RELIABILITY OF REAL TIME JUDGING SYSTEM

Maja Bučar Pajek[1], Warwick Forbes[2], Jernej Pajek[3], Bojan Leskošek[1] , and Ivan Čuk[1]

[1]Faculty of Sport, University of Ljubljana, Slovenia
[2]Australian Institute of Sport, Canberra, Australia
[3]University Medical Center, Ljubljana, Slovenia

Original research article

Abstract

The aim of our research was to analyse the implementation of a Real Time Judging System (RTJS) . In this research, 6 volunteer international level judges evaluated male parallel bars routines from Šalamun's memorial 2009 (World cup series B artistic gymnastics competition). The computer assisted system with a keyboard interface was used to record and display deductions from individual judges in real time. For validity assessment, the mean absolute and rank deviations of judges' execution scores, Kendall's W and ANOVA statistics were calculated. For consistency and reliability assessment, item-total correlations, Cronbach's alpha coefficient, intra-class correlations and Armor's theta were calculated. The overall results in terms of consistency (Cronbach's alpha mostly above 0.96) and reliability (Armor's theta 0.95, intra-class correlation for single and average measures 0.77 and 0.95, respectively) were satisfactory. As compared to results of judging analysis from a previous high level competition at Universiade 2009 higher indices of individual judge bias were found. In conclusion, RTJS shows promise as an efficient system to increase the transparency and informative value of judging while maintaining the same level of reliability.

Keywords: *artistic gymnastics; information technology; panel judging; bias.*

INTRODUCTION

Transparent and precise judging in artistic gymnastics is of paramount importance. Currently there are 6 judges (or 4 judges for competitions at levels lower than Olympic Games or World Cup) evaluating exercise execution. This results in the E (execution) score. In addition, 2 judges evaluate exercise content and they provide the D (difficulty) score (FIG, 2009). 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. Since the D score is a joint (consensus) score of both judges who evaluate exercise content, it is impossible to calculate reliability and validity, while for the E score – which is an average score of the middle four (or two) judges – this calculation is possible.

It must be stressed that currently only the sum of deductions is presented in the individual judge's score and it is not known at what time-points which deduction took place and what was its magnitude. It would be of great value if E score judges could also be evaluated according to what deduction was taken and when it was taken during the rated exercise. The differences between judges of different expertise in this regard do exist (Dallas & Kirialanis, 2010). Computerised systems to allow for such an analysis of judging are available (Čuk & Forbes, 2006) and they should be tested as a

means to reduce significant differences in judge's scores and to improve the overall quality of judging.

This work is aimed to present for the first time the results of judging with the RTJS, which enables to record the deductions of individual judges in real time. The reliability and validity indices as defined previously (Bučar Pajek, Čuk, Pajek, Karacsony & Leskošek, 2011) were examined. The results were compared to recent contemporary results of judging on parallel bars under the same FIG Code of Points regulations from 2009 (Leskošek, Čuk, Karacsony, Pajek & Bučar, 2010).

METHODS

This study was performed at the Faculty of Sport, University of Ljubljana in March, 2011. Six international judges of breve (levels) 1-4 who volunteered to participate in this study were rating the videotaped routines. The routines were chosen from the international competition Šalamun's memorial, which is a world cup competition series B and was held at Maribor, Slovenia in 2009. As for the first study, only routines on men parallel bars were selected for evaluation.

The RTJS was used to serve as an application for entry of judges' deductions, their recording and display. It was developed in Australian Institute of Sport (Warwick Forbes, Colin Mackintosh) and with collaboration of Faculty of Sport, University of Ljubljana (Ivan Čuk) . It enables the entry of judge deductions in real time during the routine execution. It is composed of a special keyboard with 4 keys (Figure 1), a computer, USB manifold, video camera, manifold of video signal, symbol generator, TV, video-recorder and video-player. System is regulated by a special software developed specifically for this application by Colin Mackintosh (Figure 2).

