USE OF OBJECTIVE METHODS TO DETERMINE THE HOLDING TIME OF HOLD ELEMENTS ON STILL RINGS

Thomas Lehmann[1], Alexander Winter[2], Alexander Seemann-Sinn[1], Falk Naundorf[1]

[1] Institute for Applied Training Science, Department of Strength and Technique
[2] singularIT GmbH, Leipzig, Germany

Original article, DOI:10.52165/sgj.13.2.181-189

Abstract
The duration of holding elements represents a critical factor for judging routines on the still rings in artistic gymnastics. Athletes can be penalized with non-recognition of an element if the hold time is too short. Dynamometric and kinematic measuring methods offer the possibility to provide support to judges in evaluating the duration of the hold time. In this study a dynamometric method with two different variants (dms10 and dms5) as well as a kinematic method (kms) based on a trained neural network were presented and examined with regard to their agreement with judges' evaluations when determining the hold time. To check the agreement, a) the percentage agreement and b) the interrater reliability were calculated using Cohen's kappa (k). The two dynamometric methods showed a percentage agreement of 83.5% (dms10) and 51.7% (dms5) with the hold time evaluation by judges. The percentage agreement of the kms was 38.8%. The interrater reliability showed for the dms10 a moderate (k = 0.58) and for the dms5 a fair (k = 0.23) agreement, while the kms showed a poor (k = 0.02) match. The results supported dms10 for its possible use as a practicable and reliable method to assist judges in evaluating hold times on the still rings. Dms5 and kms (in the current development stage) were not suitable as means of judges' support.

Keywords: *men's artistic gymnastics, still rings, judging, hold time, measurement systems*.

INTRODUCTION

Competition results in artistic gymnastics are determined by judges who evaluate routines according to clearly defined rules. Despite there being objective rules on how to evaluate a gymnastics routine, such evaluations are prone to mistakes (Leskošek, Čuk, Karácsony, Pajek, & Bučar, 2010). Making use of technical measuring systems could help to minimize these errors and thus make these evaluations more objective. In order to increase the use of measuring systems in gymnastics, researchers have developed various solutions using engineering and computer science (Fujihara, Gervais, & Irwin, 2019, Omorczyk, Nosiadek, Ambroży, & Nosiadek, 2015; Veličković, Paunović, Madić, Vukašinović, & Kolar, 2016). Some of these ideas are already being utilized in other competitions. For example, in trampoline gymnastics, time of flight (ToF) is a scored part of the competition and can be accurately measured by using a laser-based light curtain. Since the 2017-2020 Olympic cycle, determination of the take-off and take-on position of a jump (horizontal displacement) has become important. As a result, a measuring system that can determine both ToF and position is now in use (Ferger & Hackbarth, 2017). The impact scores of different horizontal displacement calculations on competition has also been investigated (Ferger, Helm, & Zentgraf, 2020).

In October 2017, the Fédération Internationale de Gymnastique (FIG) announced that it would partner with Fujitsu to develop a practical judging support system for gymnastics competitions. By using a 3-D-laser-sensor technique and artificial intelligence, the movements of gymnasts were recorded and automatically analyzed to detect elements and recognize the difficulty score (D-score) (Fujitsu Limited, 2018, Fujitsu Limited, 2019). At the Artistic Gymnastics World Championships 2019, the judging was supported by this system.

Since 2014, a dynamometric measuring system has been used at European Gymnastics Federation competitions in order to determine the hold times of strength elements on still rings (Aarts & Pluk, 2016a). The objective of this system is to help judges to more objectively estimate hold times. In this system, still rings are equipped with specially prepared force elements allowing the vertical forces on the ring cables to be measured.

The hold times are calculated in a combination of displaying objective data together with subjective data. The exact start and end time of a hold is given by a Hold Time judge. A graphical user interface assists the judging in evaluating hold times in still rings routines.

Consequently, a high level of attention must therefore be devoted to hold times during the training of athletes in order to obtain good competition scores. In summary, it can be stated that the use of objective measurement methods to support judges is gaining ground.

