Processing Judges marks and CIVA's FairPlay System (FPS)

A thorough review of why a "system" is necessary in aerobatic competition judging, and what FPS does for us



Sports Results and Judging Systems

In most competitive sports selecting the winner is easy ... it will be the first race-car past the finishing post, or the football team that scores the most goals, and so on. However some sports require experienced judges to rank the artistic and technical skills on display, and competition aerobatics is one of many activities where it takes a trained expert to tell how well each performance has met the standard required. Where such complicated judgements are required it is normal to assume that the performance can theoretically be perfect, so we simply need to count the "errors" that are seen and calculate the mark for each item by subtracting the total of errors seen from a fixed number - the winner of course is the one with the highest remaining score after adjusting for complexity and other factors.

An unfortunate aspect of these subtractive marking processes is that skill variations between judges tend to have a reversed effect. A less experienced or more timid judge is unlikely to recognise so many errors and will often award higher marks in a relatively narrow range, and these are likely to influence the result rather more than a judge with greater experience who is liable to see more downgrades - and so give lower marks and with a broader spread. It is also very difficult for any judge to prevent honest preferences and dislikes from affecting his or her decisions, whether these are applied consciously or not. At international events the influence of national characteristics can be intrusive and unusually hard to avoid.

Practical aerobatic judging

At aerobatic events Judges use their skills to cumulate the downgrades for each figure to the nearest half-mark, then subtract this total from the 'perfect' ten to give a mark which can range

from a maximum of 10.0 down to 0.0 or numeric zero. In addition there are specific occasions where fleeting hard-to-spot technical errors, such as when a snap-roll, tail-slide or spin does not display some essential characteristic, are 'perceived' and we write PZ to denote a Perception Zero, and also if the figure flown is not the one specified on the judges paperwork then an HZ is used to denote that a 'Hard Zero' has been applied. The PZ is a personal view from each judge and must be evaluated just like the numeric marks, whereas if any judge has given a HZ then the Chief



Judge must confer with the judging panel and decide either that the HZ should be applied for all judges, if possible using a video recording to guide this process, or the HZ must be rejected and

the figure fully marked. For occasional lapses of concentration a judge can also say "Oops – missed that one!" and ask for a suitable "average" mark to be generated by the system on his behalf.

Settling differences of opinion

For humans the usual way to handle collections of potentially unreliable opinions is to encourage as many observations as possible and then average them to minimise the influence of any unusual elements. This is a valid strategy as long as we can also accept the occasional disturbance that the questionable or wayout judgments will almost certainly cause. Final championship score differences between the leading aerobatic pilots however can be very small, and to accept every mark without question could easily lead



to publishing the wrong result. There should be a better way to identify marks that simply "don't fit" so that they can be given the attention that they deserve, and with FPS there certainly is.

Combining this into a plan ...

All the "raw" information from the judges goes into the scoring computer. What we need now is:

- A preparation system to overcome the effect of differences in judging styles and ability.
- A way to detect 'unusual' marks when compared to other judges marks for the same figure.
- A practical test so that we can evaluate unusual marks as either "OK" or "Not-OK", and ...
- A method for substituting a more suitable mark where a Not-OK decision requires it.
- All of this must be done in a completely 'open' way that allows Pilots and Judges to see what
 has been done, and with enough supporting information for everyone to assess just why any
 changes have been made.

Of course – the computer can <u>not</u> judge! But it can make very smart comparisons between what each judge says and, on the reasonable assumption that the dominant panel view is the 'correct' one, it can painstakingly analyse every element and employ sound mathematical techniques to reach a result that treats each judges' output in a fair and balanced way, and where necessary ensure that this always errs in favour of the pilot.

How to Compute the Results?

Over the years we have moved away from plain raw marks and its unavoidable problems, briefly through 'Bauerising', and then for some years CIVA used a statistical solution called TBLP in which a simple all-pilots/all-figures/all-judges table was used to compare all the marks together, substituting averages from the surviving judges where a mark failed the SD

based acceptance test. With TBLP however every mark from every pilot affected every other mark, and while it

provided some benefits it was said that judges could adapt their marking style to get an artificially improved result and eventually the confidence of pilots and contest administrators was lost. Rather than risk a return to using raw marks, CIVA set out to create a better solution.



CIVA's FairPlay System

The process was developed during 2005 from a completely fresh approach that combined our comprehensive championship judging experiences with a number of robust statistical testing processes to meet the very high analytical standards required. The result has proved to be a reliable scoring system which has built a good level of trust among judges and competitors alike.