Figure 1. *A special keyboard for the entry of deductions (4 keys only, for deductions of 0,1; 0,3; 0,5 and 1 points, respectively).*

Figure 2. *The outlook of the special software developed specifically for this application (computer screen reporting deductions. Note the timeline on x-axis and the deductions by specific judges on the y-axis)*

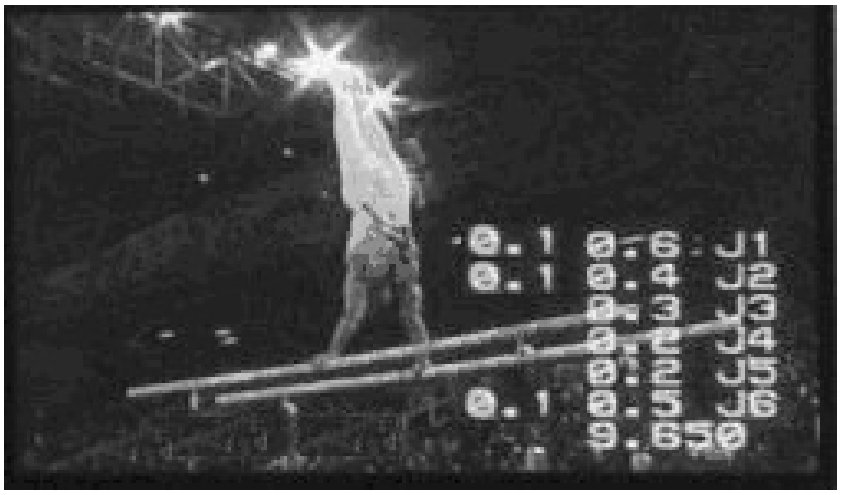

Figure 3. *The video-screen showing routine and the deductions in real-time.*

Main system features and advantages are:

- E judges cannot change deductions during the routine;
- The judges must react to each mistake during the routine with a press on the appropriate key;

- System records the time the deduction was taken and the value of deduction taken (Figure 2);
- System could be compatible with ICROS Longines® system;

Through recording of selected routine and displaying the deductions on the video of that routine at the exact time when deductions were entered it gives the competitors and coaches an invaluable feedback about the judges' evaluation (Figure 3).

After the rating of all the parallel bar routines we calculated descriptive statistics for deductions, item (individual judge) and scale (all judges together) scores. Distributional statistics (mean and standard deviation) were calculated for individual judge's deductions and for signed and absolute deviation from final E score of competitors. These two forms of deviation are measures of bias (under- or over-estimation). For each individual judge, mean rank (R_{mean}) and its deviation (dR_{mean}) from expected (unbiased) rank were also calculated. These measures of systematic deviation of E scores were used to evaluate the validity of judging (specifically, the aspect of validity which refers to the presence and extent of bias). Expected rank was calculated as (m+1)/2, where m is the number of judges (with 6 judges the expected mean rank is always 3,5). The corrected item-total correlation (r_{corr}) was also calculated (the correlation between individual judge's scores and total scores).

Cronbach's alpha coefficient as a measure of internal consistency was used to test for consistency of individual judges. For each judge the C*ronbach's alpha if item deleted* was also calculated. This is the estimated value of alpha if the given judge was removed from the model.

Armor's reliability coefficient, theta (θ), is based on the calculation of the first and largest eigenvalue (λ_1) from the principal component analysis (Armor, 1974). While the ratio of λ_1 and the number of items (in our case judges) may be interpreted as the percent of total variance in the score due to the variation in the principal component, the Armor's θ is interpreted as a measure of reliability; that is how much of the total variance is represented by the between-subject variance. The closer the value is to 1, the lower is the impact of the raters' errors.

ICC coefficients were calculated under one-way random effects model, where judges were conceived as representing a random selection of possible judges, who rated all competitors of interest. The ICC may be thought of as the ratio of variance explained by the independent variable divided by total variance, where total variance is the explained variance plus variance due to the raters plus residual variance (Shrout & Fleiss, 1979). ICC equals 1 only when there is no variance due to raters and no residual variance. There are two types of reliability analysed with ICC: the single measure reliability and the average measures reliability; both were calculated.

We also performed two analyses of between-judges differences: Kendall's coefficient of concordance W and repeated measures ANOVA. Note that in the context of this research, high (statistically significant) values of Kendall's W indicated systematic bias (under- or over-estimation) with at least one of the judges.

