21:1; 39-51, 2006

Decathlon: Towards a balanced and sustainable performance assessment method

By Wim Westera

ABSTRACT

The author argues that the current IAAF decathlon scoring tables display unacceptable bias as they favour some events over others. Performances in the sprints benefit disproportionately to those in the throwing events and the 1500m. Moreover, the system is intrinsically unstable and tends to increase the differences between disciplines over the course of time. This paper investigates alternative scoring methods. It elaborates a well-grounded procedure to express the performance scales of the events in a normalised form in order to allow comparisons. Three alternative scoring models are developed as candidates for replacing the existing model. These are based on 1) a power law description, 2) a parabolic description and 3) an exponential description, respectively. The proposed methods are uniform over the events and support self-stabilisation. They combine practical evidence and sound principles. Calibration to the current model is performed with existing data in order to enable a smooth transition from current practice. Overall effects are limited, if not negligible. Under each of the proposed models two of the current all time top 100 performers would improve their ranking substantially and all three models indicate the current number two in the ranking, Thomás Dvorák (CZE) should actually be the world record holder.

AUTHOR

Dr. Wim Westera is a physicist and educational technologist. In his role as Head of Educational Implementation at the Educational Technology Expertise Centre of the Open University of the Netherlands he combines science, educational media development and innovation. He leads a group of some 70 educational designers, media specialists and IT developers. He is also a reasonable master athlete and racing cyclist.

Introduction

The decathlon is often referred to as the ultimate athletic competition as it emphasises the versatility of the competitors, who are challenged to combine excellent physical power, explosiveness, technical (psychomotor) skills and endurance. The idea of the combined event goes back to the ancient Greeks, who introduced the pentathlon (running, discus and javelin throws, jumping and wrestling). The winner was considered the most complete athlete and was accorded almost godlike status. Today, the Olympic champion or world record holder in the decathlon is still called the 'World's Greatest Athlete'.[1]

As the core idea of the decathlon is all-roundness, a well-founded scoring model is necessary to combine the performances in the various disciplines into a total score. For this purpose official scoring tables have been

developed for each of the events. Over the years, these tables have been the subject of extended debates about their fairness and validity. On several occasions amendments have been made in order to remove manifest imbalances between disciplines. These may easily arise because of new training approaches or new materials. Indeed, some disciplines steadily evolve (like the pole vault) while others tend to stagnate (like the 100m). The present tables have been used for more than 20 years. Which makes it worthwhile to carry out a review of their appropriateness.

As will be shown in this paper, the current scoring method appears to be cursed with unacceptable bias and needs a conceptual revision. We will elaborate a well-grounded procedure to express the diverse performance scales in a normalised form and allow fairer comparisons. We will also present and evaluate alternative scoring models as candidates for replacing the existing model. In conclusion we will go into the consequences of the new models for current rankings and records.

Unbalanced decathlon score assignments

The current decathlon scoring tables have been used without modification since the 1980s. We will demonstrate that today quite some unbalance has arisen. This becomes manifest on many occasions, even though most people accept the scoring outcomes indiscriminately as a fact of life. The imbalance can be made visible by collecting the results of an exemplary group of athletes. This would enable finding out which disciplines are the most profitable for the athletes and which are the most unfavourable. In other words, where do the athletes collect their points?

To answer this question we will use the all-time top 100 decathlons as ranked by the International Association of Athletics Federations[2]. This group comprises outstanding athletes who go in for the decathlon at a sufficiently professional level to warrant reliable and integer datasets that reflect the veritable notion of the decathlon. It is essential that we consider such an exemplary group. Obviously, amateur athletes, joggers and jokers may occasionally participate in a decathlon, but they are likely to show disproportional failures at certain disciplines due to poor training, lack of technical skills or insufficient versatility. Such a sample would inevitably lead to corrupt data sets. In contrast, the very top decathletes are assumed to cover each discipline at a (world) top level: this matches the decathlon's core assertion of all-roundness over the disciplines. Figure 1 shows the average of the scores in the individual disciplines from the all time top 100 decathlons.

It turns out that there are quite significant differences between the disciplines. The athletes seem to profit disproportionately from the long jump, the 110m hurdles and the 100m, while - in contrast – the 1500m, javelin, discus throw and shot put are highly unfavourable. Apparently, top decathletes tended to specialise in sprinting, which indeed may be regarded a common denominator of the long jump, the 110m hurdles and the 100m. Throwing capabilities and endurance, however, seem to be far less profitable and may even interfere with sprint performance.