The system works within the following broad headings:

1. Separate the Raw Marks into figure Groups

First the system assembles the judges "raw" marks into groups on a figure-by-figure basis, so that like is always compared to like and different opinions of the same thing can be precisely reviewed. For Free and Free Unknown sequences where figure composition is more flexible a 'SuperFamily' system is used to group similar types of figures together to ensure that the judgement comparisons remain on a like-for-like basis.

2. Balance the Judges within each figure Group

An essential first step with each group is to re-balance the judges marks so that no Judge has more or less influence than any other. The statisticians word for this balancing act is 'normalisation', and without it comparisons between the judges would simply not be valid. In our normalisation each judges complete set of non-zero marks is moved up or down and the scatter of the marks squeezed or expanded about their centre so each then has the same overall effect as the panel average. This completely resolves the experienced / inexperienced judge dilemma, the influence of every judge now being equal. This is the move that changes the pilots marks from simple whole and half numbers to many decimal places.

3. Identify and resolve "Unusual" Marks

For each group of marks FPS calculates an idealised table of 'Fitted Value' marks that is matched to each judges own style. A "statistical confidence test" is now carried out to check the validity of each normalised mark

Processed Marks Check-Sheet - Pilot 013 Nick Onn (GBR) Sukhoi 26M G-XXV Unlimited - Power level - Programme 1 FAI 25th WAC 2009, Silverstone, Northamptonshire, 20th Chief Judge: Cramm Hill (non-scoring)
Judges: 1 - Graham Hill (non-scoring)
1 - Graham Hill (RPR) 2 - Vladimir Kolet
4 - Francis Ilie(GRR) 2 - Vladimir Kolet
7 - Tomas Korinek (CZE) 8 - Hannes Gr
11 - Lyudmila Zelenina (UKR) 81 7 6.59 6.79 6.41 7.22 6.0 Lo 3.0 6.05 5.86 5.30 8.0 8.09 7.0 **7.33** 32 7 OK 5.0 **5.77** 6.0 6.01 7.0 50 7 CHZ

against it's corresponding Fitted Value. If the test meets the FPS confidence requirement then the mark is accepted and carried-forward to the next stage, whereas if the test fails then the original raw mark is labelled 'Missing'. In this way every normalised mark is in turn either accepted or rejected. When this initial group processing is complete, if any raw mark has been set to Missing then the normalisation procedure is re-run and Fitted Values re-calculated from the very beginning - but of course now without any of the

rejected 'missing' marks. These new Fitted Values, being free of all influence from the rejected marks and correctly matching each judges own style, are now used as substitute

values in each of the Missing mark positions and in place of any "Averages" that have been requested. These substitutions are 'boxed' on the Pilots check-sheets to show where they have been made. This final set of marks can now be multiplied by the figure K-factors to build a new table of scores for each pilot by each judge ready for the next step.

4. Identify and settle any High and Low Biased Scores

The FairPlay System now uses the above table of scores as the basis for another Normalisation, Fitted Values and Missing data process very similar to that of the marks assessment procedure. This time however the process is used to detect and resolve any unusual scores that may have survived, the confidence level required here being a slightly more relaxed 90%. Biased scores are possible because even though all unusual raw marks

have been removed a judge may still have given overall an under or over-stated assessment of a competitor, and the score can thus be unacceptably high or low when compared to the other judges. Such bias can for example be the result of over-enthusiastic assessment of a home team pilot, or simply national likes and dislikes that have not been successfully kept in check. FPS as usual replaces any scores that fail their confidence test with the judges Fitted Value score, and again any such changes are clearly shown on the Pilots check-sheets.

5. Remove any possible influence from low scoring Pilots on the leaders

As a last step, it is necessary to ensure that the harder-to-judge lower scoring pilots are not able to influence the ranking of pilots at the head of the table. Pilots who have scored less than 60% in Known's and Free's or 50% in Unknown sequences are now temporarily excluded, and the entire FPS process is run again from the very first step. A results table can now be constructed from these newly calculated higher ranking scores mixed together with the previous scores for the lower scoring pilots. Finally the penalties are subtracted, and the sequence results are ready for publication.

6. Create detailed feedback for the Judges
Now the FairPlay System can turn to it's other
great strength – a thorough review of judging
performance. An individual analysis shows for
each judge how he compares to his colleagues,
while for the Chief Judge the statistics for the
whole panel are collated and ranked to show
which judge most closely matched the panel
view and by how much the other judges were

out of step with all their colleagues. In this way FPS is able to provide a great deal of easily distributed feedback for the entire judging team, something not available until the advent of this system.