According to current FIG regulations, routines on still rings have to be composed from elements of 1) kip and swing elements & swings through or to handstand; 2) strength elements and hold elements; 3) swings to strength hold elements, and 4) dismounts, to receive full evaluation of the element groups (FIG, 2017). Most of these elements end in a holding position. A count of all elements on still rings that end in a holding position from the current Code of Points (CdP) reveals they make 67% of the total elements that can be realized on this apparatus (FIG, 2017). Furthermore, these elements are the only ones with a difficulty greater than or equal to a D-score of 0.5 points, except for the "O'Neil" element and the six dismount elements (FIG, 2017).

A look at competition data from the World Championship still ring finals between 2017 and 2019 reveals 35.3% of the element part in the D-score comes from group 2 (strength elements and hold elements), and 27.0% from group 3 (swings to strength hold elements). This shows how important these elements are for a successful routine. With this in mind, it’s clearly essential to have a means of accurately determining hold times on still rings. Aarts and Pluk (2016b) confirm the importance of correct hold time determination due to the marginal differences exhibited between final scores in the 2014 and 2015 European Championships ring final. When evaluating elements, judges are required to assess the duration of the holding position, which must be held for at least two seconds for full acceptance. If a holding position is held for less than two seconds, the gymnast loses 0.3 points for this element. If there is no visible stoppage, the element will not be recognized by the difficulty judges and 0.5 points will be deducted by the execution judges (FIG, 2017). But there is no exact definition in the CdP how long a visible stop has to be. By using objective methods, a more accurate system of deductions (small, medium or large) can be utilized. However, before using technical systems in competition and training, potential measuring systems need to be evaluated in order to confirm their accuracy. There is no “gold standard” for the evaluation of hold time; for this reason the evaluation should be conducted by qualified judges (keeping in mind that judges do not always agree with each other).

The aim of this study was to investigate two measuring systems which have been developed to evaluate hold time. The two systems to be assessed are a dynamometric method and a kinematic method (which uses a trained neural network to estimate hold times in video frames). The quality of these measuring systems will be examined and compared to the corresponding decisions made by judges.

METHODS

For the data analysis, 14 still ring routines from the German Championships all-around final 2018 were used. The routines were performed by eleven German national team squad gymnasts and three non-squad gymnasts. Their age, height and body mass were 25.6 ± 2.9 years, 170.0 ± 0.062 m and 65.2 ± 4.9 kg. The body masses of the athletes were measured using the dynamometric measurement system over 0.5 seconds while in the still hanging starting position with vertical arms and legs at the beginning of the routine. The still rings routines had an average difficulty (D-score) of 5.10 ± 0.75 points. For the analysis of the hold time, 85 hold elements (also including handstands) were performed by 14 gymnasts in their routines.

Dynamometric measurement system

A dynamometric measurement system (dms), based on prepared force elements and a synchronized video camera (Basler BIP 640c, 50 Hz frame rate, 640 x 480 px), was utilized to determine hold time using dynamometry. As a compromise for the different hold elements (cross – front view; planche – side view), the camera position was about 45° (Fig. 2). Force sensors were placed in the screw connection of each cable of the rings. They were directly connected to the suspension of a FIG-certified ring apparatus by SPIETH (Fig. 1).

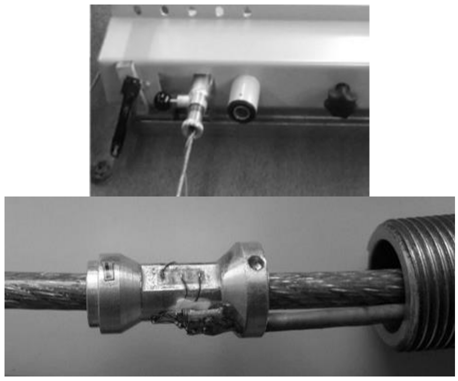

Figure 1. Position of the force sensors.