One might be tempted to infer from this pattern that performances in the throwing events and 1500m are lagging behind and thus leave more room for further improvements than the sprint-based events, but this conclusion is not tenable. First, this would reflect an embarrassing disregard of the fact that the top decathletes go to the limits of each discipline in any possible way. It would be naive to assume that substantial improvements were possible, even if radical changes in training were to be applied. It is not the athletes who should be blamed for the apparently sub-optimal performance, but the scoring method itself. In principle, the top 100 average score should be equally distributed over the events. Indeed, decathletes who have achieved scores that rank in the all time top 100 are the only candidates to set the empirical standards for genuine all-round performances. Any anomalies in the performance pattern of Figure 1 should thus be ascribed to imperfections of the scoring method.

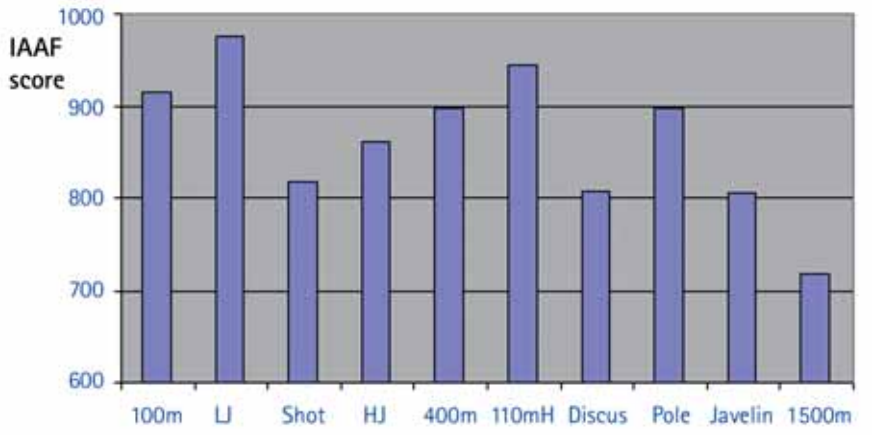

Figure 1: Average scores of the decathlon all time top 100 (version July 2005).

Secondly, the self-corrective nature of the performance pattern is refuted by the numeric gradient in the scoring tables: a 1% increase of the long jump performance yields 19 extra points, whereas the same increase in discus throw yields only 9 points and javelin and shot put will bring only 10 points. Improving the 100m performance 1% would produce 24 points! This pattern implies a positive feedback loop for sprinting-based performances at the expense of throwing skills and endurance. Therefore, the different scores in Figure 1 cannot be regarded as a temporary or coincidental deviation from equilibrium; on the contrary, the pattern seems to be highly unstable and will probably show increasing differences between disciplines in the course of time. Current decathletes are excellent sprinters. Apparently, this is a self-establishing fact, because further specialisation in the sprints pays off. As such a tendency conflicts with the premise that the decathlon champion should be the best all-round athlete rather than a solid sprinter, modifications of the scoring method are inescapable.

The current scoring method

Even though we have observed some problems with the current scoring method, we want to emphasise that the method as such is quite sophisticated. It uses objective, unambiguous, quantifiable performance data (i.e. time and distance) and avoids the subjective assessments of jurors (aesthetics, expression) that cause so many problems in the rating of gymnastics, figure skating or dressage. It also avoids complicated and probably unfair multi-stage accounting systems, like the system of rally points, games and sets in tennis. Such systems are used for historical rather than logical reasons. Furthermore, the scoring tables are progressive in kind, as will be explained below. These are far better than the linear systems that are still being used elsewhere, for instance in the combined events of speed skating.

The current scoring tables were adopted in 1984 after extensive discussions, negotiations and compromises. The process took into account an abundant amount of empirical evidence. Basically, the current scoring method for each discipline is covered by a mathematical expression of the type:[3]

$$S(P) = A.(P-B)^c \quad (1)$$

- P is the performance (i.e. the achieved distance in the long jump).
- S is the score (the number of ascribed points).
- A, B en C are event-dependent parameters that define the nature of the scoring table.

For running events (P- B) should be replaced with (B-P) because of the descending nature of performance with time. Note that the performance assessment method comprises two

stages: first the performance P is measured (in units of time or distance), next the performances are converted to a score S in order to allow addition. Clearly, it is this second stage of assessment that is problematic.

Figure 2 shows the scoring curve for the long jump, according to equation (1). It uses the following values: A=0.14354, B=220 cm, C=1.40, while P is expressed in cm.[4]

Such scoring curves have the following characteristics. The parameter B defines a threshold value (2.20m), below, which no score is assigned. This is substantiated by the assumption that any athlete is assumed to reach such distance without any effort whatsoever and therefore will not receive any points for performances below B. Above this threshold value the performances are rated through a slightly progressive curve, the nature of which is mainly determined by parameter C. The underlying idea of this nonlinearity is that an improvement at low performance levels is much easier than an improvement at high performance levels. Indeed, improving the long jump from 8.00m to 8.20m is far more impressive than the same improvement from 4.00m to 4.20m as the scoring table assigns 51 extra points against 33 extra points. The overall scaling of the curve is determined by a parameter A. Thus, the current decathlon scoring method comprises a set of 10 power laws that is specified by 30 calibration parameters (A, B and C for each of the 10 events).

