RELIABILITY AND VALIDITY OF JUDGING IN MEN'S ARTISTIC GYMNASTICS AT THE 2009 UNIVERSITY GAMES

Bojan Leskošek[1], Ivan Čuk[1], István Karácsony[2], Jernej Pajek[3], Maja Bučar[1]

[1]University of Ljubljana, Faculty of Sport, Slovenia
[2]Semmelwise University, Budapest, Hungary
[3]University Medical Center, Ljubljana, Slovenia

Original research article

Abstract

Ensuring reliability and validity of judging at artistic gymnastics competitions is difficult. Despite the FIG Men's Code of Points being changed, there is little evidence to show that these changes have had an effect on judging standards. After the last change to the Code of Points (2008) the second biggest men's artistic gymnastics competition took place in 2009 - University Games in Belgrade. Data based on judges' scores were analysed. By last change of the Code of Points the sum of the Difficulty score and the Execution score form the Final score. For the Execution score, which is evaluated by 4 or 6 judges (4-in qualifications and all around, 6 in finals) reliability and validity were calculated (intraclass correlation coefficient, Cronbach's alpha, Kendall coefficient of concordance W, and a theta coefficient; differences in mean E scores between judges were tested using repeated measures ANOVA. All data was analyzed using SPSS Statistics 17.0. Results show very high reliability (e.g. Cronbach alfa range from 0.92 up to 0.99). Systematic bias in individual judge's scores and judges' panels were frequent. Invalidity tends to decrease as competitor numbers increase. Despite good reliability and satisfactory validity of judging at the University Games it should be emphasized that judging quality differs between apparatus, sessions and judges.

Keywords: *men's artistic gymnastics, judging,reliability, validity, university games.*

INTRODUCTION

Evaluation of artistic gymnastics exercises has a long tradition. A gymnasts result is determined by a panel of judges, which should evaluate a gymnastic exercise according to clearly defined rules. Although these rules objectively specify how exercises should be evaluated, evaluation is prone to judges' errors. These errors may be unintentional or sometimes intentional, e.g. at OG 2004 in Athens where head judge was punished for biased decision in men's all around finals. In elite gymnastics the difference between competitors, especially those running for medals, is usually small and small errors can result in a big difference to the final rank of a competitor. Competitors, coaches, spectators, and the media are therefore concerned that judging is of a high standard. The Fédération Internationale de Gymnastique (FIG), which is responsible for the development of sport internationally, is continually trying to implement "fair" rules, which are interpreted by carefully chosen, well educated judges with high ethical standards (FIG, 2009a, 2009b). Some of the most important endeavors of the FIG in this direction were major changes made to the Code of Points in 2006 and the IRCOS project, which allow for evaluation of judge's performances through video analysis.

At the beginning of gymnastics judging only one judge evaluated each gymnast, today this has risen to eight judges

evaluating each gymnast (FIG, 2009c). Women's artistic gymnastics started with the World championships in 1950 following the men's tradition. Today, the Code of Points is similar for women and men in terms of the judges' panel structure and general evaluation guidelines. Both sports have 6 (or 4 for competitions at lower levels than Olympic Games or World Cup) judges evaluating exercise presentations – which results in an E (execution) score and 2 judges evaluating the exercise content – resulting in a D (difficulty) score. The E score decreases from 10 points in decrements of 0.1 point and the D score increases from 0 points up in 0.1 increments. The D score is ratio scale, while the E score is interval scale. Both scales can be used with multivariate analysis, as though especially right censoring (at 10 points) of E score may cause problems in analysis requiring multivariate normal distribution of data.

Due to the D score being a combination of two judges' evaluations reliability and validity cannot be calculated. It is however, possible to calculate reliability and validity for the E score – the average of the middle four (or two). Reliability (also called consistency or repeatability) can be defined as achieving the same results with several measurements of the same subject under identical conditions. A special case of reliability, called inter-rater reliability or objectivity is defined as achieving same results from different persons (judges, assessors, raters, observers) who evaluate the *same* performance. This later aspect of reliability is especially important in gymnastics. As most of the reliability measures are based on interitem (interobserver) correlations, they could not detect validity of judging, i.e. if there is any systematical bias in judging, e.g. systematical under- or overestimation of particular judge or competitors of certain nationalities.

