CAN AUDIENCE REPLACE EXECUTION JUDGES IN MALE GYMNASTICS?

Ivan Čuk

Faculty of Sport, University of Ljubljana, Slovenia

Original article

Abstract

The aim of research was to test the quality of spectator's execution evaluation of gymnast's performance compared to official judges. For the purpose 91 participants (spectators) evaluated 26 exercises on parallel bars on a scale from 0-10.0 points like it is used in diving (0-Completely failed, 0.5-2.0 – unsatisfactory, 2.5-4.5 – deficient, 5.0-6.5 – satisfactory, 7.0-8.0 – good, 8.5-9.5 - very good, and 10 – excellent). Following analysis were performed: Kolmogorov Smirnov test, Cronbach's Alpha, Spearmans Rank correlation, cluster analyze (square distance method), and Kruskal Wallis analyze of variance. Spectator's reliability of judging is same as for official judges; ranking is similar to the official judges and even better. Cluster analysis extracted three groups of spectators – strict ones, medium ones and permissive ones. As an average, they function well. The biggest challenge for spectators is bias towards their local (national) heroes and champions. However, in the spectators is also a group of those who are honest and strict without having biased opinion and they formed quite a big group (one quarter).

Keywords: *Judging, Code of Points, Spectators, Reliability, Validity.*

INTRODUCTION

Since the first world championship in gymnastics in 1903 judging have changed severely (Bučar 1998, Grossfeld 2014). In the beginning two judges were evaluating gymnasts, later number of judges raised to three, four, six, nine, while in 2014 there are 5 judges for execution, two for difficulty and 2 as a reference judges (Štukelj, 1989, FIG, 2014). In the past there were some extreme events, where judging was not in line with spectators. Such most referred event was during World Championship in Praque in 1962, when Boris Schahklin (Soviet Union, later Ukraine) was on parallel bars awarded with the highest score, while in public spectator's opinion Miroslav Cerar (Yugoslavia, later Slovenia) should win (Satler, Šlamberger, 1966). Whistling and clapping was so noise and long term, judges have changed their score and gymnasts tied for gold. Similar situation happened at Olympic Games in Athens 2004 when spectators did not accept Alexei Nemov (Russia) score on high bar. Judges changed the score but in such way, it did not change the ranking of the gymnasts.

At the major competitions, usually spectators know the sport they attend and they award the value (exercise as a whole) of the gymnasts exercise with appropriate

clapping; of course their heart beats with their best local (national) gymnasts and they are slightly biased what can be noticed by lauder cheering and clapping at the end of the exercise.

Judging is difficulty task (Plessner, 1999). It divides into two areas of work at the competition. The first task is to define content value (difficulty, special requirements and connections) and the second task is to define execution value (FIG, 2013). To evaluate content value it is necessary to memorize a lot of information e.g. - how element is defined (starting position, movement, final position), how difficult it is; in which group it belongs; what is the symbol of element (writing is compulsory) and at the end judge have to decide which elements will be count for content value. Content value is as an objective measure (Fink, 1986; Fink, Fetzer 1992) and can be controlled via video recording system IRCOS) in case of doubt (FIG, 2011b). On other side, judging execution is much simpler, in sense of amount of information needed to evaluate exercise. There are only four main deductions to apply – small one 0.1 point, medium one 0.3 point, large 0.5 point and fall 1.0 point (FIG, 2013)

By FIG Men's Code of Points (2013)(which includes more than 800 elements with different difficulties, special requirements, connection bonus, what is hard to memorize even for judges) it is evident spectators do not have proper information (in sense of content) what gymnasts are performing. Spectators can hardly define difficulty value; even experts without writing down whole exercise cannot always tell the exact score. Execution is easier to define by spectators. At least they can say the gymnasts succeeded or not. During our school days, we are and we evaluate on at least five level scale – unsatisfactory, satisfactory, good, very good, and excellent. So also, spectators (with some school experience) can give remarks according to their own criteria, what is e.g. excellent.