In due course, several inadequacies seem to have crept into the scoring method. Moreover, the theoretical foundation of the formula is weak. The progressive form is assumed to be associated with the kinetic energy that an athlete has to develop during the event, irrespective of whether a run, jump or throw is involved. This would suggest that performance is proportional to squared speed (v^2), which would indirectly suggest a progressive form with power 2.0. In practice, however, it was necessary to apply power functions with exponents (parameter C) well below 2.0, with some variations over the disciplines (i.e. javelin C=1.08; long jump C=1.40; 100m C= 1.81). Note that the progression of the curves is partly determined by the threshold values B (i.e. javelin B= 7.0 m; long jump B= 2.20m; 100m B= 18.0 sec). The current tables are pragmatic in kind rather than based on solid explanation. Consequently, some arbitrariness is involved. Indeed, it is difficult to explain why the long jump scoring table should start at 2.20m rather than at 2.40m, 1.80m or even at 0.00m. An additional weakness of the current system is its inability to self correct in due course: as indicated before, the current scoring system is intrinsically unstable in that differences between disciplines tend to increase rather than fade away. These observations amplify our call for a revision.

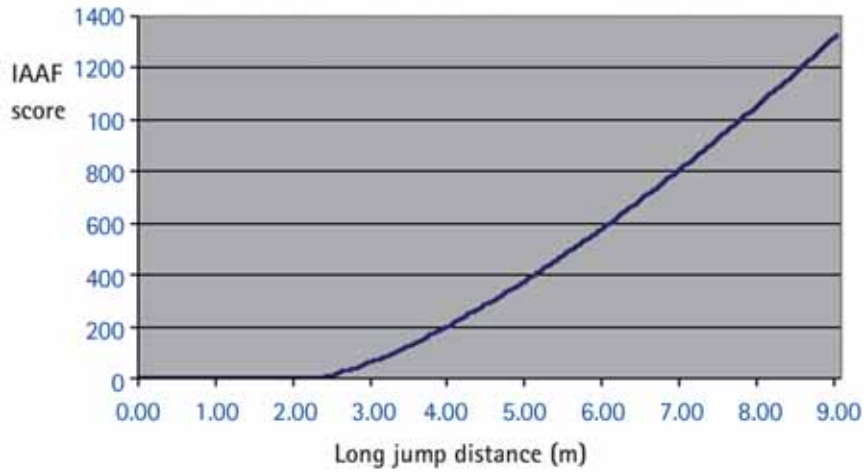

Figure 2: Current scoring curve for the long jump.

Premises for justified rating

Before we elaborate alternative scoring methods, we will list the basic requirements they should meet. The envisioned methods should:

- Allow a fair comparison between events;
- Be uniform over all events (this follows from the starting point of the decathlon);
- Use objective standards (distance and time measurements);
- Be grounded in empirical evidence from the decathlon (practical significance);
- Be based on sound principles and logic (consistent, transparent and substantiated);
- Be stable over time and thus possess self-stabilising characteristics;
- Allow smooth transitions from the existing model (acceptability).

Naturally, the method must be credible and acceptable in that it should not degrade obvious top athletes to middle-of-the-road performers. This holds even when it comprises the paradox that we reject the current method but still demand the new system to yield more or less similar outcomes.

Next, we will explore new scoring models in two stages. First, we will develop and discuss a procedure to express the diverse performance scales in a normalised form in order to allow comparisons. Second, we will develop alternative expressions for the scoring function S(P).

Normalisation of performance scales

Any scoring method for combined events is doomed to compare apples and oranges, as it combines and reckons with different processes, different variables and different types of performances. In order to enable a comparison of one type of performance with another type of performance we need a way to transform each performance scale into a normalised form. As will turn out below, such normalisation of decathlon performances can be achieved much easier than in the case of apples and oranges.

In the current system, throwing and jumping performances are expressed in a straightforward way by the achieved distance: larger distances correspond with better performances. In running events, however, performances are expressed in the length of time needed. Consequently, running performances and their quantification are inversely related, rather than linearly: the less time needed, the higher the score. In order to achieve a sensible normalisation procedure we first have to align time measurement and distance measurement. Let the performance in a certain event be quantified by a performance variable P. To be consistent in terminology, high performances should correspond with large values of P. Figure 3 displays such a performance axis.