Several authors have tried to evaluate the quality of judging at different competitions. Ansorge, Scheer, Laub, and Howard (1978) found bias in scores induced by the position in which female gymnasts appear in their within-team order. Ansorge and Scheer (1988) found biased judging towards judges' own national team and against immediate competitors' teams. Hraski (1988) analyzed judging at the World Cup in 1982 in all male disciplines; judging for floor exercises was deemed to be the poorest discipline, while still being of an acceptable standard. Duda, Brown, Borysowicz, and St. Germaine (1996) analyzed stress factors of judging; one of many concerns identified was related to the objectivity and reliability of judging. In rhythmic gymnastics, Popović (2000) found biased judging where judges scored gymnasts of their own nationality more favorably. Plessner and Schallies (2005) were determining parallax problems evaluating rings positions; experts were better evaluating position than others. Boen, Van Hoye, Auweele, Feys, and Smits (2008) found that if judges knew other judges scores it resulted in them correcting and adjusting their scores. The FIG Technical Committee is evaluating the quality of judging after all major events. In the past the ranking of gymnasts was the most important information. With the changes to the Code of points resulting in two scores – E and D – in 2006, they started to evaluate E judges by calculating the difference between the final score (average score of middle 4 or 2 judges) and the individual judges score. The 2009 Code of Points (FIG, 2009c) states that judges cannot see other judge's scores before or after they give their own score, but they do see the final E score. The aim of our research was to analyze the reliability and validity of judges' E scores on all apparatus for all sessions (qualification, all round finals, and apparatus finals) at the 2009 World University Games (Universiade) in Belgrade.

METHODS

Judges E scores were obtained from the game's official book of results. To protect judges anonymity we randomly

changed their position in the analysis from the book of results. Three analyses were carried out, one for each session of the competition. The first analyses used data from the qualification sessions on each apparatus. There were four judges for each apparatus. In the qualification session 93 gymnasts performed on the floor, 91 on the pommel horse, 91 on the rings, 113 on the vault, 94 on the parallel bars, and 89 on the high bar. The second analyses used data from the all round finals where 4 judges evaluated E score. In the all round finals 24 gymnasts competed. The third analyses used data from the apparatus finals where 6 judges evaluated E score. In the apparatus finals 8 gymnasts competed, providing 8 sets of scores for each apparatus other than the vault, where 16 sets of scores were available because each competitor performed twice. For each set of analysis we calculated statistics for the E score, item (individual judge) and scale (all judges together) scores. The following reliability and validity statistics were then calculated: intraclass correlation coefficient, Cronbach's alpha, Kendall coefficient of concordance W, and a theta coefficient (Armor, 1974), which is based on the first (largest) eigenvalue from the principal component analysis of the correlations between judges' scores. Differences in mean E scores between judges were tested using repeated measures ANOVA. All data was analyzed using SPSS Statistics 17.0 whenever possible, otherwise using Microsoft Excel.

RESULTS

Mean E scores (Table 1, Figure 1) vary between events, and for some events the data is not normally distributed due to extreme outliers (e.g. rings and high bar during qualification). There is also a large difference in the variability of scores. In general, the smallest variability in all three competition sessions is observed on vault, and the highest in pommel horse. There is a tendency of decreasing variability from first (qualification) to last (apparatus finals) session. The similar pattern of differences in variability between sessions and apparatuses is also evident in central tendency. In all three sessions vault has highest, while rings and pommel horse (except in apparatus finals) have the lowest mean and median values.