The prototype sensors consist of a bending body with strain gauges (350 Ohm). The vertical forces exerted by the gymnast were recorded at a sampling frequency of 200 Hz simultaneously with a 50 Hz reference video and stored in a measuring computer. With the help of computer software, the sum signal of the vertical forces of both left and right rings was analyzed relative to the body weight (bw) of the athlete. When conducting an automatic detection of the holding elements and the hold times, forces were detected between a wide range of a) 0.9 - 1.1 bw (±10 %; dms10) and a small range b) 0.95 - 1.05 bw (±5 %; dms5) (Fig. 2). It must be noted that due to the geometric relationship between the length of the ring cables and the gymnast’s arm spans, more force of each cable is measured for the element cross. Gymnasts with arm spans of 1.45 m achieve 1.2% and athletes with arm spans of 1.70 m even 1.9% more force in the cables. This error range is considerably lower than our two defined ranges (±10 %; ±5 %).

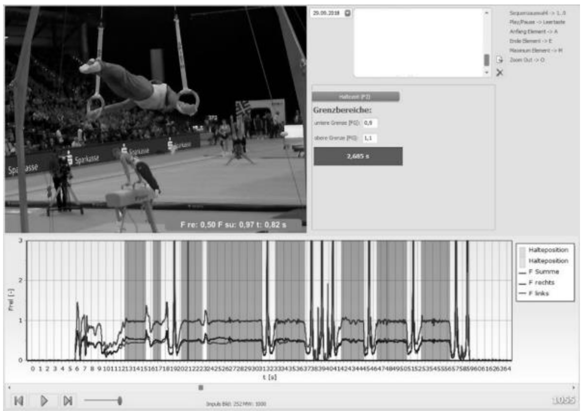

Figure 2. Example of automatic determination of the hold time (marked parts of the force-time curve) and synchronized video.

A kinematic measurement system (kms) based on OpenPose and self-programmed software was the second method used for determining hold time. Open source software OpenPose (https://github.com/CMU-Perceptual-Computing-Lab/openpose) is a trained neural network for human pose estimation in images and videos. It returns 18 body points with x and y coordinates (Cao, Hidalgo, Simon, Wei, & Sheikh, 2018). In order to avoid false negative detections, the body points below the knees (10 and 13) and from the head (14 - 17) were not considered in the evaluation of the hold times. These points were less relevant for the evaluation, plus they were more prone to errors in detection than other body points (Winter, 2019).

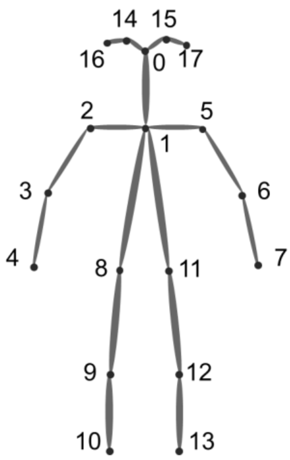

Figure 3. OpenPose body points.

The videos recorded using the dynamometric measurement system (50 Hz, Basler Bip640c), were also used for these data analyses. The x- and y-coordinates of the body points (Fig. 3) were evaluated by a specially developed software program. This program compares the x- and y-coordinates of successive frames. A holding position between two (or more) frames was detected if the difference between the x and y coordinates forming 60% of the body points was less than a defined margin of 1.5% (Winter, 2019). For a video size of 640 x 480 px the change would be a maximum of 9.6 x 7.2 px. The hold time results from counting successive frames where a holding position is detected (Winter, 2019). Using this information, the program could calculate the total hold time of the element.

For an expert comparison of these three methods, five judges (FIG brevet level; four of them category 2) evaluated the routines. They watched each routine at its original video speed and decided if the time for holding positions was accepted without deduction, with 0.3 points deduction or not accepted at all. Only hold times were the focus; holding positions were not evaluated. To compare the judges' evaluations with the other measuring methods, the majority result by the five judges was used.

In this study, all measured hold times (dms10, dms5 and kms) were classified into three judging categories. Because there is no definition of a visible stop for element recognition in the CdP, we define it as being a minimum of 0.5 seconds. Detected hold times from 0.50 to 1.99s were categorized with 0.3 points deduction and hold times ≥ 2.0s without deduction. A comparison was made between the judges' evaluations and the three measurement methods (dms10, dms5, kms). Furthermore, judges' individual evaluations were compared to their majority evaluations. In order to check both the agreement of a) the three measuring methods with the majority evaluation of the judges, and b) the agreement of the individual judge's evaluations with their majority evaluations, percentage agreements were calculated. In addition, the interrater reliability using Cohen kappa (k-value) for both a) and b) was determined to analyze the agreement. The levels for k-value were rated as either 'poor' (<0.00), 'slight' (0.00 - 0.20), 'fair' (0.21 - 0.40), 'moderate' (0.41 - 0.60), 'substantial' (0.61 - 0.80) or 'almost perfect' (0.81 - 1.00) (Landis & Koch, 1977).