Figure 3: Performance axis with threshold value P0

In accordance with the current system, we may define a threshold performance P_0 that would correspond with the performance below which no score is assigned (S=0). In the current system P_0 is given by parameter B in equation (1). The value P=0 would correspond with the ultimate inactivity. Naturally, such performance scale easily matches the distance scale of throwing events and jumping events. For running events, the performances should no longer be expressed in units of time, but rather in units of speed or, likewise, in units of reciprocal time. If so, the value P=0 would correspond with the ultimate inactivity: indeed, it would take forever.

With such alignment of the throwing-jumping events and running events in mind, the definition of a normalised performance scale can be formalised by a linear transformation of the performance variable P. We would need two calibration values, P_0 and P_1, to define the normalised performance $P_N(P)$ of a performance P in a particular event:

$$P_N(P) = (P-P_0)/(P_1-P_0) \quad (2)$$

From equation (2) it follows that:

$$P_N(P_1) = 1 \quad (3)$$

and

$$P_N(P_0) = 0 \quad (4)$$

Here, P_1 represents the high-end calibration value of the performance scale, whereas the performance threshold P_0 is the low-end calibration value.

As for the high-end calibration value P_1, we would need a stable reference value that represents high performances. While maximum performance is indefinite, per se, we propose to equate P_1 with the average of the all time top 100 performances that have been used before in Figure 1. This choice may seem somewhat arbitrary, but as it being used for the relative alignment of the performance scales of the various events, it is not critical. We might have chosen the top 50 average as a reference as well, or even the world record data. This would indeed produce different transformations (cf. Equation (2)), but it would still preserve the idea of normalisation. Actually, what matters is that the data is representative. By using the all time top 100 average existing peaks and exceptions are dimmed by the statistics. The current averages of the performances in the all time top 100 decathlons[2] are listed in Table 1.

Table 1: Average scores of the all time top 100 decathlon performances

Event	P_1 All time top 100 average
100m	$(10.76s)^{-1}$
Long Jump	7.66m
Shot Put	15.47m
High Jump	2.06m
400m	$(48.22s)^{-1}$
110m Hurdles	$(14.23s)^{-1}$
Discus Throw	46.92m
Pole Vault	4.95m
Javelin Throw	64.46m
1500m	$(4:34.12min)^{-1}$

So, when we choose the values of P_1 to correspond with the average performances listed in Table 1, we conform to the idea that athletes who achieve all time top 100 decathlon scores have the same normalised performance (e.g. $P_N(P_1)= 1$) for each event. Consequently, this means that 10.76s for the 100m is the same performance as a long jump of 7.66m and so on. In fact, starting from the principle of all-roundness, this is the only sensible decision. It also means that the associated scores $S(P_1)$ (cf. Figure 1) should be the same for each event. Note that this (arbitrary) normalisation of the performance P does not mean that P_N has an upper limit of 1; indeed, P_N may become larger than 1 if $P> P_1$, naturally when performances exceed the top 100 average (which may occur regularly).

For the low-end calibration of value P_0, the official threshold parameters B, as defined in the current scoring method (cf. Equation (1)) may seem interesting candidates. However, in contrast with the values of P_1 (the all time top 100 averages), which represent exemplary, real and reliable data, the current values of B are quite problematic, because they are the result of accumulated modifications loaded with historical bias and lack logical foundation. The origins of the existing values B are unclear and their fairness is questionable. Therefore, the indiscriminate import of these existing threshold values, which for their part may be an important cause of the unbalance in the current scoring method, is not acceptable. This becomes manifest when we list the current IAAF threshold values B relative to the high-end performances P_1 (cf. the third column in Table 2).

It appears that the relative positions of the current IAAF threshold values B are very different for the different disciplines. Relative positions spread over a factor of 7, ranging from 0.085 for discus throw to 0.598 for the 100m. Theoretically a different threshold for each event might be plausible, because, indeed, each discipline requires different techniques, different muscles and different procedures. Yet, the current thresholds seem to display quite a degree arbitrariness and break through the uniformity of the disciplines without any foundation. Our proposition here is that in the absence of any reasoning about the physical parameters that would substantiate the necessity of different thresholds, a uniform approach over the disciplines is indicated. Indeed, if uniformity over all disciplines is our starting point,

Table 2: Current thresholds B, relative thresholds B/P1 and suggested uniform thresholds P0

Event	B (IAAF threshold performances)	Relative threshold B/ P1 (B^{-1}/P_1 for running events)	Suggested uniform thresholds P_0 (using P_0/P_1=0.340)
100m	18.00 s	0.598	$(31.64s)^{-1}$
Long Jump	2.20m	0.287	2.60m
Shot Put	1.50m	0.097	5,26m
High Jump	0.75m	0.364	0.70m
400m	1:22.00min	0.588	$(2:21.82min\ (m:s)\ )^{-1}$
110m Hurdles	28.50	0.499	$(41.85s)^{-1}$
Discus Throw	4.00m	0.085	15.95m
Pole Vault	1.00m	0.202	1.68m
Javelin Throw	7.00m	0.109	21.92m
1500m	8:00.00min	0.571	$(13:26.16min)^{-1}$
Average	-	0.340	-

P_0 should be at the same position for each event. This means that we want P_0/P_1 to be a constant. A first approximation of P_0/P_1 would be the average of B/P_1, which yields a ratio of 0.340. Using this ratio produces a uniform estimate for the threshold values for each discipline (cf., Table 2, fourth column). Note the substantial differences between our uniform threshold values P_0 (fourth column) and the current IAAF thresholds B (second column), especially in the running and throwing events.