Table 1. *Distribution statistics of E score*

session	apparatus	N	M	Me	Min	Max	SD	IQR	Skew.	Kurt.
qualification	Floor	93	8.28	8.40	7.1	9.05	.47	.55	-.85	-.06
	Pommel horse	91	8.09	8.25	3.9	9.5	.90	1.00	-1.63	4.57
	Rings	91	7.69	8.00	0.35	8.85	1.07	.80	-4.03	24.27
	Vault	113	8.86	8.95	7.6	9.4	.36	.35	-1.36	1.58
	Parallel bars	94	8.46	8.60	6.35	9.5	.69	.98	-.84	.19
	High bar	89	8.06	8.30	0.65	9.35	1.08	.73	-4.29	25.78
all round finals	Floor	24	8.62	8.75	6.9	9.15	.44	.23	-2.90	10.65
	Pommel horse	24	7.61	7.85	5.1	8.75	.91	1.10	-1.18	1.36
	Rings	24	8.03	8.15	6.9	8.65	.46	.48	-1.14	.59
	Vault	24	8.89	8.90	7.95	9.6	.35	.40	-.65	1.68
	Parallel bars	24	8.43	8.53	7.6	9.15	.45	.85	-.15	-1.08
	High bar	24	8.22	8.43	6.5	9.15	.67	.89	-1.15	.64
apparatus finals	Floor	8	8.59	8.69	7.975	8.925	.31	.43	-1.13	.94
	Pommel horse	8	8.73	8.75	8.2	9.225	.39	.80	-.04	-1.57
	Rings	8	8.33	8.41	7.575	9.075	.48	.73	-.15	-.16
	Vault	16	9.06	9.15	8.025	9.425	.37	.26	-1.87	3.62
	Parallel bars	8	8.19	8.33	7.175	9.025	.61	.98	-.47	-.51
	High bar	8	8.39	8.60	6.75	8.85	.68	.30	-2.59	6.99

Legend: *N* – no. of performances; *M* – mean; *Me* – median; *Min*, *Max* – lowest and highest value; *SD* – standard deviation; *IQR* – Interquartile range; *Skew.*, *Kurt.* – coefficients of skewness and kurtosis.