RESULTS

Analysis of the various methods showed the judges on the whole and the dms10 made no deductions for hold time in the majority of the elements they evaluated. When assessing hold time using the kms (45.9%) and dms5 (51.9 %), the majority of the elements were given a deduction of 0.3 points (Table 1).

Examining the agreement between judges reveals that all five judges (100%) estimated the hold time as equal for 33 elements (Table 2). This represents 39% of the elements. Four out of five judges (80%) agreed on the hold time for an additional 31 elements (36%).

Table 1
Distribution of element evaluations across the various measurement systems.

	judges		dms10		dms5		kms	
Total number of Elements	85							
not accepted	4	4.7 %	2	2.4 %	14	16.5 %	13	15.3 %
0.3 points deduction	25	29.4 %	17	20.0 %	39	45.9 %	44	51.8 %
without deduction	56	65.9 %	66	77.6 %	32	37.6 %	28	32.9 %

Table 2
Percentage agreement between the judges.

judge assignment	elements	percentage value
100% equality (all 5 jugdes`)	33	39%
80% equality (4 to 1 jugdes`)	31	36%
60% equality (3 to 2 or 3 to 1 to 1)	21	25%

Table 3
Percentage agreement and the Cohen kappa (k-value) for the single judges with the majority.

	judge 1	judge 2	judge 3	judge 4	judge 5
total agreement with judges' majority[elements]	51	56	44	54	25
total percentage agreement [%]	60.0	65.9	51.8	63.5	29.4
k-value	0.05 'slight'	0.14 'slight'	-0.05 'poor'	0.12 'slight'	-0.11 'poor'

Table 4
Percentage agreement of the dynamometric and kinematic measurement systems with the judges' majority evaluations in determining hold times.

	dms10	dms5	kms
total agreement with judge majority [elements]	71	45	33
total percentage agreement [%]	83.5	52.9	38.8
fully recognized [%]	87.9	90.6	82.1
0.3 reduction [%]	64.7	28.2	22.7
non-recognized [%]	100.0	28.6	0

Table 5
Interrater reliability between the dynamometric and kinematic measurement systems, and the judges' majority evaluations in determining hold times.

	dms10	dms5	kms
k-value	0.58 'moderate'	0.24 'fair'	0.02 'slight'

The calculation of k-value for how an individual judge's evaluation compares to the majority evaluation of the judges resulted in values of -0.11 to 0.14. The agreement between the individual judge's evaluation and the majority evaluation is therefore rated as 'poor' to 'slight' with regards to k-value (Table 3).

The analysis showed that 71 of the total 85 elements detected using dms10 were in agreement with the majority judges' evaluations. In ten of the fourteen elements without agreement the dms10 showed longer hold times compared to the judges' evaluations. For example the judges evaluated eight elements with 0.3 points deduction while the dms10 evaluated these elements without deduction. When using dms5, 44 of 85 hold positions were shown to agree with the judges. The percentage agreements of the two dynamometric variants were therefore 83.5% (dms10) and 51.7% (dms5) when compared to the judges' evaluations. A total of 33 of the 85 elements recorded with the kms were in agreement with the judges. This corresponds to a percentage agreement of 38.8%. Dividing the elements into individual categories (fully recognized, 0.3 points deduction and not recognized) the measurement systems showed a high percentage of agreement with the judges' evaluations for the fully recognized elements (Table 4). For the non-recognized elements, dms5 and kms had a low percentage agreement with the judges' evaluations. In contrast, dms10 was 100% identical to the judges' evaluations in the category "non-recognized elements".

For the 33 elements which all 5 judges evaluated equally, a similar percentage agreement for the measuring systems is shown. The total percentage agreement for dms10 is 78.8%, for dms5 60.6% and for kms 24.2%.