Current scoring method: comparison of events

The performance normalisation procedure described above allows us to display the current scoring curves (cf. Equation (1)) at normalised performance P_N (cf. Equation (2)). Figure 4 displays the results for 5 of the events. Similar curves result for the other events.

From Figure 4 we conclude that the unbalance of the scoring is not restricted to high end performances as was inferred from Figure 1, but that it is present at all performance levels. Note that the curves not only have different scoring levels, but also very different curvatures and associated gradients. These different gradients imply that equal (normalised) performance improvements are rated differently in each discipline. The calculations confirm our preliminary conclusion that these differences cause the scoring system to be intrinsically unstable. We remark that the calculations indicate that throwing events (shot, javelin and discus) have very similar curves, which differ only up to 4%. Such resemblance might be expected with events that technically have many points in common. Similarly, running events seem to display a common pattern too: a steep rise at high performances. Yet, the differences in running scores are much greater, as is the case for the jumping events.

Note that the curves in Figure 4 only represent an intermediate stage of our analysis, because the normalisation affects only the horizontal scale, while the vertical scale is kept unchanged. As a consequence, one may signal some inconsistency while the horizontal scale uses the uniform threshold values of P0, according to Table 2, whereas the vertical scale still uses the current IAAF thresholds B, according to Equation (1). In the next section we will elaborate alternative methods to redefine the vertical scale.

Towards alternative scoring methods

So far, the divergence of the scoring curves in Figure 4 is an embarrassing confirmation of the inappropriateness of the current scoring

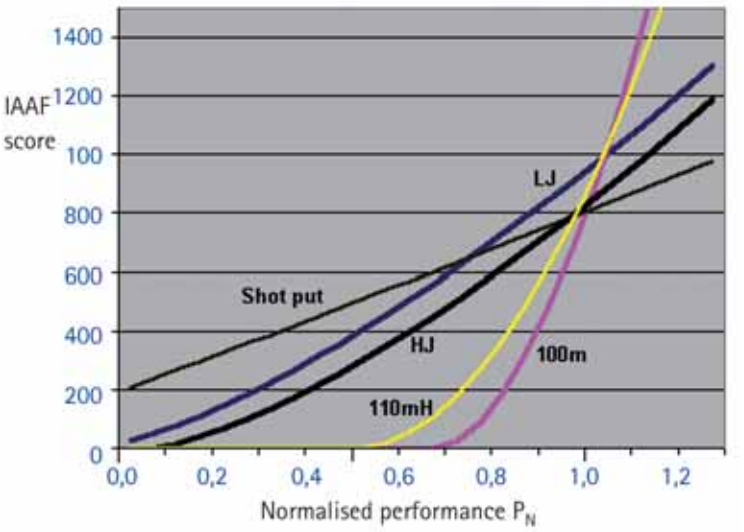

Figure 4: Current system scores at normalised performance.

method. If the normalisation procedure according to Equation (2) is accepted to be valid, the scoring curves of the various events should coincide rather than diverge. In accordance with the principles of the decathlon, the scoring should be uniform over all disciplines. This means that we have to redefine the scoring formula S (P) of Equation (1) in a uniform way. As a further constraint, we refer to the calibration values P_0 and P_1 that we have used to transform performance values P into a normalised form. For the threshold value P_0 it follows that:

$$S(P_0) = 0 \quad (5)$$

It turns out that the average all time top 100 decathlete has a score of 8639 points. Because the scoring curve S is assumed to be uniform over all events, it follows that for each event:

$$S(P_1) = 863.9 \quad (6)$$

Such empirical calibration ensures that the total scores of the all time top 100 decathletes stay in the same range as the current scores, in accordance with our premise.

Naturally, when we want to rewrite the scoring function S as a function S_N of the normalised performances P_N, according to Equations (2), (3) and (4), we obtain:

$$S_N(0) = S(P_0) = 0 \quad (7)$$
$$S_N(1) = S(P_1) = 863.9 \quad (8)$$

While uniformity over all disciplines is assumed for S_N, we have to find and substantiate a progressive curve with two fixed points, given by Equations (7) and (8). Below we will present three alternative approaches, the results of which are presented in Figure 5. The three models will be explained below.