Publication of Results

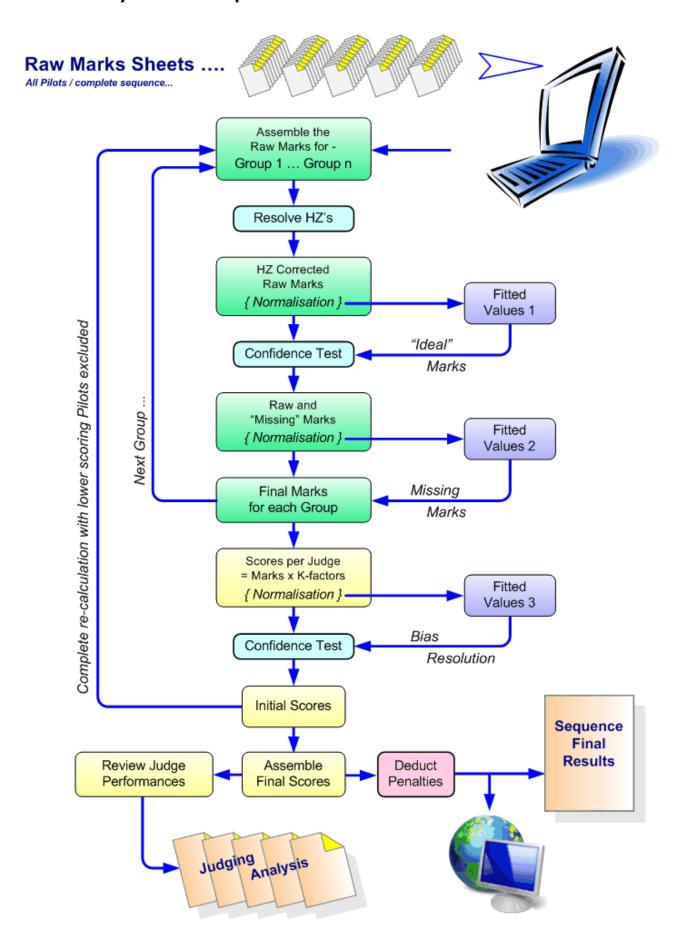
After approval from the Chief Judge and the Jury, the scorer can now publish the results on paper and to the web, and make the Chief and individual Judges sequence analysis available to the panel so the pilots and the judging panel can each see in detail just how they have performed.

The Judges Ranking Index

In an ideal world each judge would rank the pilots in the same order as the final result based upon the views of the whole panel. Whilst minor differences would generally be of little concern, significant mis-ranking of pilots compared to the panel's final conclusion would be a clear indication that a judges views are not shared and so are less likely to be correct. To measure this effect FPS determines each judges own pilot ranking from a specially prepared set of normalised raw scores, taking into account any rejected PZ's for which judges are not penalised, then builds a personal Ranking Index (RI) that will be zero if the judge is perfectly in-tune with the panel but is triggered upwards by each rank and score difference combined. At a major championship an RI value below about 10 for each sequence would indicate pretty good agreement with the published result, numbers above this level giving increasing cause for concern - a review of the judges own analysis would then be the right place to identify just where the discrepancies are being seen.

Beside the obvious advantage arising from the ease with which any judge can now review their contest performance against the published result and see where they most need to target their personal development effort, experience shows that this system can now be used as a reliable and proven basis upon which to base the selection of judges for international championship duty.

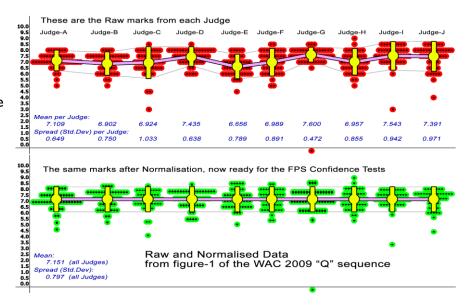
The FairPlay Process map



An example of Raw Marks Normalisation

First diagram:

Each red/black dot represents one mark given by each judge at that value. The yellow circles show the mean for each judge, the vertical yellow strips indicate the spread of the judges marks (this is the 'standard deviation'). The pink and grey lines emphasize the style differences between each judge – some judges give higher marks than others, and some judges spread their marks over a wider range than others.



