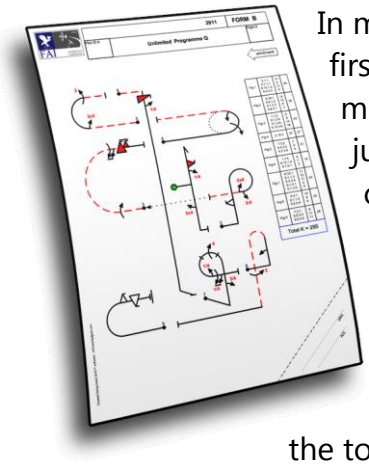
Processing Judges marks and CIVA's FairPlay System (FPS)

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 ask for a suitable "average" mark to be provided by the system; this will be a simple average of the marks from the scoring judges, to the nearest half mark.

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 way-out 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 not 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. It was revised in 2018 to include proportional assessment of any unusual marks, to smooth the effect of repeated results calculations on individual pilots' rankings.

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 statistician's word for this balancing act is 'normalisation', and without it comparisons between the judges would simply not be valid. In our normalisation each judge's 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.

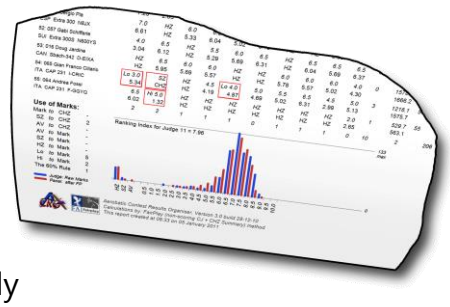
3. **Identify and resolve "Unusual" Marks**
   For each group of marks FPS calculates an idealised table of 'Fitted Value' (FV) marks that is matched to each judges own style. A statistical confidence test at 98.5% is now carried out to check the validity of each normalised mark against its corresponding FV. 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 noted for further


Processed Marks Check-Sheet - Pilot 013
Nick Onn (GBR) Sukhoi 26M G-XXVI
Unlimited - Power level - Programme 1
FAI 25th WAC 2009, Silverstone, Northamptonshire, 20th -20
Chief Judge: Graham Hill (non-scoring)
Judges: 1 - Graham Hill (GBR) 2 - Vladimir Kotelnik
4 - Francis Iller (FRA) 5 - Quintin Hawthorn
7 - Tomas Korinek (CZE) 8 - Hannes Gra
11 - Lyudmila Zelenina (UKR)

Fig No.	K Factor	SF	CJ1	J2	J3	J4	J5	J6
1	81	7	OK	7.0	7.0	6.5	6.0	7.0
				6.59	7.22	6.79	6.41	6.59
2	49	5	OK	5.5	6.0	Lo 3.0	4.0	5.5
				5.30	6.05	5.86		5.30
3	32	7	OK	6.5	7.0	8.0		6.5
				6.45	7.33	8.09		6.45
4	63	7	OK	6.0	6.0	5.0		6.0
				6.01	6.01	5.77		6.01
5	50	7	CHZ	HZ	7.0	6.0		HZ
				0.00	0.00	0.00		0.00

treatment. When this initial group processing is complete, if any mark has been set to Missing then the normalisation procedure is run again excluding the raw marks identified as Missing, and a new set of Fitted Values calculated from the very beginning. These new FV's will correctly match each judge's style but are not influenced by the raw marks that were set to Missing. Each Missing mark is now replaced either by a proportionately blended value part-way between the renormalised mark and the new FV, or if its confidence is below 95.0% the FV is adopted unchanged. These substitutions are shown 'boxed' on the Pilots check-sheets to indicate 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 levels required are now slightly more relaxed at 78.5% and 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. **Penalty handling**
   After the of the marks and scores processing has been completed for all groups the penalties are subtracted, and the sequence results are now ready for publication.