In the past many judges analysis were performed about their validity, reliability, bias, some of the ideas FIG accepted, some are still waiting to be applied (as Real time Judging System, Bučar Pajek et.al. (2011). Bias is in gymnastics quite common, Plessner (1999) found that placing of gymnast within the team competition have impact on its score. Similar found Ste-Marie (2003) that memory biases are evident according to previous knowledge of gymnasts. Another Ste- Marie (2000) analyses showed that novice judges, as compared to expert judges, spent less time looking at the gymnast perform, spent more time looking at the scoring paper, and were less able to engage in the dual-task demands required in gymnastic judging. Yet Another Ste-Marie analysis showed expert judges were significantly better at perceptually anticipating upcoming gymnastic elements from advance information; also, gymnastic elements that were correctly anticipated were judged more accurately than those that had been anticipated incorrectly; experts also exhibited significantly greater depth and breadth in their declarative knowledge base. In general we can conclude, judging in gymnastics is in general good (uniform judging education, FIG level of judges) according to reliability and validity of judges (Dallas, Kirialanis, 2010; Bučar Pajek et.al. 2012). Of course some problems persist, especially in area of execution where decisions have to distinguish between small and medium error and between medium and large error.

Reliability is defined how measurement results can be repeated with same results (IBM, 2013). It is also calculated how different results of same subject (from different evaluators, judges, spectators) are related. Cronbach's Alpha coefficient is a measure, which tells if there are reliably results. In general, Cronbach's Alpha higher of 0.9 is a reliably measurement. Reliability can be raised with higher number of measurements of subject (what FIG already done from two to 9 judges), but it would be of special interest how few tens of spectators are reliably. Up to now nobody

tested spectators how they evaluate gymnastics exercises.

Some sports with similar approach as gymnastics (difficulty and execution) have, also simplified execution evaluation (Čuk, Fink, Leskošek, 2012). The simplest is in diving where execution is evaluated on the scale from 0 to 10 (FINA, 2011). Adults understand and could use 0-10 level scale of diving in most of life situations.

Diving execution values:

0 - completely failed
0.5 - 2.0 - unsatisfactory
2.5 - 4.5 - deficient
5.0 - 6.5 - satisfactory
7.0 - 8.0 - good
8.5 - 9.5 - very good
10 – excellent.

Before OG in Beijing there were serious attempts to include audience to evaluate sport events and judge athletes performance, even some technological solutions were prepared, however the idea did not live and we have no data on how audience is evaluating athletes performance (Biggar et al., 2005).

The aim of research is to find out how spectators can evaluate gymnasts performance comparing to judges in ranking gymnasts and in compare judges reliability and spectators reliability.

METHODS

For experiment, we choose 91 participants (62 males and 29 females) - students of the first year at university, before they took any sport course. Average age was 19.5 years. They were not involved with gymnastics otherwise as in primary or secondary school during P.E: classes or as TV spectators.

Each participant received a paper with written guidelines what scores from 0-10 mean, and the table with three columns. In the first column there was number of gymnast presented on video (numbers from 1- 26), the second column was for participants remarks and the third column was for the participant score, they were not instructed to work as judge but to watch exercise from the beginning to the end, make short notes and give result.

Gymnast's exercises were from parallel bars only. We choose whole competition video recording and results from World Cup competition (FIG, 2011a). Participants had to evaluate 26 exercises. After the first gymnasts dismount, participants had 45 second to give their score (at competitions - 30 seconds for judges plus 15 seconds, as is time of results show results on panel), than next gymnasts, exercise was showed. Order of the gymnasts was same as at competition judges evaluated (if exist to have same order bias with spectators).

Participants seated apart and conversation forbidden.

Sample of variables consisted of: official judge's results (six judges scores (variables J1 to J6, final E score - EFINAL), average score from all spectators - VIEWER, scores from each spectator.

For data analysis, we used SPSS 22.0. Following analysis were performed: Kolmogorov Smirnov test, Cronbach's Alpha, Spearmans Rank correlation, hierarchical cluster analyze (between groups linkage and interval of squared Euclidean distance method), and Kruskal Wallis analyze of variance. Results were significant at $p<0.05$.

RESULTS AND DISCUSION

Kolmogorov Smirnov test showed only 24 spectators succeeded to make normal scores distribution. However, also only two judges have normal distribution of scores as well. Despite the fact most of data have not normal distribution, we calculated Cronbach's Alpha as a measure of reliability. Judges and spectators have similar reliability (0.994 versus 0.997), and comparative to results in past (Bučar Pajek et al., 2012; Leskošek, et al., 2010).