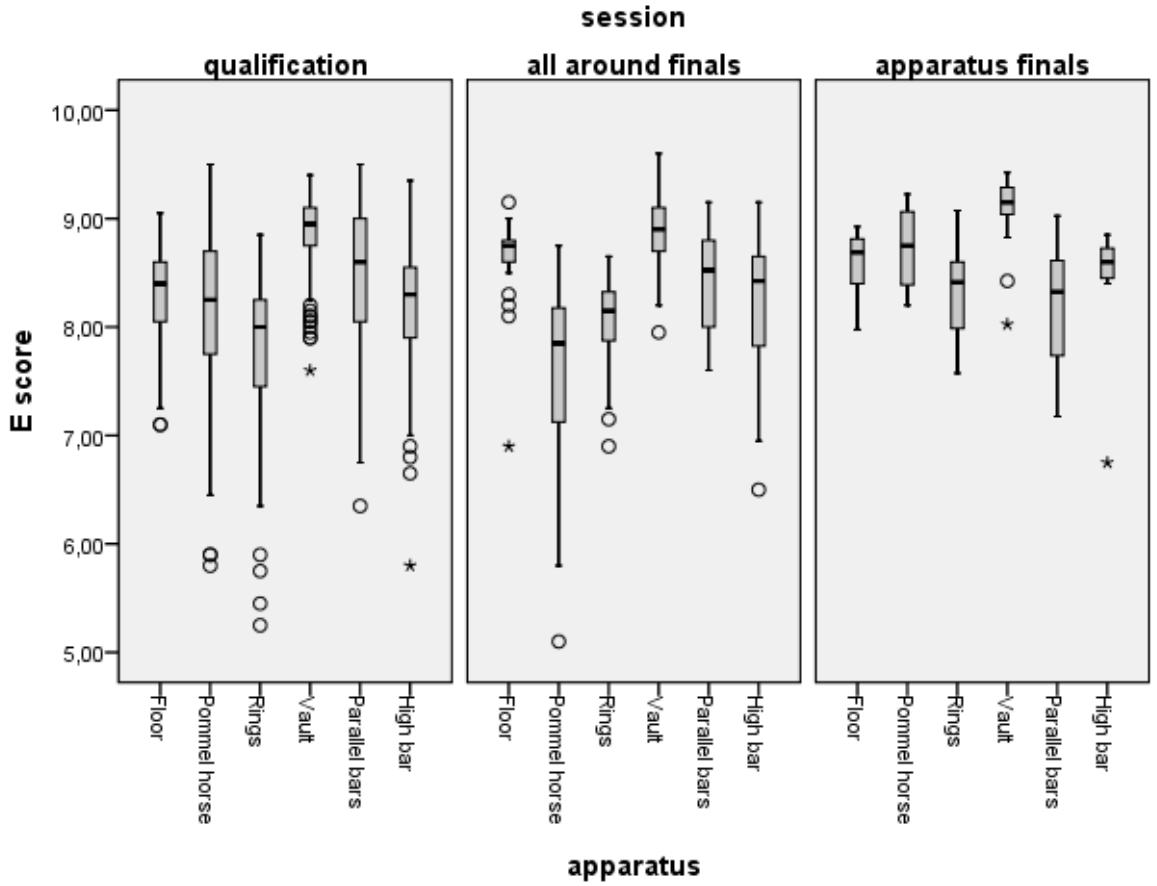

Figure 1. *Boxplots of E score. Note: in qualification, four extreme outliers (E score < 5) are excluded.*

Statistics of scores for individual judges is presented separately for each session and Distributional statistics (mean and standard deviation) was calculated for raw E scores and it's signed and absolute deviation from the final E score. These two forms of deviation are measures of bias (under/over estimation) and reliability of judge's scores. They are also transformed to mean rank (R_{mean}), and it's deviation (dR_{mean}) from expected (unbiased) rank, calculated as follows: (m+1)/2, where m is the number of judges (4 in first two and 6 in the last session). Finally, corrected item-total correlation (r_{corr}) and Cronbach's alpha *if item deleted* ($alpha_{del}$) were calculated.

Table 2. *Statistics of individual judges in the qualification session*

		E score		Dev. E score		Dev. E score abs.					
apparatus	judge	M	SD	M	SD	M	SD	R_{mean}	dR_{mean}	r_{corr}	$alpha_{del}$
Floor	1	8.28	.55	-.01	.22	.16	.15	2.5	.0	.87	.93
	2	8.30	.47	.01	.17	.13	.10	2.5	.0	.89	.92
	3	8.26	.50	-.02	.18	.12	.13	2.4	.1	.88	.93
	4	8.31	.52	.03	.22	.17	.15	2.6	-.1	.85	.93
P. horse	1	8.04	1.04	-.05	.24	.17	.18	2.5	.0	.95	.94
	2	8.00	1.01	-.09	.27	.20	.20	2.3	.2	.94	.95
	3	8.07	.81	-.02	.38	.27	.26	2.5	.0	.87	.97
	4	8.14	.81	.05	.28	.21	.19	2.8	-.3	.92	.95
Rings	1	7.79	1.18	.11	.31	.25	.21	2.9	-.4	.95	.96
	2	7.43	1.00	-.26	.36	.34	.28	1.7	.8	.92	.97
	3	7.75	1.06	.07	.28	.20	.20	2.6	-.1	.94	.96
	4	7.75	1.11	.06	.25	.19	.18	2.7	-.2	.95	.96
Vault	1	8.93	.39	.07	.16	.13	.11	2.9	-.4	.85	.93
	2	8.83	.36	-.03	.14	.10	.10	2.2	.3	.86	.92
	3	8.91	.39	.05	.15	.11	.12	2.9	-.4	.88	.92
	4	8.78	.37	-.08	.16	.12	.14	2.0	.5	.85	.92
Par. bars	1	8.42	.74	-.04	.15	.12	.10	2.4	.1	.96	.96
	2	8.47	.71	.00	.15	.11	.10	2.5	.0	.96	.96
	3	8.54	.64	.07	.23	.17	.17	2.8	-.3	.92	.98
	4	8.43	.75	-.03	.21	.15	.16	2.3	.2	.94	.97
High bar	1	8.08	1.10	.02	.21	.16	.14	2.6	-.1	.97	.98
	2	8.07	1.15	.01	.19	.14	.13	2.6	-.1	.98	.98
	3	8.06	1.07	.00	.23	.17	.15	2.5	.0	.96	.98
	4	8.02	1.10	-.04	.25	.16	.20	2.2	.3	.96	.98

Legend: R_{mean} mean rank; dR_{mean} deviation of R_{mean} from expected rank; r_{corr} corrected item-total correlation; $alpha_{del}$ Cronbach alpha if item deleted

