Report To CIVA On The New Statistical Analysis Programme

THE FAIR PLAY SYSTEM

by Alan Cassidy, UK CIVA Delegate

1. Background

1.1. Following criticism of the current TBLP system in the USA and elsewhere, the bureau of CIVA decided in Autumn 2004 to commission a review of the statistical analysis used in the processing of the results of its international aerobatic competition.

1.2. Dr. Derek Pike, formerly Head of the School of Applied Statistics at the University of Reading in the United Kingdom was appointed to carry out the review. Alan Cassidy, the UK CIVA Delegate was appointed to liaise with Dr. Pike throughout the review and to act as project manager for the review on behalf of CIVA.

1.3. Dr. Stephen Green, a qualified international judge from the UK acted in an advisory capacity on behalf of CIVA, and another professional statistician, Dr. Christopher Crocker was contracted to undertake a peer review of Dr. Pike's report. Dr. Crocker is the Principal Statistician at Mathematical Market Research Ltd, Wallingford, UK.

1.4. Following the peer review by Dr. Crocker, he and Dr. Pike worked together to finalise agreement on the new proposed system.

2. Summary

2.1. A new statistical methodology for analysing the results of international aerobatic competitions is reported.

2.2. This new method adopts internationally-recognised statistical principles in the assessment and manipulation of raw contest data.

2.3. The new system includes a novel method of grouping contest grades in order to ensure that the statistical processes applied are valid for both Free and Compulsory Programmes. A by-product of this method is that it remains robust and conservative as the numbers of pilots and judges is reduced.

2.4. Whilst the recommended minimum number of judges for international events remains seven (7), the system will continue to work technically for fewer judges. However, the process opens itself to collusion between judges and the normalisation process can give rise to larger changes in the raw scores – particularly where judges are different in the variability of their scoring methods.

2.5. The new system carries out a normalisation process of the judges' grades for each figure flown. This ensures that the opinions of all judges potentially carry the same importance. This part of the process is valid even if there are only two (2) judges in the panel.

2.6. Following this normalisation, any individual grades that exceed a nominal threshold of uncertainty are removed from the data set and replaced by fitted values based on the relevant pilot, figure and judge characteristics from the data set. This process helps to eliminate the risk of partial or inaccurate judging.

2.7. After completing this analysis at the figure level, a similar analysis is performed on the overall sequence totals derived for each pilot from each judge. This part of the process is designed to detect repeated small degrees of partiality, by a judge.

2.8. At appropriate stages of the process, reports can be printed showing a pilot his grades. During the process, data can also be extracted to facilitate further detailed analysis of judges' performance in accordance with CIVA Regulations.

3. Grouping of Figures for Analysis

3.1. In order that the statistical analysis performed on judges' grades is valid, the data must be grouped in such a way that the following criteria are met:

3.1.1. A data group contains enough points and has enough degrees of freedom[1] to permit valid calculations.

3.1.2. A data group should, to the maximum extent possible, contain data points arising from closely-related activities.

3.2. For compulsory programmes, both of the above criteria can most easily be met by considering as a coherent data set all the grades given to all the pilots for a particular figure. Grades for positioning can be treated as just another 'figure'. Thus, a data set would look like this:

Figure	Pilot	Judge 1	Judge 2	…	Judge j
f	1	Grade(f,1,1)	Grade(f,1,2)	…	Grade(f,1,j)
f	2	Grade(f,2,1)	…	…	…
f	…	…	…	…	…
f	…	…	…	…	…
f	p	Grade(f,p,1)	…	…	.Grade(f,p,j)

3.3. For free programmes, pilots may fly sequences with different numbers of figures. Also, figure 1 flown by pilot 1 may be very different in terms of its difficulty from figure 1 flown by pilot 2. Hence it is not appropriate to group Free Programme figures solely by figure number. The proposal is to list the various figures by increasing K-factor and form groups having their number of rows equal to the number of pilots. Thus a Free Programme data set could look like this:

K factor	Figure	Pilot	Judge 1	Judge 2	…	Judge j
k	f1	p1				
k + n	f4	p6				
k + n + m	f2	p14				
…	…	…				
k(max)	f3	p2				

3.4. In both the cases above, Figures compared in any particular group have identical or similar K-factors.

---

[1] [In an application, where we have for example a compulsory programme with *p* pilots and *j* judges, we can define "degrees of freedom" as (p-1)*(j-1) and we would ideally like this value not to be less than about 20].