Table 1
Spearman Rank Correlation matrix

	VIEWER	EFINAL	J1	J2	J3	J4	J5	J6
VIEWER	1.000	.927**	.759**	.885**	.933**	.901**	.892**	.878**
EFINAL		1.000	.747**	.946**	.951**	.979**	.964**	.942**
J1			1.000	.811**	.704**	.732**	.673**	.699**
J2				1.000	.913**	.921**	.846**	.893**
J3					1.000	.927**	.891**	.902**
J4						1.000	.959**	.908**
J5							1.000	.914**
J6								1.000

Rank correlations in Table 1 are all significant and very high. Average spectators score shares same ranking as other judges comparing towards final execution score. Even more spectators were much better in ranking than judge number 1. In addition, rank correlations between judges and average spectator are of the same level, while some are again much better than between judge one and other judges.

From the statistics point of view, spectators can even replace international judge, spectators reliability is adequate (number of measurements (judges) is very high and reliability increases). Even ranking is adequate, as average rank correlation between judges and final execution score is 0.921.

Further, we were interested if spectators are homogenous group. For a purpose we did cluster, analyze with groups from two to 10. The Figure

1 shows dendrogram of hierarchical cluster results. The best solution was with five groups where the first group had 59 spectators, the second 7 the third 22, the fourth 1 and the fifth 2. In next iteration with four groups the first group was composed of 81 spectator (59 from previous the first group and 22 from previous the third group).. For further analysis, the solution with five groups have been used, as the fourth group had only one spectator, statistics is not available, and the fifth group of two spectators was omitted in further statistics as number was too low. In Table 2 are averages from each, group, their standard deviations and standard errors for each evaluated gymnast. Unintentional perhaps it was good first gymnast scored by spectators opinion satisfactory value, what put spectators in position to adjust their scores later on with upgrading or downgrading their scores. As a rule, the first group have medium scores, the second group the highest scores and the third group have the lowest scores. Some particular results are interesting; the second group gave in average 10 points to Petkovšek and Fahrig, while in official results the highest score was for Wang. Petkovšek as a local hero (recognized by spectators) have higher ranking than Wang for the first and second group, while the third group placed Petkovšek and Wang as judges placed them. Kruskal Walis test showed for all gymnasts that spectators groups have different opinion, while only for Wang it showed there are all three groups of same opinion.