6. **Create detailed feedback for the Judges**
   Now the FairPlay System can turn to its 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

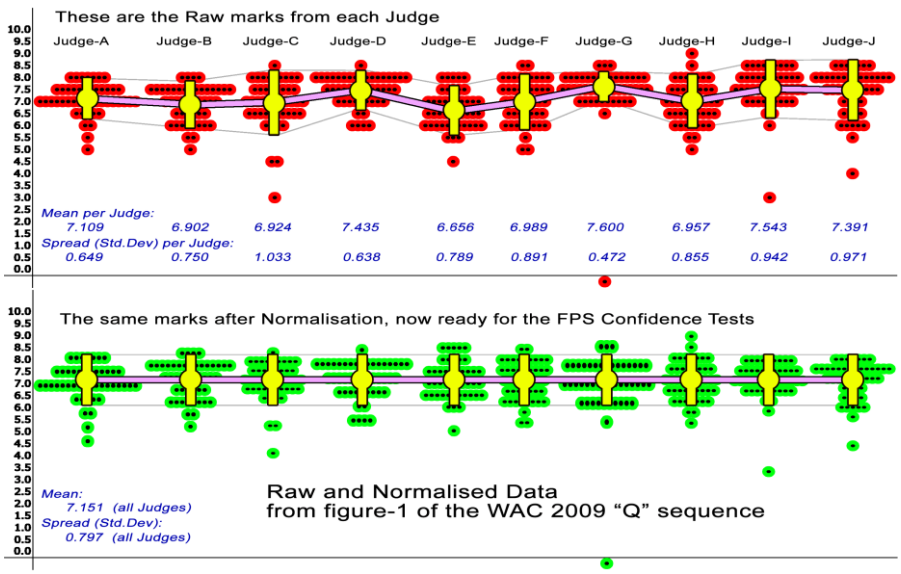
In an ideal world each judge would rank the pilots in the same order as the final result based upon the views of the whole panel. Whilst minor differences would generally be of little concern, significant mis-ranking of pilots compared to the panel's final conclusion would be a clear indication that a judge's 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.

An example of Raw Marks Normalisation

First diagram:

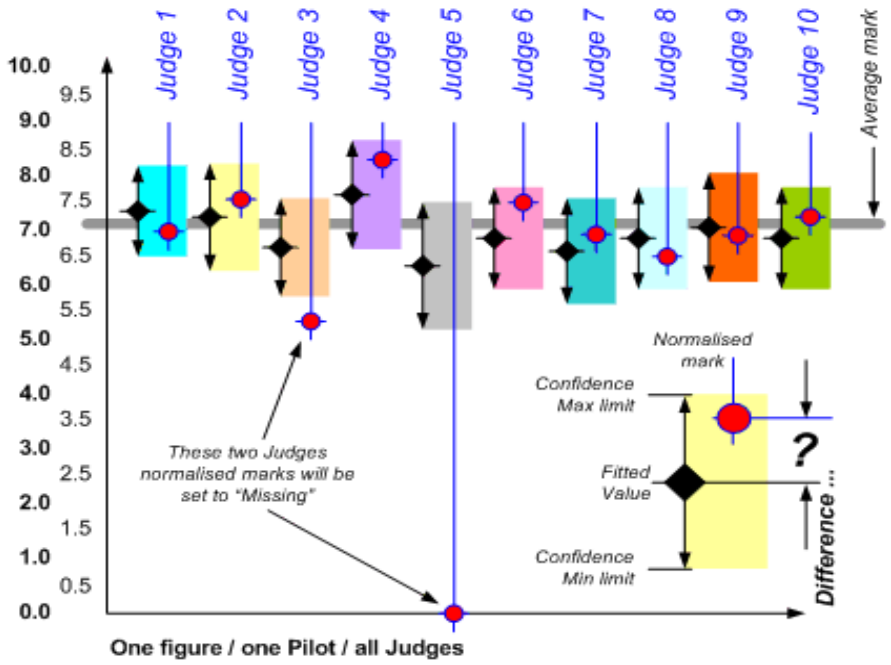
Each red/black dot represents one mark given by each judge at that value. The yellow circles show the mean for each judge, the vertical yellow strips indicate the spread of the judges marks (this is the 'standard deviation'). The pink and grey lines emphasize the style differences between each judge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



