Scoring System Revision
Agenda item 9.1

Introduction & Background

Creation of a "Scoring System Revision" Working Group was decided at the 2016 CIVA Plenary meeting, following a French proposal along the following lines:

- Task a CIVA Working Group to assess potential revisions to the scoring system with the objective that, for a given competitor, the overall scoring obtained in a programme does not depend on the scoring of other competitors: a prerequisite to widespread real-time display, which itself is a necessary step towards increased public visibility and media coverage.
- FPS would be kept to derive detailed judging analysis.

Further details can be found in appendix.

It was decided that the Working Group would be composed of Matthieu Roulet (chairman), Alan Cassidy, Doug Lovell and Gilles Guillemard.

Summary Status

In the meantime, the FPS Working Group has been assessing potential improvements to the FPS in order to bring it closer to the "stability" objective mentioned above. As expected, the FPS Working Group came to the conclusion that the FPS framework could not allow to completely eliminate ranking variability as a given Programme unfolds. Nevertheless, the FPS WG developed an improvement in the FPS which very significantly reduces this variability (see the FPS WG report, under agenda item 6.1).

As of today, the "Scoring System Revision" WG as such has not been kicked-off. Nevertheless and even considering the very significant progress proposed by the FPS WG, the goals and ideas expressed in the 'Introduction & Background' here above are still worth pursuing. Even if navigating through this topic is potentially daunting, more focus will be put on this topic in 2018.

For now I draft below the way forward roadmap.

Roadmap

1. Further brainstorming on leads / ideas that might help solve the problematics
2. Down selection for further analyses
3. Development of concepts : groundrules, theoretical descriptions and rationales
4. Experimental phase 1: Retro-testing on a set of previous championship results, tuning of parameters

5. Experimental phase 2: Retro-testing of completed system on a different set of previous championship results, assessment of variance and acceptability
6. Definition of further potential improvement steps

Only then will we be able to make a final determination of whether the pursued objective is achievable or remains elusive.

If results of this work are found promising, the new scoring system could be put in place at least for new competition formulas (reference is made to CIVA Strategic ideas), and after further tests eventually for our traditional championships.

Matthieu Roulet
31 October 2017

APPENDIX

Rebalancing CIVA Scoring Method

Matthieu Roulet, August 2016

Introduction

Reference is being made to France proposal #2 for 2017, reproduced in appendix. This proposal recommends a revision of the CIVA scoring method, "with the objective that, for a given competitor, the overall scoring obtained in a programme does not depend on the scoring of other competitors: a prerequisite to widespread real-time display, which itself is a necessary step towards increased public visibility and media coverage."

Also, new types of aerobatic competitions are gaining momentum (e.g. Sky Grand Prix, WAM series) and would pursue the same ideals – so this is the right time to put some effort on this topic.

As we know, conciling both a level of fairness (elimination of unwanted variations) that is recognized as good, and non-moving scores, has appeared to be an elusive goal in the past. This paper gives some thoughts on how this might be achieved. Those leads shall be assessed by the FPS Working Group (and also by an ad hoc Working Group if appropriate), hoping this could also foster further thoughts on possible improvements.

Assuming the principles explained in this paper are deemed worth an in-depth exploration, significant work lies ahead for those Working Group(s) to establish a detailed definition of the concept (criteria, formulas and boundary conditions), backed by simulations on previous contest scores in order to achieve the required level of confidence.

Categories of scoring methods

As I see it, there are two main parameters in a scoring method:

1. the "size" of the dataset
2. whether there is a requirement to have "fixed" scores (i.e. non-moving respective rankings)

This leads to four main categories of methods to mitigate bias or unwanted variations:

- Category 1 -- the "narrow dataset view":

This is about using marks from all judges on a single flight only. The goal is then to eliminate some potential bias using this limited set of data. This category is used in many sports, where the most common solution is to eliminate the lowest and the highest global raw scores for each performance. Everyone is well familiar with the pros (readily understood, real-time fixed scores still better than raw scores, with *some* efficiency at eliminating "country-bias" in particular) and cons (cannot fix all fairness issues) of this category.