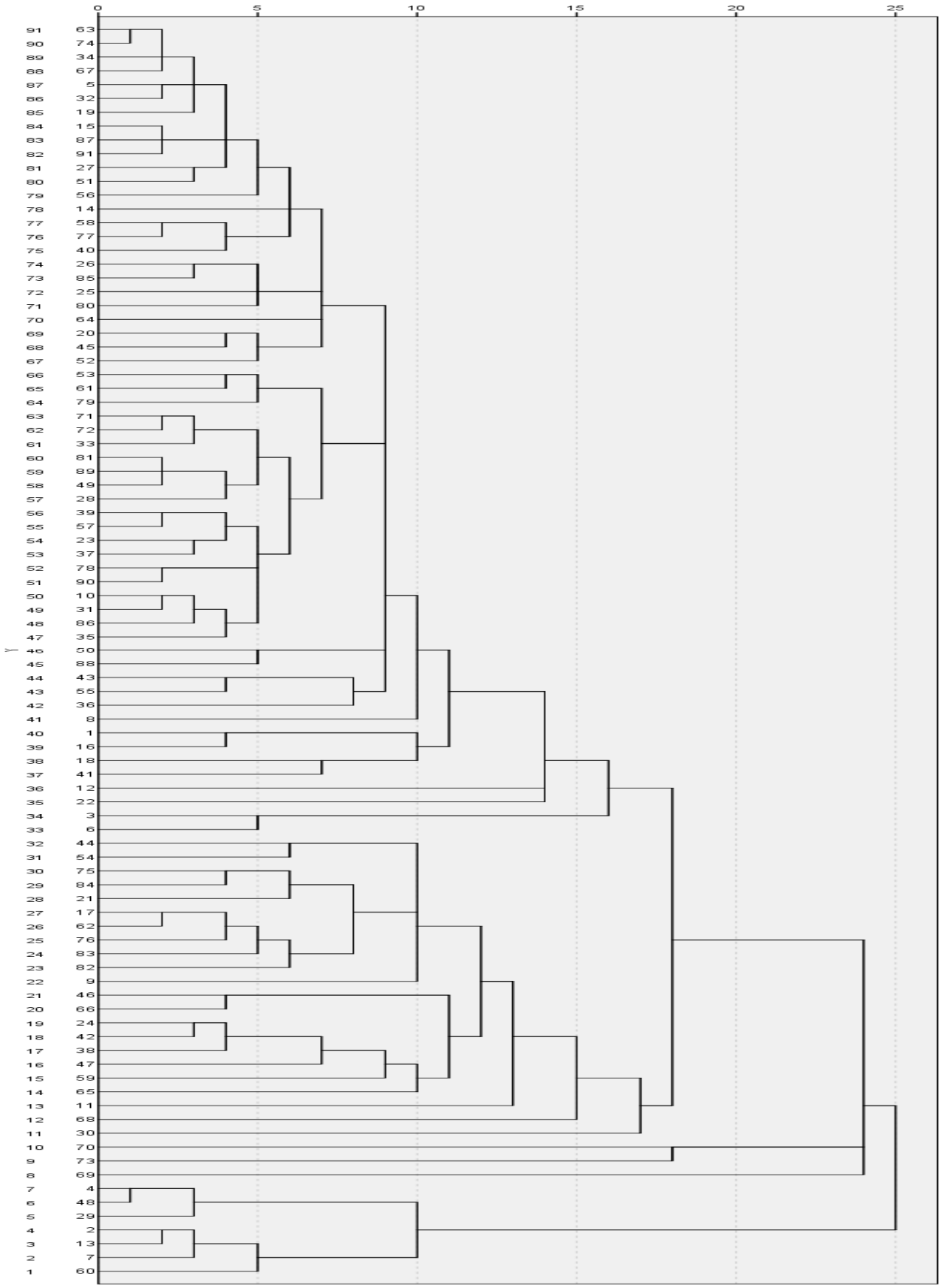

Figure 1. Dendrogram using average linkage between groups.

Table 2
Average Scores from different participants group groups (participants used 0-10 scale without introducing into gymnastics judging principles)

	N=59		N=7		N=22		N=88	
Gymnast	XA1	SE1	XA2	SE2	XA3	SE3	XAtot	SEtot
Leimlehner_AUT	7.33	.133	7.85	.508	5.18	.306	6.84	.160
Krasias_CYP	2.47	.195	4.57	.868	1.22	.262	2.32	.183
Batinkov_BUL	7.49	.193	8.71	.285	5.90	.353	7.19	.179
Babos_HUN	8.64	.122	9.42	.202	7.90	.227	8.52	.109
Alsadi_QAT	5.81	.164	7.71	.521	4.45	.320	5.62	.167
Bretschneider_GER	8.94	.103	9.57	.202	8.31	.190	8.84	.092
Masri_LUX	7.20	.146	9.00	.218	5.40	.204	6.89	.153
Palgen_LUX	8.27	.158	9.57	.297	5.95	.363	7.79	.184
Ishikawa_JAP	8.64	.089	9.28	.184	8.18	.214	8.57	.086
Behan_IRL	7.74	.140	8.85	.404	6.50	.252	7.52	.136
Neuteleers_BEL	8.54	.119	9.71	.184	7.09	.321	8.27	.138
Piasecky_SVK	4.15	.174	7.14	.142	3.09	.185	4.12	.164
Vovk_BUL	8.71	.128	9.71	.184	7.90	.185	8.59	.110
Kulesza_POL	7.69	.137	8.71	.420	7.00	.294	7.60	.129
Aoyama_JAP	9.42	.080	9.71	.184	8.77	.196	9.28	.080
Abdulrada_KUW	7.35	.159	8.57	.368	6.00	.294	7.11	.152
Britovsek_SLO	3.54	.201	6.14	.260	2.50	.225	3.48	.175
Kallai_HUN	7.86	.121	8.85	.260	7.09	.293	7.75	.121
Petkovsek_SLO	9.83	.049	10.0	.000	9.54	.143	9.77	.050
Neczli_SVK	7.16	.123	8.85	.142	4.77	.293	6.70	.169
Kierzkowski_POL	7.03	.132	8.28	.184	5.59	.370	6.77	.151
Wang_CHN	9.50	.097	9.85	.142	9.68	.121	9.57	.073
Alsaffar_KUW	5.30	.221	7.71	.420	3.00	.254	4.92	.213
Fahrig_GER	8.25	.161	10.0	.000	7.81	.260	8.28	.138
Decker_AUT	8.06	.107	9.71	.184	6.63	.233	7.84	.128
Bahlawan_SYR	1.08	.135	4.42	.972	0.50	.215	1.20	.164

Legend: N-numerus, XA – average, SE-standard error, 1,2,3-groups, tot-total

It is important to note, each spectator made his own criteria what goes with his scores. The scores were then going up or down in accordance with gymnasts presentation. As spectators were not educated gymnasts or coaches or judges, we can say they were consistent via their own scores. High number of spectators represented as items increased reliability and validity. Despite Ste-Marie (2003) results between novice and expert judges, we can not compare them in same sense; judges did have instructions and they have to evaluate according to the Code of Points with small, medium, large and fall errors, what is very different to evaluation of whole exercise within simple 0-10 points. The content of score is different.