All data were analysed with SPSS Statistics v. 17.0 software (SPSS Inc., Chicago, IL, USA) whenever possible, otherwise with Microsoft Excel v. 11.0 (Microsoft Corporation, USA).

RESULTS

There were 28 parallel bars routines evaluated by the experimental 6-judge panel. The median E score was 8,26 (range 5,68 – 9,5) and the mean±SD was 8,07±0,92. Representation of individual deductions of judges is shown in Figure 4.

The descriptive measures of performance of individual judges are shown in the Table 1. It can be seen that there were two judges with highest excursions from the expected rank (judges 2 and 4).

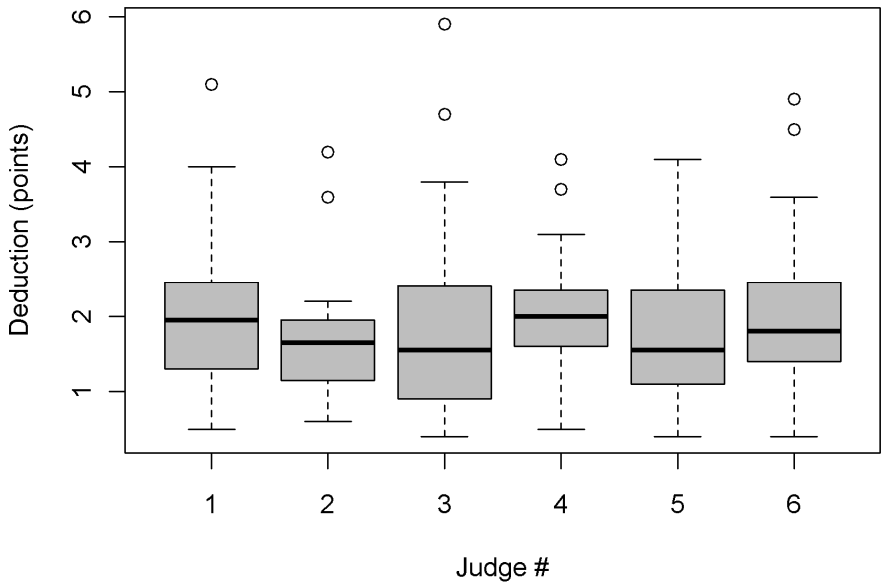

Figure 4. *The box-plots of individual judge deductions.*

Table 1. *The descriptive measures of individual judge performance. Abbreviations: rcorr - corrected item-total correlation; alpha del - value of Cronbach's alpha coefficient if judge deleted; Mean rank - mean rank of judge's E score; devRmean - mean deviation from the expected rank (expected rank always 3,5).*

Judge		Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6
Judge's deductions	Mean	2,00	1,65	1,89	2,14	1,78	2,06
	SD	0,95	0,80	1,32	0,82	0,97	1,05
Deviation from E							
score	Mean	0,07	-0,28	-0,04	0,21	-0,14	0,13
	SD	0,50	0,45	0,49	0,38	0,37	0,35
Absolute deviation	Mean	0,36	0,37	0,36	0,31	0,29	0,29
	SD	0,35	0,37	0,32	0,30	0,27	0,23
rcorr		0,82	0,83	0,95	0,89	0,88	0,91
alpha del		0,95	0,95	0,94	0,95	0,95	0,94
Mean rank		4,04	2,50	2,86	4,45	3,07	4,09
devRmean		1,43	1,57	1,71	1,45	1,32	1,38

Table 2. The correlation matrix for between individual judges' scores. The values shown are Pearson correlation coefficients.

Judge	2	3	4	5	6
1	0,69	0,83	0,80	0,70	0,78
2		0,83	0,72	0,79	0,80
3			0,87	0,89	0,88
4				0,80	0,86
5					0,86

The correlations between individual judges are shown in the Table 2. Overall high correlations were found between any two individual judges.

Chrombch alpha was 0.96. The intraclass correlation for single judge scores was 0,765 and the intra-class correlation for average values was 0,951. Kendall's coefficient of concordance was 0,18 ($p<0,001$). Armor's theta coefficient was 0,96. The F value of ANOVA for between judge differences was 4.37 ($p=0,002$).

