Evaluating race walking judges

© by IAAF
23:4; 43-49, 2008

By Nicola Maggio

ABSTRACT

One of the first questions asked after a race walking event is usually: was it well judged? As there are no absolute or certain criteria for evaluating either the work of individual judges or the race itself, this question has led to subjective assessments and sometimes to controversy. In this report the author, a long-time member of the IAAF Race Walking Judges Panel, describes a system that is being developed in Italy to assess the work of race walking judges using objective data. The system is based on the analysis of the numbers of disqualified athletes, Red Cards and cautions in a race. This data provides a set of indices leading to a Technical Assessment Index (TAI) for individual judges and a Consistency Ratio (CR) for the work of the whole judging team. The end product of the process is to indicate, through an overall Race Value, how close the judging comes to what may be considered an ideal race. Based on data collected from more than 200 races, assessments are given for a sample of Italian races and for the top international competitions since 2000. The author concludes by pointing out the limitations of the system and indicating that the development process will continue.

AUTHOR

Nicola Maggio is an Italian race walking judge who has worked at the national level since 1975 and has been a member of the IAAF Race Walking Judges Panel since 1990. From 1960 to 1972 he was a competitive race walker.

Introduction

As in any sport subject to human evaluation, one of the first questions asked after a race walking event is: was it well judged? As there are no absolute or certain criteria for evaluating either the work of individual judges or the race itself, this question has naturally led to subjective assessments, lack of agreement and, sometimes, damage to the image of the sport. However, careful examination of available statistics allows us to analyse race walking competitions more objectively and use the insight gained to improve the work of judges in the future.

Since 2004, the Race Walking Judges Commission of the Italian athletics federation (Federazione Italiana di Atletica Leggera - FIDAL) has developed and implemented a system for making such an analysis based on a design suggested by the International Association of Athletics Federations (IAAF). We have collected data from more that 250 national-level races with the aim of marking a point of departure and confrontation with this thorny issue. Those of us working on the project do not claim to have found an absolute truth, but we feel our system has merit and we will continue to work on it in the future.

To be fair here, we must point out that similar systems have been developed in

other countries (we are aware of work in this area in both Spain and the Czech Republic), but we do not know whether the data collection has been continuous over time, as is the case for our work, or if these systems are limited to single races and are ends in themselves.

As our system is by no means at a final stage, the aims of this article are to provide a kind of progress report by describing what we have developed and the thinking behind it, to discuss how the system applies to international events and to identify weaknesses to be addressed in the future.

Data analysed and the resulting indices

It is important to remind readers that according to IAAF Rule 230, when a race walking judge observes an athlete exhibiting visible loss of contact with the ground or a bent knee during the period of the stride from the first contact with ground until the vertical upright position, he/she shall send a Red Card to the Chief Judge. When Red Cards from three different judges have been sent, the athlete in question is disqualified from the race. Judges also have the option of giving an athlete one caution if the athlete seems to be in danger infringing the rule, i.e. of losing contact or bending his/her knee. Each caution must be reported to the Chief Judge at the end of the competition.

This system provides us with some objective, easy to measure, data that can be analysed in the form of coefficients. These are:

1. The number of Red Cards sent by the individual judges who contributed to the disqualification of a particular athlete
2. The total number of Red Cards coming from each individual judge
3. The number of cautions imposed by each individual judge

With this data we can calculate the following indices:

Index A: the value percentage between the value of 1 above and the total number of athletes disqualified in the race. This index (expressed as a number from 0 to 100) indicates the homogeneity of the work of the individual judge with his/her judging colleagues. It is easy to see that the higher the number of Red Cards sent by the single judge, the greater the possibility that the value of A will be high. We can point out here that the primary task of a judge is to "judge," not to raise his/her index evaluation.

As a corrective to the above we can also take into consideration the following:

Index B: the value percentage between the values of 1 and 2 above. This index (also expressed as a number from 0 to 100) has the specific purpose of mitigating the effect of a large number of Red Cards issued by a single judge compared to his/her colleagues.

Index C: the value percentage between the values of 1 and 3 above. This index (also expressed as a number from 0 to 100) indicates how many of an individual judge's cautions were followed by Red Cards to the same athlete. It also shows, if seen from the viewpoint of the judge, whether he/she has been effective with his/her cautions and, if seen from the viewpoint of the athletes, how well they took account of the message received, in the form of the caution, from the judge in question. We need to remember that the proof of a good judge does not lie in how many competitors he/she warns, but in how many he/she can help finish the race without breaking the rules through the proper use of cautions.

Technical Assessment Index

The sum of the values of Index A and Index B (thus having a theoretical value between 0 and 200) is called the Technical Assessment Index or TAI. In effect, the TAI is the translation of the work of an individual judge in a particular race into a single number.

