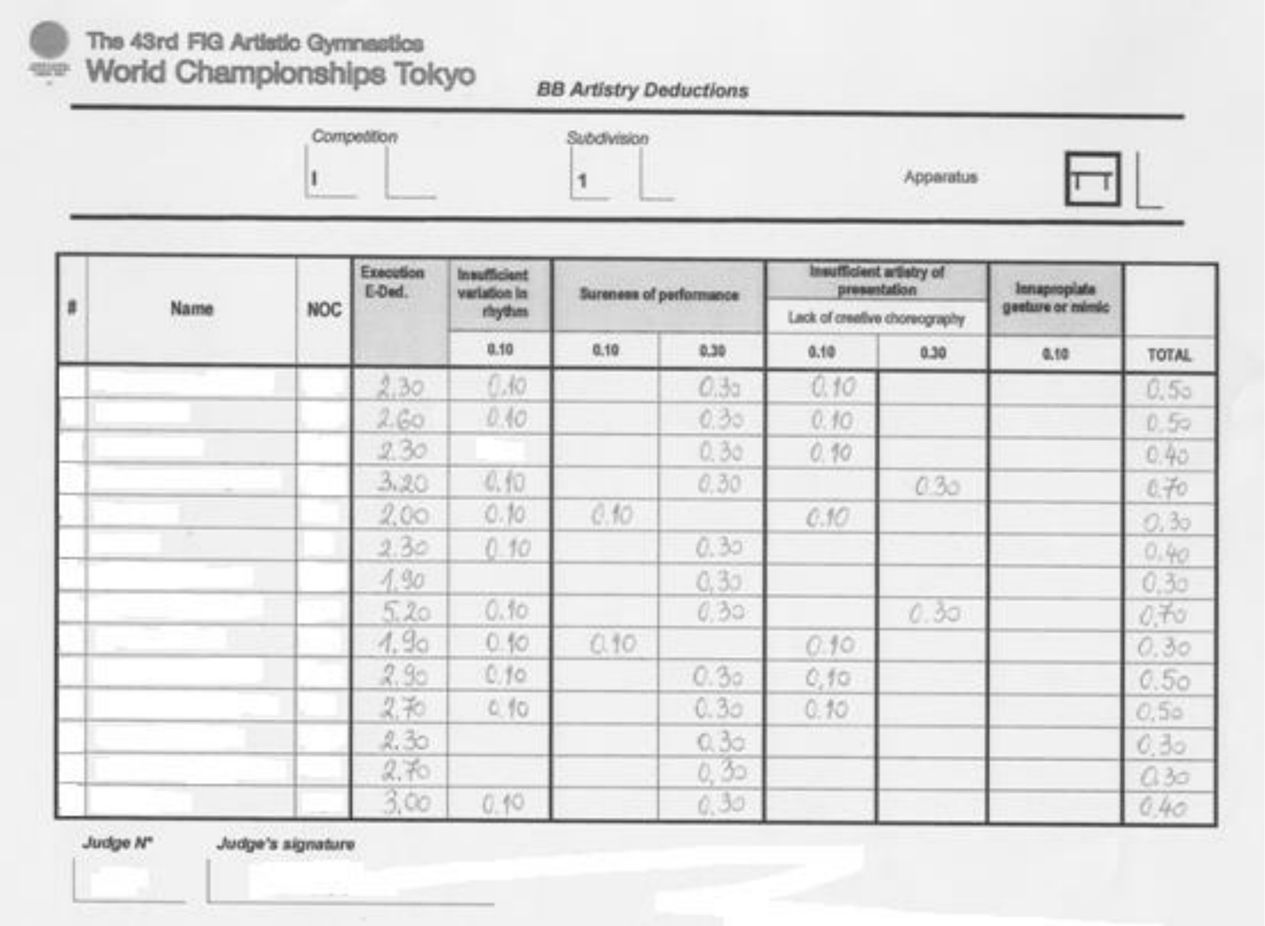
In female artistic gymnastics artistry is evaluated and judged at two apparatus: balance beam and floor. Artistic deductions are derived from the following components of artistry: inappropriate gesture and mimic, insufficient artistry of presentation, sureness of performance and insufficient variation in rhythm (Table 1). The deductions are given in the magnitude of 0.1 or 0.3 points and the final artistry deduction is included in the final E score.

But it seems, that is not very clear for the judges and coaches what is the artistry and how the judges are expected to judge. In theory, artistry at balance beam and floor is defined as mastery of execution (the judges should move away from the personal taste of beauty and follow the definition in the Code of points). But in the Code of points (FIG, 2009), there was no clear definition of mastery, just deduction for artistry mistakes (Table 1).