Table 3. *Statistics of individual judges in the all round finals*

		E score		Dev. E score		Dev. E score abs.					
apparatus	judge	M	SD	M	SD	M	SD	R_{mean}	dR_{mean}	r_{corr}	$alpha_{del}$
Floor	1	8.58	.54	-.04	.18	.13	.14	2.3	.2	.90	.94
	2	8.69	.45	.07	.14	.10	.12	2.9	-.4	.93	.93
	3	8.60	.46	-.02	.14	.09	.12	2.6	-.1	.90	.94
	4	8.56	.50	-.06	.22	.14	.18	2.2	.3	.85	.95
P. horse	1	7.40	1.16	-.20	.40	.30	.33	1.8	.7	.94	.96
	2	7.63	1.11	.03	.33	.25	.21	2.7	-.2	.94	.96
	3	7.65	.79	.05	.22	.16	.15	2.6	-.1	.96	.96
	4	7.70	.89	.10	.21	.18	.15	2.9	-.4	.94	.96
Rings	1	8.08	.61	.05	.31	.24	.21	2.9	-.4	.77	.93
	2	8.05	.47	.03	.21	.14	.16	2.7	-.2	.86	.89
	3	7.94	.50	-.09	.20	.17	.14	2.0	.5	.83	.90
	4	8.00	.45	-.03	.17	.14	.11	2.4	.1	.87	.89
Vault	1	8.88	.36	-.01	.14	.10	.10	2.6	-.1	.85	.89
	2	8.83	.31	-.06	.12	.10	.08	2.0	.5	.89	.89
	3	8.90	.36	.01	.21	.11	.18	2.4	.1	.75	.93
	4	8.96	.41	.07	.16	.14	.11	3.0	-.5	.84	.90
Par. bars	1	8.38	.56	-.05	.19	.14	.13	2.2	.3	.90	.89
	2	8.50	.39	.08	.25	.19	.19	2.9	-.4	.77	.94
	3	8.37	.49	-.06	.24	.18	.17	2.2	.3	.81	.92
	4	8.45	.46	.02	.11	.08	.08	2.7	-.2	.93	.88
High bar	1	8.25	.69	.03	.24	.20	.13	2.8	-.3	.89	.94
	2	8.27	.85	.05	.29	.23	.18	2.8	-.3	.93	.93
	3	8.12	.66	-.10	.22	.18	.15	2.1	.4	.91	.94
	4	8.19	.68	-.02	.26	.22	.15	2.3	.2	.86	.95

Legend: R_{mean} mean rank; dR_{mean} deviation of R_{mean} from expected rank; r_{corr} corrected item-total correlation; $alpha_{del}$ Cronbach alpha if item deleted