The determination of k-value showed the level of agreement with the majority of the judges' hold time evaluations: it was moderate for dms10 (k = 0.58) and fair for dms5 (k = 0.23) (Table 5). For the kms (k = 0.02), there was a slight agreement with the judges' evaluations.

DISCUSION

The use of objective, scientifically evaluated and practicable methods to assess the quality of gymnastics elements is of interest to athletes, coaches and scientists alike. The international gymnastics federations UEG and FIG as well as local organizing committees and media and spectators are also interested in a fair competition for all gymnasts.

Therefore, a dynamometric and a kinematic method were presented in this study as a way of providing support to judges when it comes to evaluating still ring hold times. It investigated how well the dynamometric and kinematic systems agreed with the judges' evaluations of hold times. The results of the study show that for the dynamometric measuring system (dms10), 71 of the 85 hold time evaluations matched those of the judges. However, since the percentage of agreement did not take random agreement into account, Cohen's Kappa was additionally calculated to indicate the agreement of the methods (McHugh, 2012). Based on the calculated k-value (0.58), the result is a 'moderate' agreement between the two measuring methods. By comparison, if one compares the hold time evaluations of individual judges to their majority evaluations, the result shows only a 54.1% mean percentage agreement and a 'low' to 'slight' k-value. The dynamometric system with a 10.0% range of body weight (dms10) was therefore shown to have the highest concordance with the majority of the judges - higher than individual judges themselves.

These results contrast with the experiences of Aarts and Pluk (2016b) who concluded that a fully automatic evaluation of hold time based on force measurements would not be accurate enough. A possible explanation for these different results could be the range of force values that were accepted as hold times, or the comparison with the judges' observations instead of the comparison with the Smart Rings system made by Aarts and Pluk (2016b). The dms5 method, which had a smaller range of force values (0.95 - 1.05 times of the body weight), resulted in a fair agreement with the judges' evaluations. These results support the view of Aarts and Pluk (2016b). Nevertheless, even with the dms5 method, a greater agreement with the judges' majority evaluation is observed when compared to the evaluations of individual judges. A weakness in the dms method must be explained. A constant force in the defined bw can be determined, if the gymnast performs at a constant velocity through the holding position. At constant velocity the acceleration is zero and thus the force is equal to the body mass (cf. $F=m\cdot a$). This weakness in the dms method could be the reason why dms10 evaluates fourteen elements without agreement with the judges, in most cases with a longer hold time. The kms method, which uses human pose estimation through trained neural networks to detect human key points in images and takes an algorithmic approach for calculating hold times, resulted only in a 38.8% agreement with the judges' evaluations. The k-value of 0.02 confirms that the agreements between the kms method and the judges' majority evaluations were only 'slight'. One reason for this poor correlation can be found in the deficiency of the trained neural network. Since the algorithm is highly dependent on the quality of the neural network output, improving the network would increase the quality of the method itself. Such an optimization could be realized through specific training of the neural network on sports related images or an optimized camera perspective of the input videos (Winter, 2019). Gymnastics elements on rings, especially positions like head down/ feet up, were not yet used for training the neural network. Another reason for the 'slight' agreement was that the kms method is based on only 50 Hz video with a resolution of 640 x 480 px. In further studies, video material in higher resolution and/or higher frame rate should be used.

Due to the high significance of hold times in competitions, it's important they are accurately examined during training.

Athletes can therefore be given important information about their performance level in the strength holding elements. For example, a short underrun of the hold time (< 0.5s) would probably be caused by an insufficiently pronounced sense of time, while large underruns (> 0.5s) are likely due to weaknesses in strength levels. There were many possibilities to determine the duration of holding elements with video methods in training. Unfortunately, these methods did not work automatically, making it difficult for coaches to use them as a stand-alone system. Therefore, to effectively check the hold time during a training session, a semi-automatic system should be used whereby both the coach and the athlete receive direct feedback on how long the corresponding elements have been held. The dms10 method meets this requirement and is therefore suited for training. For competition requirements however, the demonstrated 16.5% disagreement with the judges' majority evaluations means this method is considered unacceptable.