CONCLUSIONS

Experiment with spectators to serve as a judge showed interesting results. Reliability of exercise presentation judging is same as for official judges; ranking is similar to the official judges and even better. Cluster analysis extracted three groups of spectators – strict ones, medium ones and permissive ones. As an average, they function well. The biggest challenge for spectators is bias towards their local (national) heroes and champions, but in the spectators is also a group of those who are honest and strict without having biased opinion. They form quite a big group (one quarter).

With modern technology, e.g. smart mobile phones FIG could perform some experimental judging among spectators. With further data collection and data analysis, perhaps some new horizon open as there is not much sports events by now, where spectators could have active role in competition.

REFERENCES

Biggar D., Hoffman A. Jordan C., & LasnikV. (2005). *DiscAthon: A Rating Device for Olympic Events.* Retrived from: www.uxgoodbadugly.com/papers/DiscAThon_OlympicRatingDevice.doc.

Bučar Pajek, M. *(1998) Primerjalna analiza tekmovalnih pravil v moški in ženski športni gimnastiki [Comparative Analsis of Men and Women Code of Points in Gymnastics]: diplomsko delo*. Ljubljana: Fakulteta za šport.

Bučar Pajek, M., Čuk, I., Karacsony, I., Pajek, J., & Leskošek, B. (2012). Reliability and validity of judging in Women' artistic gymnastics at the 2009 University games. *European Journal of Sport Science*, 3, 207-215.

Bučar Pajek, M., Forbes, W., Pajek, J., Leskošek, B., & Čuk, I.(2011). Reliability of real time judgeing system. *Science of Gymnastics Journal*, *3*(2), 47 – 54.

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*(1), 73-82.

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

Fink, H. (1986). *Towards an 'Alternative' Code of Points for Men and Women.* Rome: FIG Judges' Symposium The Future of Gymnastics.

Fink, H., & Fetzer, J. (1992). *Code of Points for Men's Artistic Gymnastics (draft proposal for implementation after 1996).* Vancouver: FIG-MTC.

FIG. (2013). *MAG Code of Points 2013-2016*. Lausanne: FIG.

FIG. (2011a). *Results from World Cup competition in Maribor*. Lausanne: FIG.

FIG. (2014). *Technical Regulations*. Lausanne: FIG.

FIG. (2011b). Use of IRCOS decision. Lausanne: FIG.

FINA (2011). Fina Rules and regulations. Retrived on 15.4.2011 from http://www.fina.org/H2O/index.php?option=com_content&view=section&id=17&Itemid=184

Grossfeld A. (2014). Changes during the 110 years of the World Artistic Gymnastics Championships. *Science of Gymnastics Journal*, *6*(2) 5-37.

IBM. *SPSS 22.0*, 2013.

Leskošek, B., Čuk, I., Karacsony, 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*(2), 25-34.

Plessner, H. (1999). Expectation biases in gymnastics judging. *Journal of Sport & Exercise Psychology, 21*(2), 131-144.

Satler M., Šlamberger V. (1966). *Poslednji romantik na konju [Last romantic on horse]*. Ljubljana: Nedeljski dnevnik.

Štukelj, L. (1989). *Mojih sedem svetovnih tekmovanj* [*My Seven World Competitions]*. Novo mesto: Dolenjska založba.

Ste-Marie, D. M. (1999). Expert-novice differences in gymnastic judging: An information processing perspective. *Applied Cognitive Psychology, 13*, 269-281.

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

Ste-Marie, D. M., Bottamini, G. (2003). Memory influenced biases in gymnastic judging: The influences of surface feature changes. *Applied Cognitive Psychology, 17*, 733-751.

Corresponding author:

Ivan Čuk
Faculty of Sport
Gortanova 22
1000 Ljubljana
Slovenia
Tel.: +386 1 5707700
E-mail: ivan.cuk@fsp.uni-lj.si