Model I: Power law

In accordance with the current scoring method, we assume that S_N can be described by a power law:

$$S_N (P_N) = A.(P_N)^c \quad (9)$$

From Equations (2) and (9) we find that the regular scoring function S(P) can be written as:

$$S (P) = A.((P - P_0)/(P_1 - P_0))^c \quad (10)$$

Note that this power law approach significantly differs from the current IAAF power law in that performance in the running events is expressed in units of reciprocal time, rather than in units of negative time (cf. Equation (1)). Also, it follows from the uniformity of S_N that A and C are constant over the events, in contrast with the current IAAF scoring method which demands different values of A and C for each discipline.

The constraint in Equation (6) gives:

$$A = 863.9 \quad (11)$$

The only remaining unknown in Equation (10) is the power C. Naturally, the value of C determines the progressive form of the scoring curve, so it follows that C>1. A simple estimate of C can be obtained by conforming to the 10 IAAF power parameters C that are used in the current scoring method[3]. When we equate C with the average of these current powers we find:

$$C = 1.479 \quad (12)$$

The resulting score curve is displayed in Figure 5. When we compare the suggested power law curve of Figure 5 with the scoring curves in Figure 4, it turns out that the new curve has an intermediate position. Coincidentally, the new curve almost coincides with the high lump curve in Figure 4. Relevant data for this suggested power law curve are summarised in Table 3.

Model II: Parabolic

It was mentioned above that the progressive form of the scoring curve may be associated with the role of the kinetic energy that is developed by the athlete. Along this line of thought the resulting scoring curve should be parabolic, because the performance P is always expressed in units of distance or units of (reciprocal) time. This argumentation, however, is not very specific, and it omits the effects of the different techniques and constraints of the disciplines. Yet, there is another reasoning that underpins the likelihood of a parabolic scoring curve. To find a solid basis for the progressive behaviour we should return to the basic idea that progression reflects that the gradient of the scoring curve increases with performance. In mathematical terms we state the premise that the extra score $dS_N(P_N)$ that follows a performance improvement dP_N is proportional with the performance P_N:

$$dS_N(P_N) \sim P_N \,.\, dP_N \quad (13)$$

Indeed, achieving a performance increment dP_N at a high performance level P_N produces more points dS_N than the same increment at a lower level. Integrating Equation (13) gives a parabolic dependence:

$$S_N(P_N) = A \,.\, (P_N)^2 \quad (14)$$

Note that this parabolic curve is a special case of the power law of Equation (9), i.e. C=2. The scaling constant A is given by Equation (11).

As can be seen from Figure 5, the parabolic curve (C=2) is slightly more progressive than the

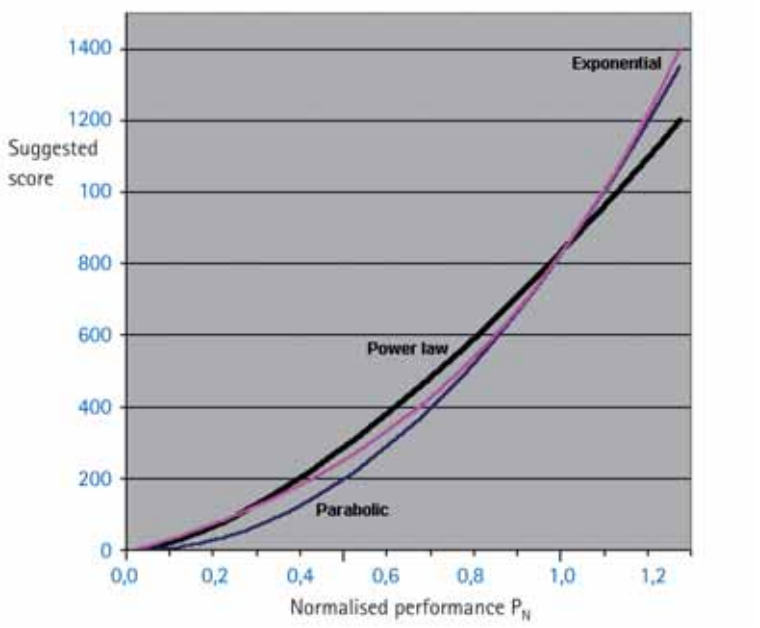

Figure 5: Suggested uniform scoring curves in accordance with three alternative models.

power law curve, which uses only a power of C=1.479. Differences between the two scoring curves are up to a few percent (0 – 40 points) in the high performance area (P_N>0.9) and rise up to 100 points at low performances (P_N=0.55). Relevant data for this suggested parabolic curve are summarised in Table 3.