In Artistic Gymnastics we are concerned with the problem of a systematic bias and inconsistency of judges which may influence the final ranks of competitors. Continuous monitoring of the quality of judging (incorporating reliability and validity) is a necessity. Therefore we designed this study with the aim to analyze the reliability and validity of judging artistry in female gymnastics. We have used the judging results from one of the world's largest competitions and examined them for indices of inter-rater reliability and validity. On the basis of results we proposed several lines of concern regarding the performance of judging and justified the need for further exact and thorough reevaluation of this field.

Table 1. *Artistry Deductions at Balance Beam (FIG, 2009).*

FAULTS	**0.1**	**0.3**
Insufficient variation in rhythm	X	
Sureness of performance	X	X
Insufficient artistry of presentation throughout the exercise including:		
Lack of creative choreography originality of composition of elements and movements	X	X
Inappropriate gesture or mimic not corresponding to the movements	X	



The 43rd FIG Artistic Gymnastics
World Championships Tokyo

BB Artistry Deductions

Competition 1

Subdivision 1

Apparatus

#	Name	NOC	Execution E-Ded.	Insufficient variation in rhythm	Sureness of performance		Insufficient artistry of presentation Lack of creative choreography		Innapropiate gesture or mimic	
				0.10	0.10	0.30	0.10	0.30	0.10	TOTAL
			2.30	0.10		0.30	0.10			0.50
			2.60	0.10		0.30	0.10			0.50
			2.30			0.30	0.10			0.40
			3.20	0.10		0.30		0.30		0.70
			2.00	0.10	0.10		0.10			0.30
			2.30	0.10		0.30				0.40
			1.90			0.30				0.30
			5.20	0.10		0.30		0.30		0.70
			1.90	0.10	0.10		0.10			0.30
			2.90	0.10		0.30	0,10			0.50
			2.70	0.10		0.30	0.10			0.50
			2.30			0.30				0.30
			2.70			0,30				0.30
			3,00	0.10		0.30				0.40

Judge N° Judge's signature

Figure 1. *Judge sheet for artistry (to protect judges and gymnasts identity we erased identifications data from presented artistry sheet (Majer, 2013).*

METHODS

This evaluation of artistry is based on results at World Championship in Tokyo 2011. The evaluation forms for artistry deductions were inspected for all competitors on balance beam qualifying session (N=194). Each competitor was evaluated by 5 judges of international level. For each competitor the deduction score for each component of artistry and final artistry deduction score given by each judge was noted. Final difficulty, execution and total score were monitored as well for each competitor. The identity of judges was not revealed and was kept anonymous for the purpose of this report.

The reliability of judges in monitoring artistry was evaluated by counting the frequency of missing scores and by distribution of deductions at various components of artistry.

The compliance and coherence of judges was evaluated through calculation of mean artistry deduction and mean rank of the artistry deduction for each individual judge. Ranks of the judge's artistry deduction for each competitor were analyzed using the Kendall's coefficient of concordance W. In this specific application of Kendall's W, the higher (and more significant) W values denote systematic over or under-rating of artistry deductions and are therefore a reflection of a special case of judging bias. Kendall's W was calculated for final artistry deduction and separately for each component of artistry.

Kendall's coefficient of rank correlation tau-b between judges for total artistry deductions was compared to tau-b for final total deductions without artistry deductions.

This evaluation was used to compare the concordance of judges at artistry and other components of judging execution. Finally, the Kendall's tau-b correlation coefficient between total artistry deductions and final D, E and total scores were calculated for separate judges.

Used set of variables included: FREQUENCIES OF DEDUCTIONS for components of artistry evaluated by the judges, TOTAL ARTISTRY DEDUCTIONS with distribution by judges, MEAN RANK OF ARTISTRY DEDUCTIONS given by individual judges and TOTAL ARTISTRY DEDUCTION MEAN RANK, CORRELATION COEFFICIENTS of total artistry deductions and total deductions between judge pairs.

RESULTS

There were 194 competitors on balance beam qualification session with artistry deductions included. The frequencies of missing deductions and distribution of deductions for various artistry components are given in table 2.

For inappropriate gesture or mimic there was no deduction for vast majority of competitors. Judge No. 4 stands out with the highest number of deductions and the highest number of missing values at all components of artistry. In general, there are large differences in the distribution of no deduction, 0.1 and 0.3 deductions for sureness of performance and insufficient artistry of presentation.

When the data on individual judge's artistry evaluation forms were inspected, several cases were found, where the judges gave artistry deductions, but calculated the sum of separate deductions in a wrong way (the final artistry deduction was different than the sum of separate components).

Total artistry deductions with distribution according to individual judges are given in table 3.