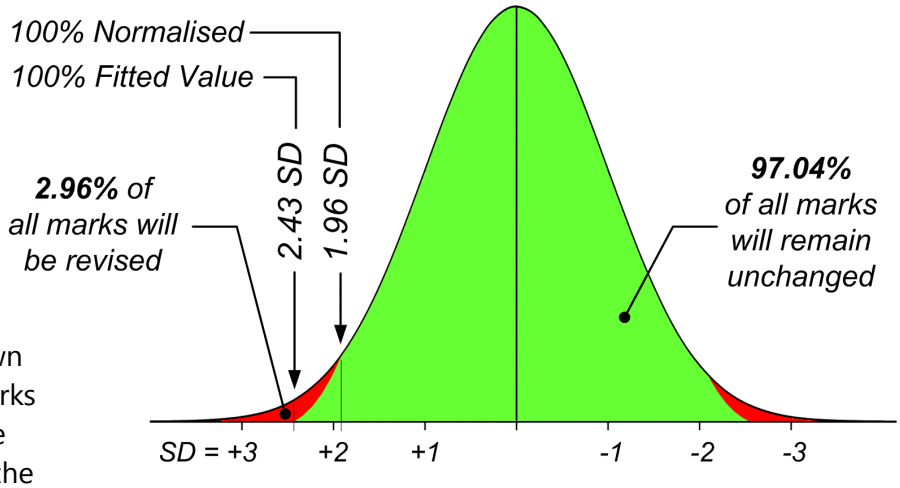
Second diagram:

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 (FV) 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 98.5% (2.43 SD) confidence range within which the mark can be accepted. Any that are outside this range are too different to the value we should expect the judge give, and they must be adjusted.