Model III: Exponential

The progression of the scoring curve can also be approached with statistics. Indeed, progression may be assumed to reflect the reduced chance of success at increased performances. To define progression statistically we state that the extra score $dS_N(P_N)$ that follows a performance improvement dP_N is inversely proportional with the occurrence or frequency $f(P_N)$ of performance P_N in the population of decathletes:

$$dS_N(P_N) \sim 1/f(P_N) \,.\, dP_N \qquad (15)$$

While the frequency $f(P_N)$ may be assumed to descend monotonously - indeed fewer and fewer athletes will be able to achieve better performances -, a performance improvement dP_N is more greatly rewarded at high performance levels.

The next question would be: what evidence is available about the frequency $f(P_N)$? The standard approach to sort out $f(P_N)$ would be to take a random sample to represent the population of all decathletes. This, however, introduces two severe conceptual problems. First, while gathering results from decathlon competitions, national and international ranking lists and so on, we would be taking biased samples that only represent the local top 10 or top 50 participants and disregard large groups of modal athletes who make up the majority of the decathlon population. Secondly, the combination of data in different performance intervals, e.g. the combination of international data and sets of regional data, is not straightforward but should be linked with the relative occurrences in the performance intervals. Obviously, combining the results of the World Championships with the data of some unimportant event would not produce a representative sample for the decathlon population. This problem is circular in kind and thus irresolvable: to derive the performance distribution $f(P_N)$ from combined results in various intervals we would need to know the relative occurrences, which are given by $f(P_N)$ itself. Therefore, empirical occurrence data will not be of any help here.

A second approach would be to suggest a theoretical probability distribution, by investigating the conditions of the probability process. Although various well-known distribution functions like the Poisson distribution or the normal distribution may be interesting candidates, we would still need good empirical estimates of the distribution's mean and variance. Also, we have to consider that only (part of) the descending tail of such distribution is of relevance, because only the descending tail reflects increasing failure; in contrast, the ascending tail at low performances represents the fact that most athletes easily exceed these low performances.

These observations indicate severe difficulties in a straightforward and successful application of a statistical analysis. However, we have some indications that the performance distribution function might be approximated by the negative exponential distribution:

$$f(P_N) \sim e^{-\lambda P_N} \qquad (16)$$

where λ is a constant.

This choice is underpinned by the following arguments:

- The exponential distribution is often associated with the survival of species in biology or similar processes that account for failures and drop-outs, for instance the reliability of technical components. The process of survival has many things in common with sports events. Consider, for example, the high jump and pole vault, where the requested performance of athletes is incremented in steps, until eventually all competitors have dropped out. Theoretically all decathlon events can be mapped on to this approach and thus match a regular survival pattern.
- As will be demonstrated below, the premise of Equation (16) provides a monotonous progressive scoring curve and thus fits our objectives.

- Clearly, the probability function f in Equation (16) can be regarded the solution of the following differential equation:
  $df(P_N) \sim f(P_N).\ dP_N$ (17)
- This equation establishes the sensible premise that a performance increment dP_N causes a frequency change $df(P_N)$ that is linearly proportional with $f(P_N)$.
- The exponential distribution is simple in its form, it has only one parameter (λ) and it can be integrated analytically.

When we combine Equations (15) and (16) and integrate and make use of Equations (7) and (8), we obtain the following progressive expression:

$$S_N(P_N) = A\ .\ (e^{\lambda P_N}-1)/\ (e^{\lambda}-1) \quad (18)$$

Again A is given by Equation (11). To decide on the value of l we set the pragmatic requirement that the exponential curve has an intermediate position between the power curve and the parabolic curve. By minimising the total squared differences between the curves at the interval [0.1] we find λ=1.602. The resulting exponential curve is shown in Figure 5. Although our fitting procedure implies an intermediate curve, the exponential relationship creates a relatively strong progression at high-end performances (P_N>1). Differences with the power law curve are up to 15 per cent in the midrange (up to 60 points). Note that the inverse value of l represents that expect value of the performance PN. This would indicate an average performance of $\langle P_N \rangle = 1/\lambda = 0.62422$. Relevant data for this suggested exponential curve are summarised in Table 3.

Conclusion

All three suggested models meet the requirements for a justified rating for which we have expressed a need. The normalisation procedure allows a fair comparison between events. The proposed scoring methods are uniform over the events and support self-stabilisation. They combine practical evidence and sound principles. Various calibrations to the existing model would allow smooth transitions from the current method. As a consequence, overall effects are limited if not negligible.