A first evaluation of an individual judge's work therefore is offered by the stratification of the TAI value he/she achieves. This value can then be weighted with other parameters and or objective values related to the race, such as:

- The theoretical difficulty of the race
- The number of athletes at the start
- The conduct of the race on the road or track

Measuring "Consistency"

By the term consistency we want to indicate the degree of uniformity of the work of a team of judges with respect to a particular race. It is obvious that the work of a team in which some of the individual judges have some very high TAI values and others have TAI values that are very low will be completely different from that of a team whose members all have high values, and even more different from one whose members all have very low values.

We recognise that there may be several possible technical or mathematical definitions of consistency. Many seminars on race walking have been about trying to give a solution to how this concept can measured. Some proposals go so far as to consider the type of error ("loss of contact" or "bent knee") as an element of measurement. For our part, we found that parameters that are too sophisticated do not help, as they can be difficult to understand even for trained personnel.

Our aim was to create a definition that was both suitable and very simple. In the end we arrived at the following ratio:

$$\text{Consistency} = \frac{\text{Total number of DQed athletes} \times 3}{\text{Total of Red Cards coming from all Judges}}$$

Our thinking is based the fact that three Red Cards from three different judges are needed for the disqualification of an athlete in a race walking competition. Consequently the value given by the formula above, which we call the Consistency Ratio or CR, may simply be considered a numerical value indicating how the behaviour of a homogenous team of judges, as a whole, compares within a particular race.

The value that the CR gives may vary from a maximum of 1, where all the Judges have only given Red Cards to athletes who are eventually disqualified, to a minimum of 0, where all the judges have given Red Cards to different athletes and no athlete collects the three required for disqualification.

Interpretation and evaluation

The question that arises at this point is: how can we use the values of the variables identified above to evaluate race walking judges, particularly those working at the international level?

Work in this area was carried out with the help of a graduate of the faculty of Motor Sciences at the University of Milan, Francesco Casanova, a student of Prof. Antonio La Torre, who, in a specific thesis analysed the data from the most important international race walking competitions between 1990 and 2004. We accepted the conclusions he reached as benchmarks for comparison of the values being examined and these are given below. They are now applied to the races that normally take place in Italy. However, the accepted benchmarks do not preclude new calibrations for comparison in the future, if analysis identifies changes to the averages of the indices over time.

Technical Assessment Index

As stated above, the TAI has a theoretical value that varies between a minimum of 0 and a maximum of 200. Based on the work of Mr. Casanova, the following classifications were agreed:

1. **Very good (Excellent)** is considered a value exceeding 160
2. **Good** is considered a value between 125 and 159
3. **Normal** is considered a value between 90 and 124
4. **Poor (Inadequate)** is considered a value of less than 90

Ratio of Red Cards to cautions

If the goal of the judge is to help athletes finish the race without breaking the rules, it is clear that the lower the ratio of Red Cards to cautions (Index C) the more the athletes have benefited from the "advice" received from the judge in the form of cautions. The following classifications were agreed:

1. **Very good (Excellent)** is considered a value less than 40
2. **Good** is considered a value between 55 and 40
3. **Normal** is considered a value between 99 and 56
4. **Poor (Inadequate)** is considered a value exceeding 100

Values above 100 indicate that the judge in question found it necessary to directly impose Red Cards (since he/she was absolutely sure of non-compliance of the athlete's technique, as called for in IAAF Rule 230.5) without the prior use of the instrument of a caution (a possibility provided in IAAF Rule 230.4). In these cases, the athlete finds out about his/her technical situation only by looking at the Posting Board(s), where the number of Red Cards given to each athlete is shown, and has a reduced chance to modify his/her technique during the race.