Second diagram:

During the Normalisation

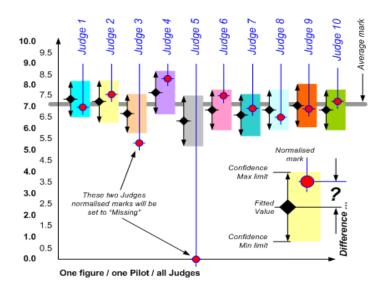
process each judges block of marks has been moved up or down so that their average is equal to the average for the all of the judges, and the spread of each judges marks has been squeezed or expanded to be equal to the average spread for all judges. Because all the judges now have an identical style of marking it is possible to start comparing any judge against the others in a meaningful way.

How does the FairPlay confidence test work?

Taking each normalised mark in turn through the whole group, FPS carries out a statistical test on each one to obtain an 'Uncertainty' valuation for it. This is done by taking the numeric difference between the mark and the 'Fitted Value' that FPS has calculated for it and dividing by the Residual Standard Deviation (SD) for the group. In the upper diagram each judge's mark is shown as a red circle and the Fitted Value as a black diamond. The height of the black arrow indicates the 97.5% confidence range within which we can accept the mark. Any that are above or below this range are too different to the value we should expect the judge to have given, and they can't be used.

If the result of the confidence test exceeds 2.24 then we can say that the uncertainty of the mark is greater than 97.5% and it must be discarded. To understand this look at the idealised distribution of marks shown in the lower diagram. In FPS the marks in the central 97.5% area between the +/- 2.24 Standard Deviation boundaries are accepted as OK, while those in the extreme left/right red areas are the 2.5% that are most different to all the rest and thus are most likely to be unacceptable.

For the rejected marks in the red areas the judges original raw mark is set to "Missing", the blank in the normalised table being replaced in the next step by a new Fitted Value that is now entirely free of any unwanted anomalies.



97.5% FPS confidence area = +/- 2.24SD

1.05%

1.05%

1.05%

1.05%

1.05%

1.05%

1.05%

1.05%

1.05%

1.05%

1.05%

2.5% of all marks are outside the 97.5% FPS confidence limits and are discarded

Decoding the Pilots FairPlay Check-Sheet

Processed Marks sheet Pilots ACRO FairPlay

How to 'de-code' your

nationality and aeroplane details, the level and sequence being flown and The header describes the pilot, the event, place and date

Figure numbers and K-factors for each figure. All marks that FPS determines as "wrong" are shown boxed. This HZ was not matched by a Confirmed Hard Zero from the Chief Judge and must be replaced by a Fitted Value.

For a non-scoring Chief Judge, the gradable figures are labelled "OK" whilst figures given a Confirmed Hard Zero are shown as "CHZ"

figure a "Confirmed Hard Zero" and The Chief Judge has awarded this any non-HZ Judges marks are accordingly boxed

applicable the Harmony mark are shown by their first 4 characters. The Positioning mark and if

sequence have been completed, and the scores given here have been approved by the Jury for Explains that all flights in this inal publication. The ACRO version and FairPlay settings are shown above the time and date of the report.



259.00 45.42 420.00 47.50 257.20 49.91 299.00 306.09 361.44 37.74 251.50 266.64 347.08 45.43 228.16 227.11 125.62 0.00 257.47 221.76 212.08 306.87 43.54 343.86 1-Nick Buckenham (GBR) 2-Francis litier (FRA) 3-Gabor Talabos (HUN) 4-Algis Orlickas (LTU) 5-M. Bezdenezhnykh (RUS) 0.00 6.43 6.64 7.92 0.00 90. 6.50 6.65 6.62 .57 6.51 7.5 7.04 0.9 2918.88 2924.05 3046.74 3203.07 2915.77 3032.50 2921.06 7.0 H 6 - Vladimir Kotelnikov (RUS) 7 - Georges Brocard (SUI) 7.0 8.0 8.0 8.0 7.68 0.0 6.5 7.67 6.51 H 6.5 Joe Bloggs (GBR) Sukhoi 26 G-ABCD Unlimited level - Programme 2: Free Unknown 6.5 7.86 7.56 6.5 6.78 8.0 0.0 6.5 4.07 HZ Processed Marks Check-Sheet - Pilot 005 Nick Buckenham (non-scoring) 17th FAI EAC, Touzim CZE, 02. - 12.Sep.2010 6.51 5.0 6.20 7.69 8.0 8.20 8 - Lyudmila Zelenina (UKR) 6.0 6.5 Sequence has been completed: Sequence scores here are FINAL 5 4 - Algis Orlickas (LTU) 5.00 8.0 3 6R 5.0 7.56 5.00 8.5 2.5 07 6.0 6.0 4.83 7.52 H 6.0 2 Chief Judge: 8.53 7.0 8.0 8.22 7.0 7.0 7.22 8.0 8.06 0.0 7.0 6R 7.0 6.5 6.67 ZΗ 6.5 ΖH 6.5 Judges: 2 CHZ દું š š š š š ð ð š š š š š š š ų, 2 99 38 6 66 7 31 7 22 2 46 7 44.7 40 7 Posi 60 1 6.5 6 7 6 7 200 No. Factor Anomalies: 37 Scores: 9 7 2 3