Note that in aerobatics competition, the dataset – while limited – is an order of magnitude larger than in most sports using this category: For each judge we have one mark per figure rather than a global mark. This opens the door to various options in our case, e.g.:

- elimination of the highest and the lowest score at sequence level
- elimination of the highest and the lowest score at figure level
- cap or gradual elimination depending on delta vs average / mean deviation, at figure level
- …

- Category 2 -- the "broad local dataset view":

This is about using all marks from all judges for all pilots, on a single programme only. With this considerably extended set of data, it becomes possible to make corrections based on detailed statistical analysis. Obviously our current FPS, as well as the previous TBLP, fall into this category, with the well-known pros (extensive fairness) and cons (non real-time score, moving rankings and therefore not appealing for the public & media).

- Category 3 -- the "broad deferred dataset view":

This is about using detailed judge performance based on Category 2 solutions applied on the previous programme(s), in order to apply a weighing factor on each judge for the current programme. Therefore in this Category we also deal with a considerably extended set of data (even potentially larger than in Category 2), which makes it possible to benefit from a detailed statistical analysis.

This Category is the major innovation proposed in this paper, and I strongly recommend to explore it further. I believe it has the potential to combine the pros of Category 2 together with the pros of having fixed scores, while the potential traps of this method could be fixed with a set of additional methodology rules.
This is what is discussed further below, after a brief description of Category 4 for the sake of completeness.

- Category 4 -- the "broad global dataset view":

This is about using all marks from all judges for all pilots, on several programmes, e.g. on all Programmes of the competition, before declaring rankings on each programme. It could be argued that by using such an extended set of data, a proper statistical analysis would give results of even higher fairness level than Category 2. This Category is not discussed further here, because of its cons which obviously collide against the wanted direction.

Discussion on Mixed Category 3

Category 3 solutions must address at least the following issues:

a. If the judge bahaviour in the current Programme is not taken into account in the processed scores for that Programme, then how do we prevent adverse impacts in an “unethical judge scenario” whereby the said judge would be tempted to give a gross advantage or disadvantage to a given pilot, even at the expense of a lower weight in subsequent Programmes ?
b. Even worse, what if this unethical behaviour (or simply unintentional gross bias) occurs in the last Programme, when there is no weight-factor incentive on this contest anymore because there

is no “next Programme”, and where the stakes are at their climax (e.g. to determine the World Champion) ?

c. What is the weight factor for the first Programme ? In case it is determined from past judge performance, then how is the weight factor established for a new judge ?

Issues a. and b. justify a case for introducing a Category 1 solution into the Category 3 method – what we could call a “mixed Category 3”: First, a Category 1 method is applied on each flight. Then, a Category 3 method is applied on the remaining judges or marks after normalisation of the overall weighting factors. Such a “mixed Category 3” method would clear the “gross bias” cases. In issue a. above, each judge has a double incentive to perform well: in case of significant deviation on any current flight, the judge faces the risk of having his/her marks affected by a lower weight (up to totally discarded) on that flight, plus the risk of a detrimental impact on his/her weighting factor for the next Programme.

Issue b., in addition, points to the necessity of carrying forward the judge’s weighting factor for the next contest he/she would judge. Behaviour in the last Programme could be subject to an extra impact on this weighting factor. And behaviour in the last Programme could be a major criteria for the next selection of judges.

This also solves issue c., except initial weighting factor for a new judge. This is further tentatively addressed below.

Creation of a Judge Index

The discussion above points to each judge having at any time a “Judge Index” (JI), subject to continuous evolution, programme after programme, competition after competition. In each programme, the respective JI in the Board of Judges determine each judge’s weighting factor after a simple normalisation process.

The JI scale and direction are tbd, and so are its detailed calculation rules. Nevertheless a few tentative guidelines can be mentioned :

- The JI is not what we currently call the RI: Not only is it an evolutive index along CIVA contests judged, but it does also take into account possibly more elaborate aspects than the RI.
- Example of more precise/more meaningful aspects than current RI could be:
  - ✓ Biased behaviour for or against a given country;
  - ✓ Number of “A” (Averages) as discussed in a rule change proposal this year;
  - ✓ Mean deviation in using marks, compared to the average mean deviation of the Board of Judges: if the method is deemed to create a risk that judges would score figures on a narrow set of marks believing it would help to achieve a better JI, then this could be fixed by introducing a bonus for judges using a larger range of marks;
  - ✓ …
- Judge performance on the most recent programme has more impact on the JI than less recent ones.
- The JI is subject to an “oxydization” system (degradation along time when a judge spends a long time without judging).
- The JI calculation rules can be easily adapted to take into consideration lessons learned / experience.