Consistency Ratio

Following the definition of consistency given above, the CR, we can say that our

Technical assessment of judges & events										
Judges	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**Athletes DQ**
50M	0	0	0	0	0	0	0	0	0	0
20W	2	2	1	4	4	5	4	5	27	2
20M	6	8	7	6	9	13	8	8	65	8
10 Jun. M	0	0	0	0	0	0	0	0	0	0
5 Jun. W	0	0	0	0	0	0	0	0	0	0
Total Red Cards	**8**	**10**	**8**	**10**	**13**	**18**	**12**	**13**	**92**	**10**
50M	0	0	0	0	0	0	0	0	0	0
20W	3	6	5	8	12	8	8	3	53	
20M	7	8	9	17	18	10	15	2	86	
10 Jun. M	0	0	0	0	0	0	0	0	0	
5 Jun. W	0	0	0	0	0	0	0	0	0	
Total Cautions	**10**	**14**	**4**	**25**	**30**	**18**	**23**	**5**	**139**	
50M (DQ's)	0	0	0	0	0	0	0	0	0	
20W (DQ's)	1	2	1	1	2	1	0	2		
20M (DQ's)	3	6	5	4	5	4	1	4		
10 Jun. M A index	0	0	0	0	0	0	0	0		
5 Jun. W →	0	0	0	0	0	0	0	0		
Total Red Cards DQ's	**4**	**8**	**6**	**5**	**7**	**5**	**1**	**6**	→	B index
A = (Red Cards DQ's/Athletes DQ)	40%	80%	60%	50%	70%	50%	10%	60%		
B = (Red Cards DQ's/Total Red Cards)	50%	80%	75%	50%	54%	28%	8%	46%		
Technical Assessment Index (A + B)	**90**	**160**	**135**	**100**	**124**	**78**	**18**	**106**		
Evaluation of judges team										
Average of Judge's points									**101**	**Normal**
Standard deviation of Judge's points	C index →								**46**	**Sufficient**
Consistency Ratio									**0.33**	**Normal**
% Red Cards/cautions	80%	71%	57%	40%	43%	100%	52%	260%	**66%**	**Inadequate**
	Inadequate	Inadequate	Normal	Excellent	Good	Inadequate	Good	Inadequate		

Figure 1: Evaluation of individual judges and the judges team at the Italian championships

observations of the past years and comparison with the study mentioned above, leads us to classify the values of the CR as follows:

1. **Very good (Excellent)** is considered a value greater than 0.60
2. **Good** is considered a value between 0.45 and 0.60
3. **Normal** is considered a value between 0.275 and 0.40
4. **Poor (Inadequate)** is considered a value of less than 0.275

Figure 1 shows the relevant statistics for the eight judges (columns 1-8) in five Italian championship races. These are the number of Red Cards and cautions sent by each judge and the totals, the number of athletes disqualified in each competition, and the number of Red Cards sent by each judge who determined the disqualification of an athlete. The bottom line of the main box shows the TAI for each judge and in the lower box we can see the evaluation of the judges team, including the CR.

Race Value

An overall evaluation of the judging of a race can be expressed as the product of the average of TAI values of the individual judges and the CR of the judging team. This is called the Race Value. The greater the figure for the Race Value, the better the race was judged.

It has been found, fortunately not too frequently, that a judging team's CR can be high but the average of the members' TAI values is very low, which leads to the logical conclusion that the team was not up to judging the race or somehow lost control. On the other hand, there are cases where the impact of a race on the outside world may not have been all that positive and the subjective evaluation by uninformed observers was that the judging was not good, despite the fact that both the CR and the average of the TAI values are high, indicating the team's mastery of the race. This is the argument for an indicators like the Race Values that combines both parameters and provides a strong, objective counter-argument in such cases as mentioned here.

Table 1 is an example of the Race Values given for races in Italy during the months of May and June 2008. Keep in mind that the database contains more than 200 races.

The ideal race

As the end product of this process is to indicate what may be considered the parameters of a hypothetical ideal race from a judging point of view, we have suggested the following values as a target for a team of judges:

1. Consistency Ratio greater than 0.60
2. Technical Assessment Index value average for all judging team members exceeding 160
3. Ratio of Red Cards / cautions less than 40

When a team of judges achieves these values we can make the following statements:

1. The team was highly homogeneous in its view of the race.
2. The members of the team were highly qualified individually.
3. The team as whole tried to use prevention, rather than repression, to ensure the athletes used legal technique.

In terms of the athletes, any disqualification imposed by a judging team achieving the target values can be said to represent:

1. An issue related to the actual need.
2. A decision carefully considered, even if in some cases it was taken quickly.
3. The protection of athletes with good technique over those with irregular technique.

Table 1: Evaluations of the judging of ace walking competitions in Italy during May and June 2008 (CR=Consistency Ratio, TAI=Technical Assessment Index)

	Date	Level	Road/ Track	Place	Race	CR	TAI	Race Value
	01.05.2008	IAAF	Road	Sesto San Giovanni	IAAF Race Walking Challenge	0.49	180	88
	01.05.2008	Normal	Track	Alessandria	Youth Meeting	0.26	100	26
	11.05.2008	Normal	Track	Bressanone	Brixia Meeting	0.40	110	44
Avg.						**0.38**	**130**	**53**
Min.						**0.26**	**100**	**26**
Max.						**0.49**	**180**	**88**
	07.06.2008	Normal	Track	Cinisello Balsamo	U18 Club Championships Final A	0.38	108	41
	07.06.2008	Normal	Track	Imola	U18 Club Championships Final A1	0.67	137	92
	07.06.2008	Normal	Track	Maiano in Riviera	U18 Club Championships Final A2	0,38	85	32
	07.06.2008	Normal	Track	Bastia	U18 Club Championships Final A3	0.55	71	39
	08.06.2008	High	Road	Borgo Valsugana	Italian 20 km road championships	0.29	108	31
	13.06.2008	High	Track	Torino	Italian Junior – U23 Championships	0.19	106	20
	20.06.2008	Normal	Track	Bressanone	Italian Masters Championships	0.11	108	12
	21.06.2008	Normal	Road	Molfetta	Italian Grand Prix	0.10	65	6
	27.06.2008	High	Track	Firenze	Club's Top Challenge	0.32	121	39
Avg.						**0.33**	**101**	**35**
Min.						**0.10**	**65**	**6**
Max.						**0.67**	**137**	**92**