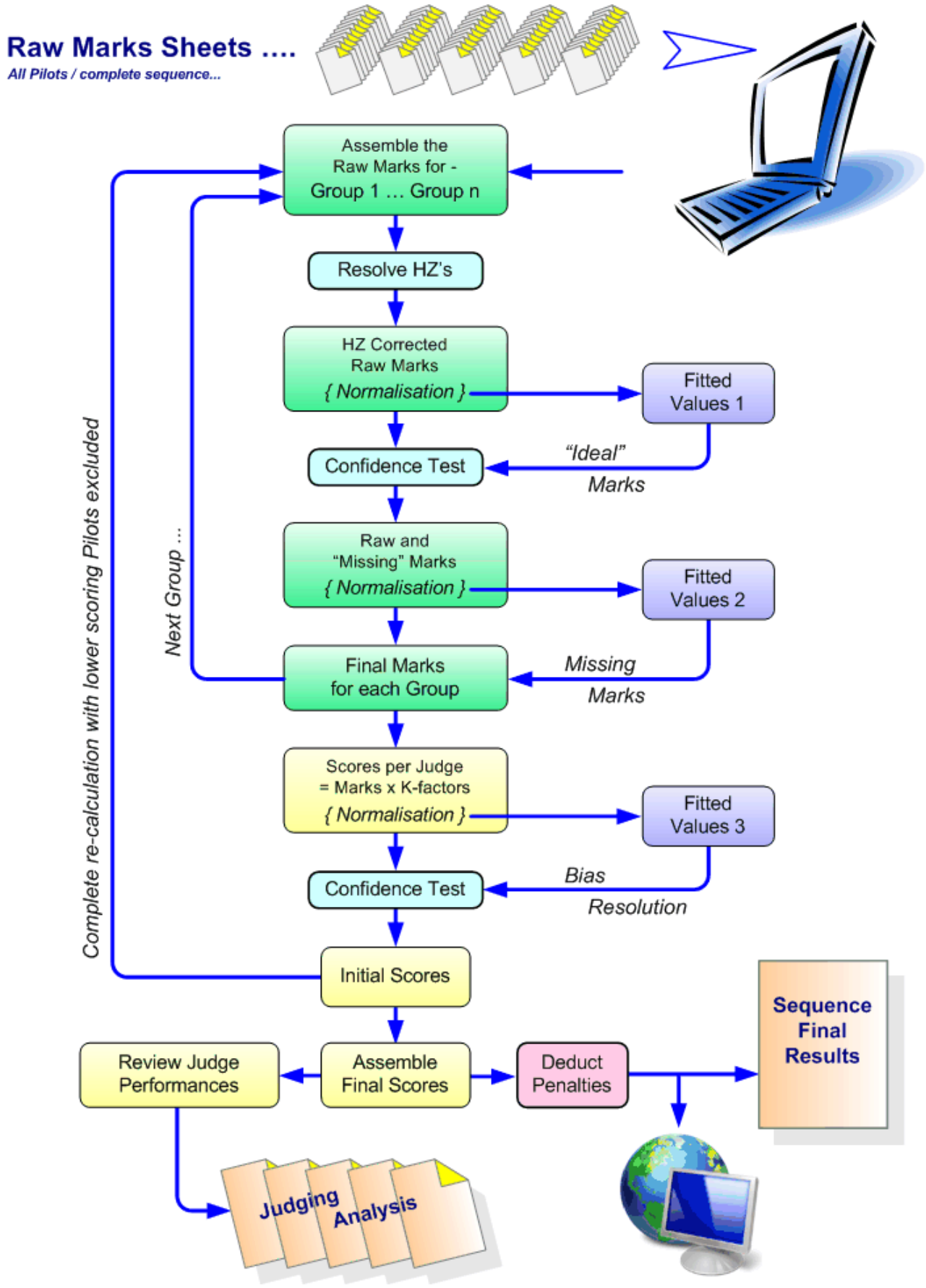


If the result of the confidence test is between 98.5% / 2.43 SD and 95% / 1.96 SD the mark is proportionately blended between its normalised value and the FV. Any that are beyond 95% are simply replaced by the FV.

To understand this look at the idealised distribution of marks shown in the lower diagram. In FPS the marks in the central 97.04% green area are accepted without change, those in the left/right red areas between 1.96 and 2.43 SD are blended proportionately from their normalised value to the FV, while those with SD above 2.43 are directly replaced by the FV.

In response to feedback from pilots over many years regarding the extent of individual pilot rank changes that are an inevitable feature of statistical systems when results calculations are repeated as the number of pilots marks entered gradually increases, the FairPlay System was thoroughly reviewed and developed for the 2018 competition season to incorporate the above proportionate blending process. In practice this mimics the subjective methodology that humans apply to these situations as confidence in a comparison slides from high to low, and the degree of minor rank changes has now been reduced by about 50%.

The FairPlay Process map



How to 'de-code' your

Pilots ACRO FairPlay Processed Marks sheet

The header describes the pilot, nationality and aeroplane details, the level and sequence being flown and 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".