DISCUSSION

In this first report of the results of judging using the RTJS we have found overall satisfactory indices of reliability. When we compare the values to the report of judging analysis on the Universiade 2009 (Leskošek et al, 2010) we can see that the values of reliability are quite comparable, see Table 4.

Table 4. *The comparison of this study and judging analysis of universiade 2009 in indices of reliability (objectivity)*

Study	Leskosek et al 2010 (Universiade 2009)	This study
Rcorr (median and range)	0.93 (0.77 – 0.96)	0.88 (0.82-0.95)
Cronbach's alpha (median and range)	0.98; 0.93; 0.96	0.96
ICC (single measures) †	0.91; 0.77; 0.81	0.77
ICC (average measures) †	0.98; 0.93; 0.96	0.95
Armors theta †	0.98; 0.94; 0.97	0.96

† The three values for Universiade 2009 denote are derived from the qualifications, all around finals and apparatus finals sessions, respectively.

When the indices of validity (concerning systematic bias) are regarded, there is a trend towards higher maximal deviations from E score at individual judge level with RTJS. Also, when compared to all around finals and apparatus finals at Universiade 2009 it can be seen, that RTJS yielded higher and statistically significant indices of bias (systematic deviation).

To put the observed differences from Table 5 in proper perspective two notions about the current judging process must be

made. First, currently there is a possibility for judges to correct their E-scores when they inspect the final sum of their deductions at the end of routine. Second, during the competition (going from routine to routine) it is possible for judges to correct their judging according to the final e-scores from previous routines. These two possibilities were prevented in our experiment. The judges had to make their deductions during the routine without the possibility for further corrections once the deduction was perceived by computer system (i.e. after the click on the keyboard). This important difference reduced the regression towards the mean and accentuated the differences between judges.

When we regard the possible use of RTJS in future it is more than obvious that such system would be of great value for the training of judges. It would enable a faster and more efficient inspection the judging output during the routine. Additionally, with this application the feedback to coaches and competitors would be much more informative giving them as much detail as possible about when and where the deductions were taken within their routine. Finally, we believe that this system should be tested also at gymnastics competitions to compare it with the current one and to see if it would enable us to further improve the judging performance in terms of reliability and validity and transparency of judging.

Table 5. *The comparison of this study and judging analysis of universiade 2009 in indices of validity (systematic bias of judges).*

Study	Leskosek et al 2010 (Universiade 2009)	This study
Maximal mean deviation from E score	0.07; 0.08; -0.18	-0.28
Maximal mean absolute deviation from E score†	0.17; 0.19; 0.26	0.37
Kendall's W†	0.03 ($p<0.05$); 0,08; 0.12	0.18 ($p<0.001$)
ANOVA F value for between judge differences (p)†	5.7 ($p<0.05$); 1.87; 0.89	4.37 ($p=0.002$)

† The three values for Universiade 2009 denote are derived from the qualifications, all around finals and apparatus finals sessions, respectively.

To conclude, we have reported for the first time the performance of RTJS. The system was able to record the deductions of individual judges in real time during the execution of routines. The results of the judging panel composed of 6 international level judges evaluating male parallel bars routines were comparable to the highest level competition (Universiade 2009) in the terms of reliability indices. Higher values of bias indices were found RTJS probably as a consequence of reducing the process of self stimulated regression towards the mean. This system shows great promise as a computerised application to provide more transparent, informative and reliable judging performance in the future.

REFERENCES

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

Bučar Pajek, M., Čuk, I., Karacsony, I., Pajek, J. & Leskošek, B. (2010). Reliability and validity of judging in Women' artistic gymnastics at the 2009 University games. *European Journal of Sport Science.* Manuscript accepted for publication at 14. december, 2010.

Čuk, I. & Forbes, W. (2006). Kam greš, sojenje? [Where is judging headed?]. In Kolar, E. & Piletič, S (Eds.), *Gimnastika za trenerje in pedagoge 2* (pp. 76-86). Ljubljana: Gimnastična zveza Slovenije.

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

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

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*, 25-34.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-427.

Coresponding author:

Maja Bučar Pajek
Faculty of Sport
University of Ljubljana
Gortanova 22
1000 Ljubljana
Slovenia
E-mail: maja.bucar@fsp.uni-lj.si