Practical application to international competitions

For comparison with the hypothetical ideal race, the data from the most important international competitions since 2000 are shown in Table 2. The reader may reflect and try to identify those events coming closest to the ideal we have identified, but should also take into account the following variables:

- The World Youth and World Junior Championships include two races that take place on the track while the other championships include three races that are staged on the road;
- The World Race Walking Cup has since 2004 included races for junior men and women, giving a total of five races.

Conclusion

As we stated above, the system we have developed is not the absolute truth when it comes to evaluating the work of race walking

Table 2: Most important international race walking events since 2000

	Total Red Cards	Total DQs	Cautions		Total Cautions	% RC/ Cautions	Average Judge's points	Consistency Ratio	Race value
Olympic Games									
Sydney 2000	132	14	142	150	292	45%	78.12	0.32	24.86
Athens 2004	111	10	187	91	278	40%	59.79	0.27	16.16
Beijing 2008	108	10	148	109	257	42%	69.08	0.28	19.34
Sub Total	**351**	**34**	**477**	**350**	**827**	**42%**	**69.00**	**0.29**	**20.05**
World Championships in Athletics									
Edmonton 2001	160	34	157	126	283	57%	109.21	0.64	69.62
Paris 2003	173	33	156	120	276	63%	108.61	0.57	62.15
Helsinki 2005	154	28	239	94	333	46%	104.24	0.55	56.86
Osaka 2007	130	20	186	112	298	44%	88.01	0.46	40.62
Sub Total	**617**	**115**	**738**	**452**	**1,190**	**52%**	**102.52**	**0.56**	**57.32**
World Cups									
Torino 2002	243	33	188	241	429	57%	83.31	0.41	33.94
Naumburg 2004	279	37	246	153	399	70%	81.84	0.40	32.56
La Coruna 2006	247	30	313	141	454	54%	79.22	0.36	28.87
Cheboksary 2008	243	26	200	224	424	57%	74.36	0.32	23.87
Sub Total	**1,012**	**126**	**947**	**759**	**1,706**	**59%**	**79.68**	**0.37**	**29.76**
World Youth Championships									
Ostrava 2007	31	2	55	45	100	31%	104.29	0.19	20.19
Sub Total	**31**	**2**	**55**	**45**	**100**	**31%**	**104.29**	**0.19**	**20.19**
World Junior Championships									
Kingston 2002	39	5	25	61	86	45%	99.16	0.38	38.14
Grosseto 2004	31	3	62	34	96	32%	89.40	0.29	25.95
Beijing 2006	48	5	50	28	78	62%	108.99	0.31	34.06
Bydgoszcz 2008	53	6	43	46	89	60%	99.13	0.34	33.70
Sub Total	**171**	**19**	**180**	**169**	**349**	**49%**	**99.18**	**0.33**	**33.06**
European Championships									
Munich 2002	110	19	121	113	234	47%	92.42	0.52	47.89
Goteborg 2006	50	5	64	77	141	35%	81.12	0.30	24.34
Sub Total	**160**	**24**	**185**	**190**	**375**	**43%**	**86.77**	**0.45**	**39.05**
Total	**2,342**	**320**	**2,582**	**1,965**	**4,547**	**52%**	**92.59**	**0.41**	**37.95**

judges. Although they do not affect the primary function and analysis of the system, there will tend to be weaknesses with regard to the quality of the data, which also exist in many other fields. These weaknesses are as follows:

1. The collection of statistical data involves the examination of many races and reliability is only ensured if a statistically sufficient number of races are considered (in no case less than 50 races) and the more the better. We recommend the continued monitoring of values and the use of mathematical tools to test the data, such as the ANOVA Test (Analysis of Variance).
2. The use of statistics cannot be the only tool for evaluating the judging of race walking events. There are parameters that our system has not yet taken into consideration. These include timeliness of decision making, the different positioning of judges on the road, and other factors that may change the sense of what is obtained by examining the statistics.

Taking these points into consideration, we can continue our work to correct and improve the system in the future.

Please send all correspondence to:
Nicola Maggio
email : maggio.nicola@alice.it