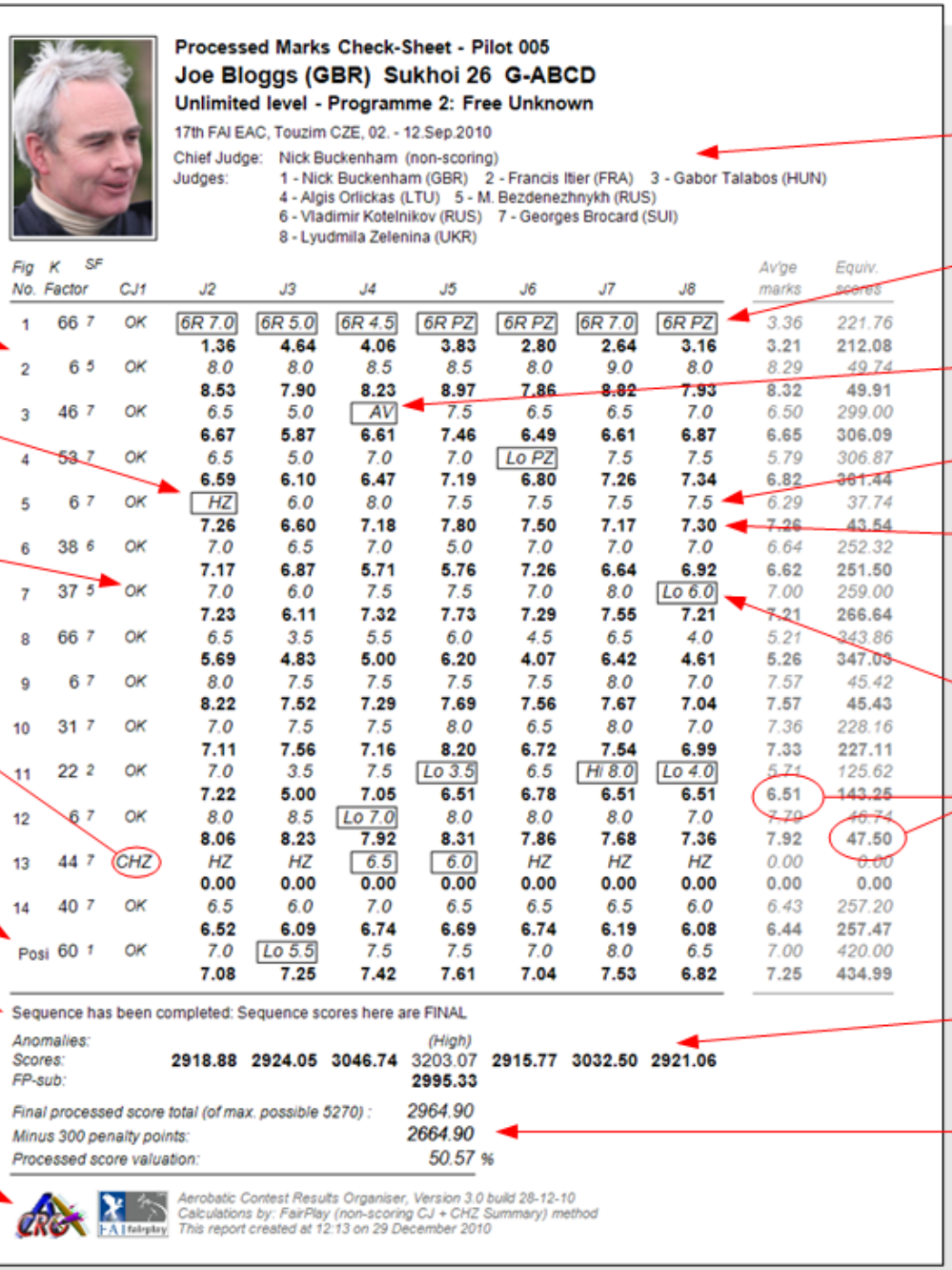
The Chief Judge has awarded this figure a "Confirmed Hard Zero" and any non-HZ Judges marks are accordingly boxed.

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

Explains that all flights in this sequence have been completed, and the scores given here have been approved by the Jury for final publication.

The ACRO version and FairPlay settings are shown above the time and date of the report.

Example ONLY !



Processed Marks Check-Sheet - Pilot 005

Joe Bloggs (GBR) Sukhoi 26 G-ABCD

Unlimited level - Programme 2: Free Unknown

17th FAI EAC, Touzim CZE, 02. - 12.Sep.2010

Chief Judge: Nick Buckenham (non-scoring)

Judges: 1 - Nick Buckenham (GBR) 2 - Francis Itier (FRA) 3 - Gabor Talabos (HUN)
4 - Algis Orlickas (LTU) 5 - M. Bezdenezhnykh (RUS)
6 - Vladimir Kotelnikov (RUS) 7 - Georges Brocard (SUI)
8 - Lyudmila Zelenina (UKR)