Table 4. *Statistics of individual judges in the apparatus finals*

		E score		Dev. E score		Dev. E score abs.					
apparatus	judge	M	SD	M	SD	M	SD	R_{mean}	dR_{mean}	r_{corr}	$alpha_{del}$
Floor	1	8.74	.41	.15	.24	.24	.13	4.5	-1.0	.72	.94
	2	8.48	.28	-.11	.11	.13	.10	2.6	.9	.91	.91
	3	8.39	.39	-.20	.22	.28	.08	1.8	1.8	.73	.93
	4	8.69	.39	.10	.16	.15	.10	4.4	-.9	.89	.91
	5	8.63	.17	.04	.18	.11	.14	3.6	-.1	.90	.93
	6	8.66	.40	.08	.12	.11	.08	4.1	-.6	.96	.90
P. horse	1	8.80	.38	.07	.12	.11	.09	4.0	-.5	.94	.91
	2	8.79	.23	.06	.23	.17	.16	3.4	.1	.82	.94
	3	8.85	.32	.12	.15	.15	.12	4.3	-.8	.91	.92
	4	8.74	.54	.01	.23	.21	.06	3.8	-.3	.88	.92
	5	8.59	.55	-.14	.27	.22	.20	2.9	.6	.80	.93
	6	8.49	.65	-.24	.33	.28	.29	2.6	.9	.89	.92
Rings	1	8.26	.40	-.07	.21	.14	.16	2.9	.6	.86	.96
	2	8.36	.48	.03	.19	.11	.14	3.4	.1	.89	.95
	3	8.39	.40	.06	.20	.16	.11	3.8	-.3	.88	.96
	4	8.24	.57	-.09	.24	.19	.16	3.1	.4	.88	.95
	5	8.25	.64	-.08	.25	.19	.17	3.6	-.1	.91	.95
	6	8.43	.53	.09	.15	.11	.14	4.1	-.6	.93	.95
Vault	1	8.94	.46	-.12	.17	.15	.14	2.7	.8	.92	.98
	2	9.11	.38	.04	.10	.09	.07	4.1	-.6	.95	.97
	3	9.06	.32	-.01	.11	.08	.07	3.2	.3	.94	.97
	4	9.02	.40	-.05	.11	.09	.08	2.8	.7	.94	.97
	5	9.13	.33	.07	.09	.09	.07	4.3	-.8	.95	.97
	6	9.10	.45	.04	.14	.10	.10	4.0	-.5	.94	.97
Par. bars	1	8.15	.63	-.04	.34	.24	.22	3.1	.4	.78	.96
	2	8.01	.67	-.18	.36	.26	.31	3.1	.4	.81	.96
	3	8.21	.84	.02	.30	.26	.11	4.2	-.7	.94	.95
	4	8.23	.62	.03	.14	.11	.09	3.6	-.1	.96	.95
	5	8.14	.64	-.06	.18	.16	.08	2.8	.7	.95	.95
	6	8.30	.56	.11	.22	.19	.15	4.3	-.8	.90	.95
High bar	1	8.45	.55	.06	.18	.13	.12	3.8	-.3	.97	.98
	2	8.44	.80	.04	.19	.14	.13	3.7	-.2	.97	.98
	3	8.28	.75	-.12	.20	.19	.12	3.0	.5	.94	.98
	4	8.51	.52	.12	.21	.16	.17	4.4	-.9	.96	.98
	5	8.49	.81	.09	.21	.16	.16	4.2	-.7	.96	.98
	6	8.20	.61	-.19	.15	.21	.13	1.9	1.6	.97	.98

Legend: R_{mean} mean rank; dR_{mean} deviation of R_{mean} from expected rank; r_{corr} corrected item-total correlation; $alpha_{del}$ Cronbach alpha if item deleted

Pearson's correlations between judges (Table 5) are, in the main, higher than .8. There are few exceptions, usually when the number of competitors in an event is low. One very low correlation (.39) in the floor apparatus finals between judges 1 and 3 is due to judge no. 3 awarding the highest score in this event to the competitor who finished in 7th (second last) place. Without this score correlation would be .81. Except for the vault qualifications and parallel bars all round finals, the floor has the lowest correlations between judges, with horizontal bar providing the highest correlations. In general, average correlation is the highest in qualification session and lowest in the all round finals, with two clear exceptions (high correlations in pommel horse all round finals and vault apparatus in the finals).