In the all time top 100 ranking the average change is 10 positions for each of the models, which corresponds with relative improvement (or degradation) of 30%. The biggest leap is observed for the number 59 athlete in the current ranking (Mike Smith (CAN)), who may be assumed to be greatly underrated and put at a disadvantage by the current system because of relatively poor sprinting (100m in 11.23; 110m hurdles in 14.77). Both the parabolic method and the power method allocate

Table 3: Summary of alternative scoring methods

Event	P0	P1
100m	$(31.64s)^{-1}$	$(10.76s)^{-1}$
Long Jump	2.60m	7.66m
Shot Put	5.26m	15.47m
High Jump	0.70m	2.06m
400m	$(1:41.81min)^{-1}$	$(48.22s)^{-1}$
110m Hurdles	(41.85s)-1	$(14.23s)^{-1}$
Discus Throw	15.95m	46.92m
Pole Vault	1.68m	4.94m
Javelin Throw	21.92m	64.46m
1500m	$(13:26.16min)^{-1}$	$(4:34.12min)^{-1}$
I. Power law	$S(P)=A\ .((P-P_0)/(P_1-P_0))^c$	with A = 863.9 en C= 1.479
II. Parabolic	$S(P)=A\ .((P-P_0)/(P_1-P_0))^c$	with A = 863.9 en C= 2.000
III. Exponential	$S(P)=A\ .\ (e^{\lambda P_N}-1)/\ (e^{\lambda}-1)$	with A = 863.9 en l=1.602

Smith a rank of 8th; the exponential yields a rank of 4th. Likewise number 21 in the IAAF ranking (Uwe Freimuth (GER): 11.03, 14.66) enters the all time top 10: 6th (parabolic), 7th (power) or 5th (exponential).

From this we see that the alternative models seem to counteract the sprint bias of the current model. Remarkably, all three models indicate a new world record holder, or rather a reinstatement of the old record holder, as Thomás Dvorák's (CZE) 1999 performance in Prague outstrips Roman Sebrle's (CZE) subsequent mark from 2001 in Götzis, which is unanimously ascribed to 2nd. Dvorák's record would read 9232 points using the power law, 9469 with the parabolic or 9777 for the exponential curve. Note that these scores greatly exceed Sebrle's currently recognised mark of 9026, especially in the case of parabolic and exponential scoring due to the relatively high progression of the curves at world level performances. The medallists at the 2005 World Championships in Athletics[4] would remain unchanged under the three alternative methods, although Brian Clay's (USA) winning margin would be even more pronounced, due to the same effect.

In this paper we have shown that the current decathlon scoring method suffers from severe bias and produces unfair outcomes. It would need a revision to become more balanced and stable. We have demonstrated that it is possible to devise alternative scoring methods that are uniform, balanced and substantiated and that avoid the negative effects of the current method. On several occasion we have chosen to estimate or calibrate data by falling back on existing habits or data (e.g. performance thresholds P_0, power C) in order to connect to existing practice. One may wonder about the exact value of the power parameter C, or one may question the necessity to define thresholds P_0 at all. Indeed, other choices are possible and arguable, possibly with different outcomes and consequences, but the quintessence of this paper is to present a proof of concept of appropriate alternatives.

The presented models not only have greater plausibility, they also are much simpler and need fewer parameters. Instead of 30 parameters in the current model, the power law method uses only 22 (magnitude A. power C and 10 times P_0 and P_1), as does the exponential model (magnitude A, rate λ and 10 times P_0 and P_1); the parabolic method uses 21 (magnitude A and 10 times P_0 and P_1). This reduction is an improvement as, according to "Ockham's Law of Parsimony" or "Principle of Economy" (called "Ockham's razor") one should make no more assumptions than needed to explain ascertained facts.[5] It supposes that the same principle of simplicity prevails in the physical cosmos, since the laws of nature are governed by the tendency towards minimum energy and a minimum number of degrees of freedom.

Such a principle of economy would indeed fairly suit the efforts of decathletes who seek to challenge the limits of performance, equally in all events.

Please send all correspondence to:
Dr. Wim Westera – Wim.westera@ou.nl

REFERENCES

1. DECATHLON ASSOCIATION (2005) Retrieved 09-09-2005 at http://www.decathlonusa.org/index.html
2. IAAF (2005): All time ranking decathlon, Retrieved 09-09-2005 at http://www.iaaf.org/statistics/toplists/inout=O/ageGroup=N/season=0/gender=M/discipline=DEC/legal=A/index.html
3. KNAU (2004) Formules en constanten, commissie Wedstrijdreglement, cf art. 89, art.116 en hoofdstuk 17 art. 4 van het wedstrijdreglement, Retrieved 09-09-2005 at http://www.knau.nl/ .
4. IAAF (2005):Tenth World championships in Athletics Helsinki, Retrieved 09-09-2005 at http://www.iaaf.org/documents/pdf/3365/AT-DEC-M-10--0--.RS2.pdf
5. Dictionary of Philosophy (2005) Retrieved 09-09-2005 at http://www.ditext.com/runes/p.html