Fig No.	K Factor	SF	CJ1	J2	J3	J4	J5	J6	J7	J8	Av'ge marks	Equiv. scores
1	66	7	OK	6R 7.0	6R 5.0	6R 4.5	6R PZ	6R PZ	6R 7.0	6R PZ	3.36	221.76
				1.36	**4.64**	**4.06**	**3.83**	**2.80**	**2.64**	**3.16**	3.21	212.08
2	6	5	OK	8.0	8.0	8.5	8.5	8.0	9.0	8.0	8.29	49.74
				8.53	**7.90**	**8.23**	**8.97**	**7.86**	**8.82**	**7.93**	8.32	49.91
3	46	7	OK	6.5	5.0	AV	7.5	6.5	6.5	7.0	6.50	299.00
				6.67	**5.87**	**6.61**	**7.46**	**6.49**	**6.61**	**6.87**	6.65	306.09
4	53	7	OK	6.5	5.0	7.0	7.0	Lo PZ	7.5	7.5	5.79	306.87
				6.59	**6.10**	**6.47**	**7.19**	**6.80**	**7.26**	**7.34**	6.82	361.44
5	6	7	OK	HZ	6.0	8.0	7.5	7.5	7.5	7.5	6.29	37.74
				7.26	**6.60**	**7.18**	**7.80**	**7.50**	**7.17**	**7.30**	7.26	43.54
6	38	6	OK	7.0	6.5	7.0	5.0	7.0	7.0	7.0	6.64	252.32
				7.17	**6.87**	**5.71**	**5.76**	**7.26**	**6.64**	**6.92**	6.62	251.50
7	37	5	OK	7.0	6.0	7.5	7.5	7.0	8.0	Lo 6.0	7.00	259.00
				7.23	**6.11**	**7.32**	**7.73**	**7.29**	**7.55**	**7.21**	7.21	266.64
8	66	7	OK	6.5	3.5	5.5	6.0	4.5	6.5	4.0	5.21	343.86
				5.69	**4.83**	**5.00**	**6.20**	**4.07**	**6.42**	**4.61**	5.26	347.03
9	6	7	OK	8.0	7.5	7.5	7.5	7.5	8.0	7.0	7.57	45.42
				8.22	**7.52**	**7.29**	**7.69**	**7.56**	**7.67**	**7.04**	7.57	45.43
10	31	7	OK	7.0	7.5	7.5	8.0	6.5	8.0	7.0	7.36	228.16
				7.11	**7.56**	**7.16**	**8.20**	**6.72**	**7.54**	**6.99**	7.33	227.11
11	22	2	OK	7.0	3.5	7.5	Lo 3.5	6.5	Hi 8.0	Lo 4.0	5.71	125.62
				7.22	**5.00**	**7.05**	**6.51**	**6.78**	**6.51**	**6.51**	6.51	143.25
12	6	7	OK	8.0	8.5	Lo 7.0	8.0	8.0	8.0	7.0	7.79	46.74
				8.06	**8.23**	**7.92**	**8.31**	**7.86**	**7.68**	**7.36**	7.92	47.50
13	44	7	CHZ	HZ	HZ	6.5	6.0	HZ	HZ	HZ	0.00	0.00
				0.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	0.00	0.00
14	40	7	OK	6.5	6.0	7.0	6.5	6.5	6.5	6.0	6.43	257.20
				6.52	**6.09**	**6.74**	**6.69**	**6.74**	**6.19**	**6.08**	6.44	257.47
Posi	60	1	OK	7.0	Lo 5.5	7.5	7.5	7.0	8.0	6.5	7.00	420.00
				7.08	**7.25**	**7.42**	**7.61**	**7.04**	**7.53**	**6.82**	7.25	434.99

Sequence has been completed: Sequence scores here are FINAL

Anomalies:				*(High)*			
Scores:	**2918.88**	**2924.05**	**3046.74**	3203.07	**2915.77**	**3032.50**	**2921.06**
FP-sub:				**2995.33**			

Final processed score total (of max. possible 5270) : *2964.90*
Minus 300 penalty points: *2664.90*
Processed score valuation: *50.57 %*

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

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

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

An "Average" mark has been requested by the Judge, and FPS has given a Fitted Value.

In each pair of data rows the upper italic number 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 normalised mark fails the confidence test Hi or Lo, or the raw mark is already boxed.

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

For information purposes only and to assist Pilots to quantify the relative value achieved 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.