2995.33 FP-sub:

Final processed score total (of max. possible 5270) Processed score valuation: Minus 300 penalty points:

Aerobatic Contest Results Organiser, Version 3.0 build 28-12-10 Calculations by: FairPlay (non-scoring CJ + CHZ Summary) method This report created at 12:13 on 29 December 2010

50.57

2964.90 2664.90

The Chief Judge (with non-scoring or scoring status) and the Judges details are listed in full

relevant confidence test, and all marks for the =PS has determined for this figure that more igure have been replaced by Fitted Values. than 60% of the Judges marks failed the

In each pair of data rows the upper italic number the Judge, and FPS has given a Fitted Value.

An "Average" mark has been requested by

is the raw mark given by the Judge. The lower bold figure is either -

 the Normalised mark when this satisfies the relevant FPS confidence test, or ...

the Fitted Value, substituted by FPS when the Hi or Lo, or the raw mark is already boxed. normalised mark fails the confidence test

with their Hi or Lo prefix, and the Fitted Value is calculated figure confidence limits are boxed Marks that are above or below the FPS substituted below.

Pilots to quantify the relative value achieved For information purposes only and to assist in each figure, the two right-most columns display the average mark for all Judges together with the "equivalent" score.

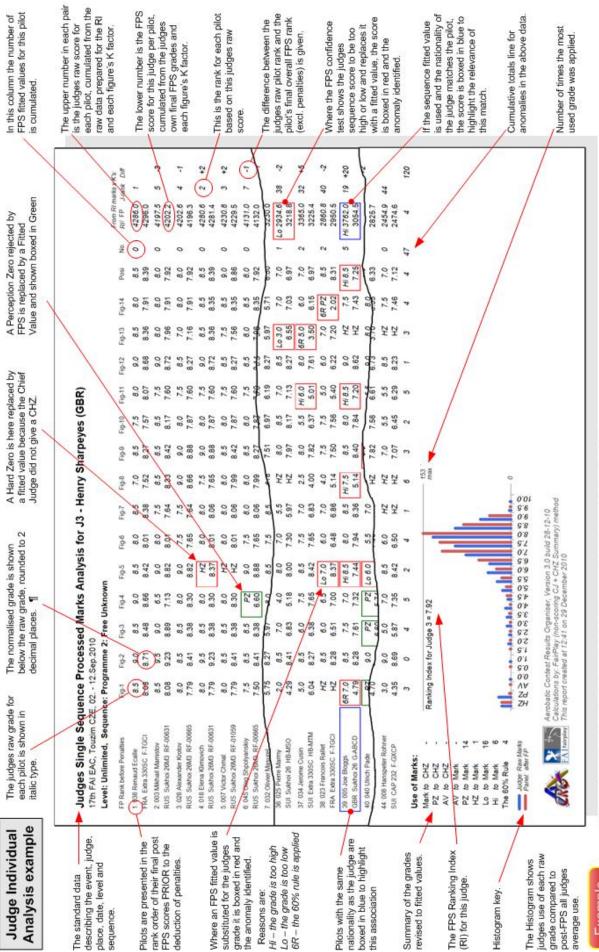
These are NOT formal FPS steps, nor is this data used in any subsequent calculations.

-PS detects a High or Low sequence anomaly The FPS score for each Judge is shown. If the score is annotated and the Fitted Value substitution shown below.

The final FPS processed score is shown, and are deducted. The overall percentage scored any penalty points that have been awarded is shown on the last line.

www.exploit-design.com

Decoding the Judges Individual Analysis Sheet





Note the system works in double-precision mode using 16 significant characters, viz. 123.456789012345

The printed output is normally rounded to two decimal places for clarity

www.exploit-design.com

2011

Decoding the Chief Judges Overall Analysis Sheet