Table 2. *Frequencies of Deductions and Missing Values for Components of Artistry Evaluated*

Artistry component	Deduction level	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
Inappropriate gesture or mimic	No deduction	194	175	190	138	179
	Deduction 0.1	0	0	0	9	1
	Missing value	0	19	4	47	14
Insufficient variation in rhythm	No deduction	88	138	48	46	10
	Deduction 0.1	106	37	142	102	171
	Missing value	0	19	4	46	13
Sureness of performance	No deduction	2	21	87	57	24
	Deduction 0.1	34	78	94	60	64
	Deduction 0.3	158	76	9	30	92
	Missing value	0	19	4	47	14
Insufficient artistry of presentation	No deduction	88	106	124	74	112
	Deduction 0.1	89	62	59	67	47
	Deduction 0.3	17	7	7	6	22
	Missing value	0	19	4	47	13

Table 3. *Number of Competitors with Given Total Artistry Deduction and Their Means by Individual Judges.*

Total artistry deduction	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
No deduction	1	17	31	31	1
Deduction 0.1	11	46	53	27	15
Deduction 0.2	13	33	62	26	52
Deduction 0.3	49	57	35	31	22
Deduction 0.4	58	26	3	6	50
Deduction 0.5	46	6	8	18	24
Deduction 0.6	4	2	1	3	1
Deduction 0.7	12	4	1	3	16
Deduction 0.8	0	0	0	3	0
Missing	0	3	0	46	13
Mean total deduction	0.39	0.24	0.18	0.24	0.34

Table 4. *Mean Ranks of Judge's Artistry Deductions and Kendall's Coefficient of Concordance W.*

Artistry component	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	N	Kendall's W[a]	Sig.
Insufficient variation in rhythm	2.85	2.02	3.36	3.08	3.70	133	0.314	<0.001
Sureness of performance	4.11	3.42	1.86	2.13	3.48	132	0.532	<0.001
Insufficient artistry of presentation	3.47	2.88	2.74	2.90	3.00	133	0.054	<0.001
Total artistry deduction	4.3	2.8	1.87	2.12	3.91	143	0.527	<0.001

Table 5. *Correlation Mmatrices for Total Artistry Deductions Between All Judge Pairs. Correlations between total deductions (but without artistry deductions, which were subtracted from total deductions) are also shown.*

	Item	D score	E score	TAD 1	TAD 2	TAD 3	TAD 4	TAD 5
Correlations with final scores	Final score	0.68	0.78	-0.61	-0.66	-0.66	-0.62	-0.71
	D score		0.44	-0.49	-0.52	-0.53	-0.60	-0.51
	E score			-0.61	-0.63	-0.63	-0.49	-0.71
Correlation matrix for artistry deductions	TAD 1				0.55	0.59	0.46	0.60
	TAD 2					0.70	0.62	0.61
	TAD 3						0.61	0.63
	TAD 4							0.58
Correlation matrix for total deductions without artistry				TD 1	TD 2	TD 3	TD 4	TD 5
	TD 1				0.73	0.73	0.69	0.70
	TD 2					0.83	0.73	0.70
	TD 3						0.74	0.73
	TD 4							0.67

The coefficients of variation of the artistry deductions for the individual judges 1-5 were: 0.36, 0,63, 0,73, 0,84 and 0,48. Mean ranks of judges for components of artistry and total artistry deductions mean rank are presented in table 4. Ranks were tested for concordance with Kendall's W coefficient of concordance. These results are also given in table 3. No data is given for inappropriate gesture or mimic component, since there were no deductions for this component for any of the competitor in 3 out of 5 judges.

The correlations in total artistry deductions between separate pairs of judges are given in the table 5. This table also holds correlation matrices for various correlations of artistry deductions with other variables for all judge pairs.

TAD - total artistry deduction, the numbers denotes judges; TD - total deduction without artistry deduction, the number denotes judges.

It can be seen, that all correlation coefficient for judge pairs in total deductions (TD) are higher than coefficients for total artistry deductions (TAD), average TAD correlations coefficient was 0.6±0.06 and average TD correlation coefficient was 0.73±0.04, the difference between TAD an TD being statistically significant, $p < 0.001$. In general, the magnitude of correlations between TAD and final scores, D scores and E scores are expectedly negative, but also of relative low magnitude.

DISCUSSION

In the present analysis we have evaluated the artistry deductions on balance beam qualifying session at the World Championship competition in Tokyo 2011. We have found serious deviations in reliability of monitoring the artistry of competitors and significant values of systematic under- or over-rating denoting suboptimal validity.

For the component of inappropriate gesture and mimic it can be seen, that to a vast majority of competitors no deduction was given from 3 out of 5 judges. Only a single competitor was penalized from judge 5 and 9 competitors (not including the competitor of judge 5) were penalized from judge 4. These findings throw a major doubt on the relevance of this artistry category to be judged, when no deduction in this category is given from majority of judges to any of competitors. Additional source of problems when judging gesture and mimic comes from the fact that the judges may be inspecting the competitors mostly from the flank position and from the substantial distance, which prevents the appropriate gesture and mimic assessment. Additional problem, especially for less experienced judges, is that judges may spend significant amount of time looking at scoring sheet and therefore missing some of the less important features of the routine, such as mimic and gesture (Ste-Marie, 2000).