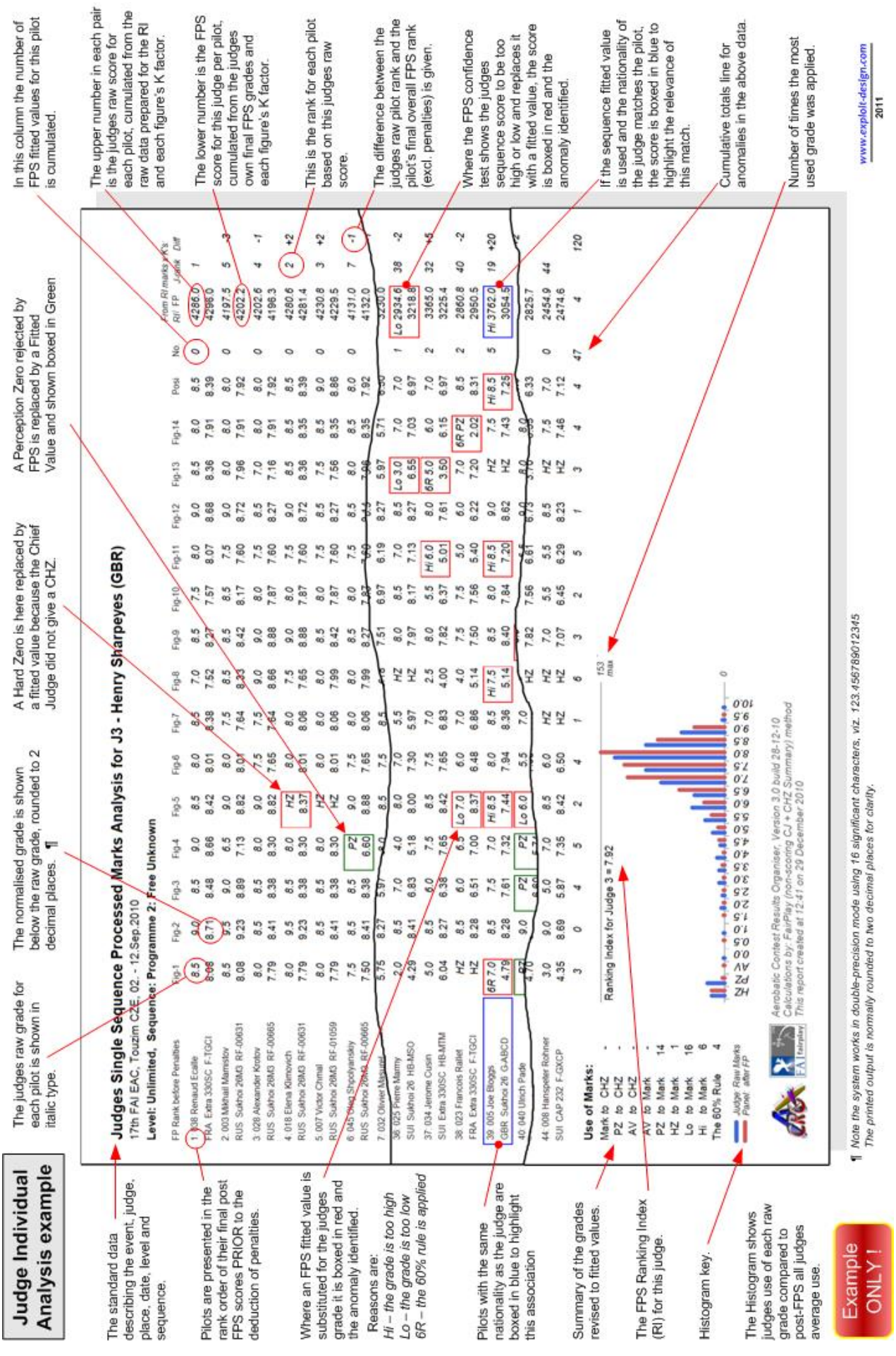
The FPS score for each Judge is shown. If FPS detects a High or Low sequence anomaly the score is annotated and the Fitted Value substitution shown below.