A new CIVA judge would be allocated a default “beginner” JI of tbd (obviously not as good as a “good” experienced judge, but better than a “bad” judge).

Beyond its direct use in a “Category 3 method”, introduction of such a Judge Index brings several advantages:

1. It measures judge performance along time and across contests.
2. It creates a sound emulation between judges.
3. It visibly rewards judge performance.
4. It eases introduction of new judges, where they can gradually build up their influence on scores as they gain positive experience.

Calculation of weighted scores : simplified example

In this simplified example let’s assume the following:

- 5 judges
- Judge A: JI=1
- Judge B: JI=1
- Judge C: JI=2 (weight = twice that of a judge with a JI of 1)
- Judge D: JI=1.3
- Judge E: JI=2.1
- Category 1 method is applied on each figure, with a highest and a lowest mark removed (note that a Category 1 method may be much more elaborate than in this simplified example, using e.g. mean deviation in combination with JI, …)

=> For a given competitor, on Figure 1, Raw Marks (RM) are:

A	B	C	D	E
7.5	7	6.5	8.5	6.0

=> Step 1: Marks from D and E are removed:

A	B	C
7.5	7	6.5

=> Step 2: Weighting Factors (WF) are computed (formula WF = JI/ΣJI):

A	B	C
0.25	0.25	0.5

=> Step 3: Weighted Mark (WM) for that figure is computed (formula WM = Σ(RM*WF)) :

6.875

Appendix – France Proposal #2 for 2017

NP2017-2 / FRANCE PROPOSAL #2

Subject: **Scoring System**

Proposal

Task a CIVA working group to assess potential revisions to the scoring system with the objective that, for a given competitor, the overall scoring obtained in a programme does not depend on the scoring of other competitors: a prerequisite to widespread real-time display, which itself is a necessary step towards increased public visibility and media coverage.

FPS would be kept to derive detailed judging analysis – which would be used for instance in two different ways:

- as a judge selection tool for the next contest;
- to apply a weighting factor on each judge's marks in the next programme(s) (with weighting factor on the first programme determined by past performance, and performance on the last programme heavily impacting judge index for next selection, etc).

Rationale

As shown by plenty of other sports popularized by the media, a paramount requirement to enthrall the audience is a real-time scoring system: After each competitor's performance, a time or distance or number of points etc, is displayed and the audience immediately gets relative rankings.

If we want to make our sport more appealing to the public and media, it is therefore crucial to display real-time scoring e.g. on giant screens at the contest site. But here we face a fundamental credibility issue in our sport: In which other sport can we have a situation where competitor A is said to have performed better than competitor B but then after competitor C performed, finally no…wait…B did better than A ? This is not understandable for any public or media and we believe goes against the increased visibility we are looking for.[1]

[1] The current FPS scoring system has been designed to maximize fairness, and we certainly do not challenge the current system in this respect – it does the job marvelously and we are convinced that no system with non-moving overall score could prove as effective as FPS in resolving bias and unwanted variations
However we believe another requirement -- credibility of our sport – must now come into play in addition to fairness. What we suggest is a rebalance between fairness and credibility. We have been putting all our eggs in the basket of absolute fairness and statistician expertise – and again this led to a great achievement in this respect. Yet statistics experts could not care less about other factors that might hinder progress of our sport. We must take ownership of what we really want to achieve in this respect. CIVA owns the vision of where we want to go and what balance is right for our sport, not experts in statistics.
The fact that all media-appealing sports we can think of, which have a similar "judging" issue – and that all those sports have had a long history, money and the means to develop absolutely fair systems like FPS -- chose not to take this path, maybe should trigger some further thoughts.
There are a number of potential schemes (more or less simplistic, more or less fair, etc) to achieve an "acceptable level of fairness" with non-moving scores, hence the need to investigate and assess.

Therefore, we believe a prerequisite to real-time display is the development of a scoring system whereby the number of points obtained by a given competitor is not subject to modifications after his/her performance (while still keeping fairness as a key requirement). We believe there are several ways to achieve that, and would like to promote the idea that a working group be tasked by CIVA to come up with an agreed solution.