When looking at inter-judge variability, we have found large differences in the distribution of magnitudes and the mean total artistry deductions. The dispersion of mean deductions was relatively large, going from 0.18 points for judge 3 to 0.34-0.39 points (twice the amount) for judges 1 and 5. This is supplemented by the significantly ($p<0.001$) lower correlations between judge pairs in total artistry deductions as compared to correlations in total deductions from E score (without artistry deductions). Furthermore, the number of competitors without deduction for separate components of artistry is highly variable between the judges and even some calculation mistakes in summation of artistry deductions were noted. Taken together, these facts point to an insufficient inter-rater reliability of artistry judging, the finding which is substandard for general judging performance at major gymnastic competitions (Leskošek, Čuk, Karácsony, Pajek, & Bučar, 2010; Pajek, Cuk, Pajek, Kovac, & Leskosek, 2013).

Serious flaws in validity of artistry judging were also found. Here we focused on a special case of validity, which deals with the presence of systematic over or under-rating or scoring of competitor's artistry (what is also called bias). Table 3 clearly shows that we found a significant amount of systematic under- or over-rating in every artistry component examined. We speculate, that this has a different origin than national bias, where judges give better scores to gymnasts of same nationality (Ansorge & Scheer, 1988). This may better be explained by differences in character and personal characteristics (personal taste, culture), judging education and relatively high frequency of changes in FIG rules regarding the judging of artistry (FIG, 2009). The judging of artistry was also relatively poorly defined in FIG rules. In Code of points 2013 – 2016 artistry is better defined (FIG, 2013). 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.

In conclusion, we have analyzed the judging of artistry on balance beam at World Championship 2011 competitions and found worrying results. The inter-rater reliability was poor with large differences in number of competitors penalized and in average artistry deductions. For the artistry component of inappropriate gesture and mimic, majority of judges gave no deduction and other judges differed significantly. This puts the inclusion of this artistry component in the present code of points (FIG, 2013) under question. Validity of judging was substandard with systematic under- or over-rating found in all examined components of artistry and total artistry deductions as well. Due to the limitation of data to a single competition these results may be regarded as pilot and hypothesis generating. We propose that the performance of judging artistry should be repetitively examined in present Olympic Cycle (2012-2016) and if our results are confirmed a thorough reevaluation of the way and scope of artistry evaluation should be made by FIG.

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 womens gymnastics induced by expectations of within-team order. *Res Q, 49*, 399-405.

Aronson, R.M. (1970). *The art and science of judging men's gymnastics*. Lowell: Lowell Technological Institute.

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*, 621-628.

Bucar, M., Cuk, I., Pajek, J., Karacsony, I., & Leskosek, 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.

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

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

GymnasticsResultsCom. Gymnastics Results, 2012. Available at: http://www.gymnasticsresults.com; Accessed on 07.01.2012

FIG. (2009). *Code of Points for Women Artistic Gymnastics Competitions.* Retrieved 19 January 2012, 2012, from http://figdocs.sportcentric.net/external/public.php?folder=661

FIG. (2013). *Code of Points for Women Artistic Gymnastics Competitions.* Retrieved 12 May 2014, from https://www.fig-gymnastics.com/site/page/view?id=471

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.

Majer, N. (2013). *Zanesljivost in skladnost sodnic pri sojenju artističnosti na gredi[Reliability and validity of judging artistry at Balance beam]*. Bachelor degree. University of Ljubljana: Faculty of sport.

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

Plessner H. (1999). Expectation biases in gymnastics judging. *J Sport Exercise Psy*, *21*, 131-144.

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

Popović, R. (2000). International bias detected in judging rhythmic gymnastics competition at Sydney-2000 Olympic Games. *Facta universitatis-series: Phys Educ Sport*, *1*, 1-13.

Ste-Marie, D. M. (2000). Expertise in women's gymnastic judging: An observational approach. *Perceptual and Motor Skills*, *90*(2), 543-546.

Ste-Marie, D.M., Valiquette, S.M. & Taylor G. (2001). Memory-influenced biases in gymnastic judging occur across different prior processing conditions. *Res Quarterly Exercise Sport*, *72*, 420-426.

ACKNOWLEDGE

Authors are grateful to Nelli Kim, *FIG* Women's Technical Committee President who supported this study during and after the World Championship in Tokyo 2011.

Corresponding author:

Maja Bučar Pajek
Faculty of Sport
University of Ljubljana
E-mail: maja.bucarpajek@fsp.uni-lj.si