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

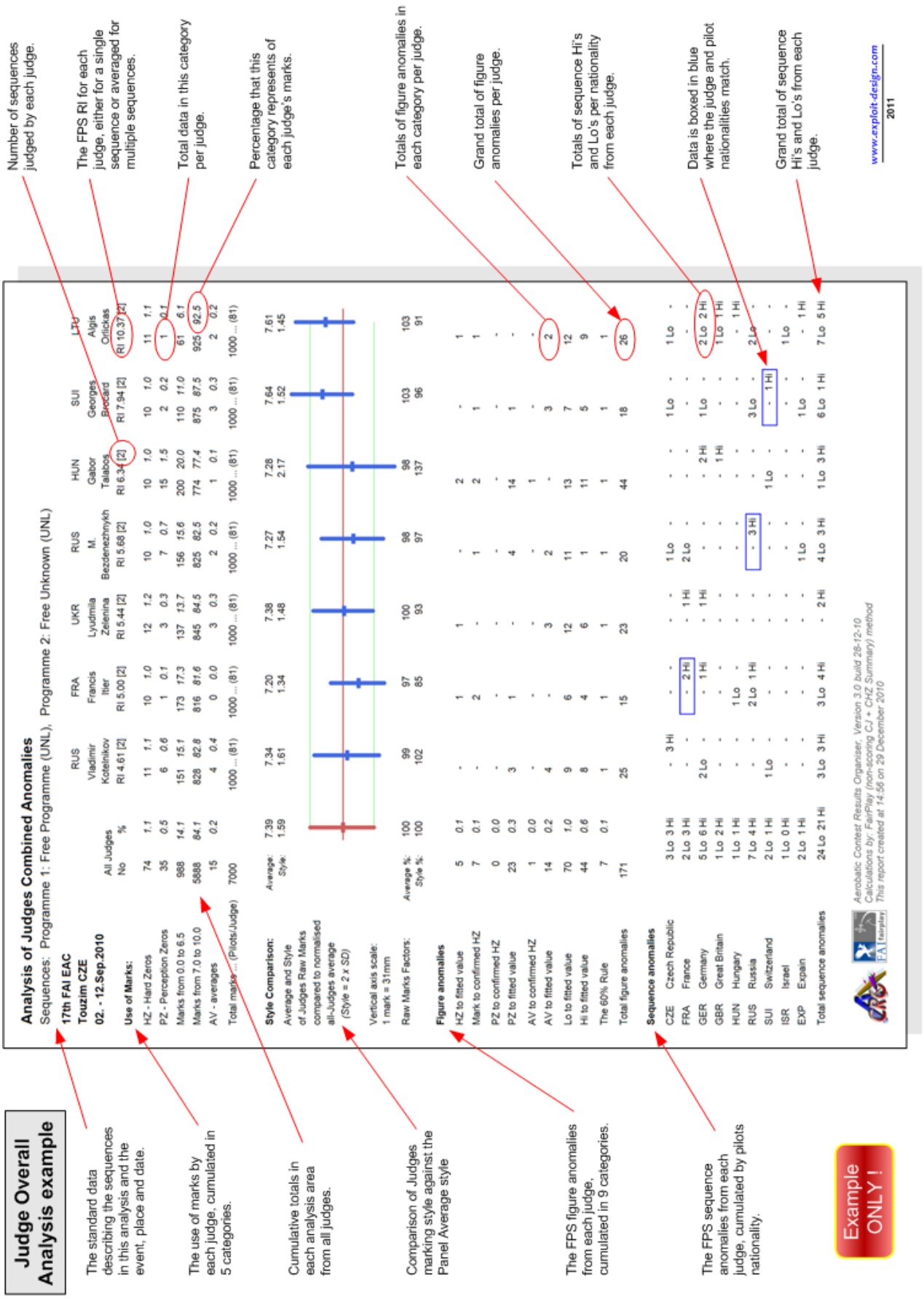
www.exploit-design.com
2011

Decoding the Pilots FairPlay Check-Sheet

Decoding the Judges Individual Analysis Sheet



Decoding the Chief Judges Overall Analysis Sheet



NHB – November 2019