Table 5. *Pearson correlation coefficients between judges' E scores*

		qualification				all around finals				apparatus finals					
	judge	1	2	3	4	1	2	3	4	1	2	3	4	5	6
Floor	1		.85	.81	**.79**		.91	.89	.78		.65	**.39**	.73	.75	.87
	2	.85		.83	**.79**	.91		.84	.86	.65		.88	.79	.90	.90
	3	.81	.83		.82	.89	.84		.83	**.39**	.88		.75	.79	.71
	4	**.79**	**.79**	.82		.78	.86	.83		.73	.79	.75		.77	.90
	5									.75	.90	.79	.77		.84
	6									.87	.90	.71	.90	.84	
Pommel horse	1		.94	.87	.92		.90	.93	.91		.77	.86	.81	.88	.90
	2	.94		.85	.90	.90		.94	.92	.77		.92	.90	**.60**	.70
	3	.87	.85		.83	.93	.94		.94	.86	.92		.90	.71	.83
	4	.92	.90	.83		.91	.92	.94		.81	.90	.90		.70	.83
	5									.88	**.60**	.71	.70		.78
	6									.90	.70	.83	.83	.78	
Rings	1		.90	.93	.92		.73	**.69**	.75		.81	.85	.75	.86	.77
	2	.90		.88	.90	.73		.83	.83	.81		.75	.83	.82	.93
	3	.93	.88		.93	**.69**	.83		.81	.85	.75		.80	.85	.87
	4	.92	.90	.93		.75	.83	.81		.75	.83	.80		.86	.87
	5									.86	.82	.85	.86		.85
	6									.77	.93	.87	.87	.85	
Vault	1		**.80**	**.79**	**.79**		.82	.73	.78		.90	.90	.90	.87	.88
	2	**.80**		.83	**.77**	.82		.71	.87	.90		.89	.94	.93	.90
	3	**.79**	.83		.82	.73	.71		**.67**	.90	.89		.88	.93	.94
	4	**.79**	**.77**	.82		.78	.87	**.67**		.90	.94	.88		.93	.89
	5									.87	.93	.93	.93		.93
	6									.88	.90	.94	.89	.93	
Parallel bars	1		.95	.91	.92		.74	.79	.93		**.57**	.76	.86	.77	.70
	2	.95		.90	.93	.74		**.66**	.78	**.57**		.85	.81	.82	.71
	3	.91	.90		.88	.79	**.66**		.81	.76	.85		.90	.89	.89
	4	.92	.93	.88		.93	.78	.81		.86	.81	.90		.92	.91
	5									.77	.82	.89	.92		.95
	6									.70	.71	.89	.91	.95	
High bar	1		.96	.95	.94		.86	.88	.80		.95	.93	.94	.97	.94
	2	.96		.96	.96	.86		.90	.87	.95		.90	.93	.98	.96
	3	.95	.96		.94	.88	.90		.80	.93	.90		.97	.91	.94
	4	.94	.96	.94		.80	.87	.80		.94	.93	.97		.92	.94
	5									1	.98	.91	.92		.93
	6									.94	.96	.94	.94	.93	

Note: coefficients lower than .8 in qualification session, .7 in all around finals and .6 in apparatus finals are printed in bold

Table 6. *Reliability and validity measures of competition events*

Session	apparatus	Alpha	$ICC_{average}$	ICC_{single}	Armor's theta	Kendall W	ANOVA F
qualification	Floor	.94	.94	.81	.95	.01	.92
	P. horse	.96	.96	.87	.97	.02	2.83
	Rings	.97	.97	.88	.98	.20	23.06
	Vault	.94	.93	.77	.94	.16	19.51
	Par. bars	.98	.98	.91	.98	.03	5.70
	High bar	.99	.99	.95	.99	.02	1.25
all around finals	Floor	.96	.95	.84	.96	.06	2.08
	P. horse	.97	.97	.88	.98	.14	3.92
	Rings	.92	.92	.75	.93	.08	1.47
	Vault	.92	.92	.74	.93	.13	2.27
	Par. bars	.93	.93	.77	.94	.08	1.87
	High bar	.95	.95	.84	.96	.07	1.30
apparatus finals	Floor	.93	.91	.62	.95	.36	3.92
	P. horse	.94	.92	.67	.96	.13	2.49
	Rings	.96	.96	.80	.97	.06	1.01
	Vault	.98	.97	.86	.98	.16	4.24
	Par. bars	.96	.96	.81	.97	.12	.89
	High bar	.98	.98	.89	.99	.26	2.92

Note: coefficients W and F that are significantly different from zero at p<.05 are underlined.

DISCUSSION AND CONCLUSIONS

As many of the reliability measures of judges' performances are based on Pearson's correlations (*r*) it's important to evaluate these before evaluating derived measures. The size and sign of *r* can be heavily affected by the presence of outliers, especially if the number of outliers is high compared to the total number of cases e.g. in the high bar apparatus finals (8 competitors) *r* between the first two judges is .95; if we omit the outlier (competitor with score 7.1 and 6.5 given from first and second judge, respectively), *r* is not only much lower but also of negative sign (*r*=–.33). As a consequence of this, ICC changes from very high (.98) to moderate (.63).

Despite this, indices of reliability are generally quite high. In different sessions and apparatus all reliability measures (Cronbach's alpha, ICC, Armor's theta) are higher than .90. Those indices tend to be a little lower in the all round finals than in qualification and apparatus finals. There appears to be no systematic differences in reliability between apparatus. Vault scores tend to have lower reliability than other apparatus in qualification and all round finals, but not in apparatus finals. High bar scores have the highest reliability in qualification session and apparatus finals, but only average in all around finals.