CONCLUSION

Reliable measurement systems can help to minimize errors made by judges when evaluating a gymnastics routine. The correct determination of hold time on still rings is one area that is prone to judges' errors . Technical measurement systems can also be used as immediate feedback systems in training. In this study, two measuring systems with the capability to determine hold time (two dynamometric system variants and a kinematic system) were examined by comparing them with judges' evaluations. The results support the use of the dynamometric method (dms10) as a reliable measurement system for the determination of hold times in training sessions. Due to the automated evaluation of hold time, this method was suitable as a feedback system in training. At this moment, the approach via a kinemetric method using a trained neuronal network for human pose estimation does not represent a satisfactory solution yet. However, special training of the neural network may possibly optimize this method and potentially make it an acceptable system too. The integration of measurement systems is an important approach for improving objective evaluations in gymnastics and can additionally improve the quality of information in this sport.

ACKNOWLEDGMENTS

This research was funded by the Federal Ministry of the Interior, Building and Community and was supported by a decision of the German Bundestag. We would like to thank the German Gymnastics Federation and SPIETH Gymnastics for their support in preparing and carrying out our measurements.

REFERENCES

Aarts, M., & Pluk, A. (2016a). *Hawkeye" for gymnastics: Automatic evaluation of hold time in a still rings exercise during the European Championships gymnastics 2014*. Paper presented at the Science and Engineering Conference on Sports Innovation, Amsterdam.

Aarts, M., & Pluk, A. (2016b). *Making Every Second Count: Implementation of the Smart Rings Jury Tool to Evaluate Hold Time during the Men's Artistic Gymnastics European Championships 2016*. Paper presented at the 4th International Congress on Sport Sciences Research and Technology Support, Porto.

Aarts, M., & Pluk, A. (2017). *Implementation of the Smart Rings Jury Tool during the Men's Artistic Gymnastics European Championship 2016*. Paper presented at the Science and Engineering Converence of Sports Innovation, TU Delft.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi:arXiv:1812.08008

Ferger, K., & Hackbarth, M. (2017). New way of determining horizontal displacement in competitive trampolin. *Science of Gymnastics Journal*, *9*(3), 303-310.

Ferger, K., Helm, F., & Zentgraf, K. (2020). Estimating horizontal displacement deduction in trampoline gymnastics by means of constant and variable errors of landing positions: A new gold standard? *Science of Gymnastics Journal*, *12*(2), 203-216.

FIG. (2017). *2017 Code of Points - Men`s Artistics Gymnastics*. Lausanne: FIG.

Fujihara, T., Gervais, P., & Irwin, G. (2019). Head-toe distance as a simple measure to evaluate amplitude of circles on pommel horse. *Science of Gymnastics Journal*, *11*(2), 189-200.

Fujitsu Limited. (2018). The International Gymnastics Federation to Implement Fujitsu's Judging Support System. Retrieved from https://www.fujitsu.com/global/about/resources/news/press-releases/2018/1120-01.html

Fujitsu Limited. (2019). Changing the world of sports with judging support system. Retrieved from https://www.fujitsu.com/global/about/resources/case-studies/vision/gymnastics-sport/

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, *33*(1), 159-174.

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.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, *22*(3), 276-282.

Omorczyk, J., Nosiadek, L., Ambroży, T., & Nosiadek, A. (2015). High-frequency video capture and a computer program with frame-by-frame angle determination functionality as tools that support judging in artistic gymnastics. *Acta of bioengineering and biomechanics*, *17*(3), 85-93.

Veličković, S., Paunović, M., Madić, D., Vukašinović, V., & Kolar, E. (2016). Proposed method of identification of technical errors in artistic gymnastic: Case study. *Science of Gymnastics Journal*, *8*(1), 43-56.

Winter, A. (2019). *Applicability of machine learning methods for determining the human position on the basis of video data, for example gymnastics on the apparatus Rings*. BA Thesis [unpublished]. University Leipzig. Leipzig.

Corresponding author:

Dr. Thomas Lehmann
Institute for Applied Training Science
Department of Strength and Technique
Marschnerstr. 29
04109 Leipzig
Phone: +49-341-4945339
Email: thlehmann@iat.uni-leipzig.de

Article received: 14.8.2020
Article accepted: 19.12.2020