Although these results are not directly comparable with results from the 1982 World Cup in Zagreb (Hraski, 1988) it seems that reliability is improving over time, and through the introduction of new rules, especially splitting judges' panel into judges for exercise presentation and exercise content. In Zagreb, only 20 gymnasts competed, all in one session; they were evaluated by 5 judges (head judge and four score judges), which were judging exercise difficulty and exercise presentation together; Armor's theta ranged from .92 (on the floor) to .98 (rings and high bar), whereas in Belgrade Armor's theta ranged from .93 (rings and vault all round finals) to .99 (high bar qualifications and apparatus finals).

High reliability of E scores is not always accompanied by high validity. Systematic bias in individual judge's scores (as measured by dR_{mean}, a deviation of mean rank from expected unbiased rank) and judges' panels (as measured by Kendall W and ANOVA F) were frequent.

Surprisingly, the second highest dR_{mean} (1.6) appeared in the event with the highest reliability (i.e. high bar finals). Poor validity tended to decrease as the number of competitors increased, this was particularly evident in the apparatus finals, where each judge only gives 8 scores (16 on vault). It's seems that in sessions with more competitors, judges have an opportunity to adjust their criterion of judgment after the first few competitors.

Despite good reliability and generally satisfactory validity of judging at the University Games it should be emphasized that the quality of judging differs between apparatus, sessions, and individual judges. There are numerous objective and subjective factors for these differences e.g. the number of competitors in a session, judge's seat positions and view angle to the gymnast, and the judge's experience. At the moment as there is only sum of deductions presented in the judge's score it would be advisable if E judges could be evaluated according to what deduction was taken in time of gymnast's exercise. Such a computerized system already exists (Čuk and Forbes, 2006) and it would be good if it could be tested to overcome significant differences in judge's scores.

REFERENCES

Armor, D. J. (1974). Theta reliability and factor scaling. In H. Costner (ed.), Sociological methodology (pp. 17-50). San Francisco: Jossey-Bass.

Ansorge, C.J, Scheer J.K., Laub, J., and Howard, J (1978). Bias in Judging Women's Gymnastics Induced by Expectations of Within-Team Order. *Research Quarterly*, 49(4), 399-405.

Ansorge, C.J. and Scheer, J.K. (1988). International bias detected in Judging Gymnastics Competition at the 1984 Olympic Games. *Research Quarterly for Exercise and Sport*, 59(2), 103-107.

Boen, F., van Hoye, K., Vanden Auweele, Y., Feys, J. and Smits, T. (2008). Open feedback in gymnastic judging causes conformity bias based on informational influencing. *Journal of Sports Sciences*, 26 (6), 621-628.

Čuk, I. and Forbes, W. (2006). Kam greš, sojenje? In E. Kolar, Piletič S. (eds.), Gimnastika za trenerje in pedagoge 2 (pp. 76-86). Ljubljana: Gimnastična zveza Slovenije.

Duda, J.L., Brown Borysowicz, M.A., and St. Germaine, K. (1996). Women's Artistic Gymnastics Judges' Sources of Stress. *Technique*, 16 (10), 1-5.

FIG (2009a). 2009 FIG General Judges' Rules. FIG. Retrieved December 3, 2009, from *http://figdocs.lx2.sportcentric.com/external/serve.php?document=658.*

FIG (2009b). 2009 FIG Judges' Rules Specific Rules For Men's Artistic Gymnastics. FIG. Retrieved December 3, 2009, from *http://figdocs.lx2.sportcentric.com/external/serve.php?document=529*.

FIG (2009c). Code of Points for Men Artistic Gymnastics Competitions. FIG. Retrieved October 1, 2009, from *http://figdocs.lx2.sportcentric.com/external/serve.php?document=1205*.

Hraski, Ž. (1988). Valorizacija suđenja u muškoj sportskoj gimnastici. *Kineziologija*, 20(2), 143-152.

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

Popović, R. (2000). International Bias Detected in Judging Rhythmic Gymnastics Competition at Sydney 2000 Olympic Games. *Facta Universitatis* (Series: Physical Education and Sport), 1(7), 1–13.