3.5. Whilst this process is primarily designed for international contests with numerous pilots and judges, it can also be applied at domestic regional contests where the numbers of pilots and judges may be reduced. The proposed method of grouping will still produce valid data groups in such circumstances as long as the number of judges is not less than three (3). This desirable result is achieved by setting a minimum number of rows per data group at eleven (11).

3.6. For example, consider a Primary level contest with 2 pilots each flying 5 figures and observed by 3 judges. The data group table could look like this (figure 0 is positioning):

K factor	Figure	Pilot	Judge 1	Judge 2	Judge 3
3	0	1			
3	0	2			
3	5	1			
3	5	2			
4	3	1			
4	3	2			
10	2	1			
10	2	2			
10	4	1			
10	4	2			
15	1	1			
15	1	2			

3.7. The data set above has 22 degrees of freedom (12-1)*(3-1), which is acceptable for the normalisation process, subject to the qualification that the range of K-factors is not large compared with the possible range of K-factors. It will be unusual for any individual grade to show as anomalous with only three judges, but it would do so, for instance, if a single judge gave a soft zero grade to a figure graded highly by the other two judges. Of course, collusion between two judges to distort the results cannot be detected with such a small panel.

4. Pre-Processing of HZ and A Grades

4.1. On their grade sheets, judges may mark HZ or A to figures that they believe to be wrongly flown or which they did not observe. In the former case, the Chief Judge must determine whether a Hard Zero is confirmed. When HZ is not confirmed, the HZ grades will be designated as “Missing” and treated in the same way as the A grades.

4.2. CIVA Regulations currently call for unconfirmed HZ grades and all A grades to be changed to the average of the other judges’ Raw Grades. Such a plain average, however, is not the most appropriate replacement, as it takes no account of the style of the judge whose missing grade is to be replaced. For example, if a judge who is normally relatively high-scoring fails to observe a figure, the average of other, lower-scoring, judges will not do the pilot proper justice.

4.3. It is an important part of the new system that such replacement values are properly ‘fitted’ to the data set. Therefore such grades must not be corrected on the judging line but must be input as originally written. Similarly, when a Hard Zero is confirmed, the setting to HZ of all relevant numerical grades must be done inside the computer at the pre-processing stage. When numerical grades are set to HZ, the software must increment the total of Hard Zero errors for the judge concerned for use in the Hard Zero Index (HZI) calculation element of the Judges Performance Index (JPI).

4.4. Some minor changes will be required to CIVA Regulations to describe the handling of HZ and A grades at the judging line.

4.5. At this stage it is normal practice to print each pilot’s score sheet for checking. This should still be done, but it must be appreciated that, at this stage, HZ and A grades that are to be replaced will be left unchanged.

4.6. It is also possible at this stage to derive and publish histograms illustrating the number of each possible numerical grade awarded by each judge over the whole programme. These illustrations are a useful aid in providing feedback to the judges on their individual styles. Currently, these data are used to derive a Discrimination Index (DI) for each judge, to be included in the overall JPI.

4.6.1. Following extensive simulation of possible judging styles, I have concluded that this element of the judges performance is not suitable for inclusion in an overall assessment of competence. This is because the ideal range of scores is exactly the same as the ‘actual’ range of figures flown. The range of grades applied by a judge should be appropriate to the figure flown, and not simply cover a wide range of numbers.

4.6.2. A by-product of this process review, therefore, is to recommend that CIVA Regulations be amended to remove the DI from inclusion in the overall JPI.

5. Figure Grade Normalisation

5.1. The grades in each data group must be normalised to ensure that each judge has the same chance to decide the outcome. Without normalisation, it is possible for a judge who grades figures very differently to overwhelm those who grade over a smaller range. During the normalisation process, Raw grades of HZ and A will be excluded from the calculation of any Means and Standard Deviations.

5.2. The normalisation process aims purely to modify the grades in such a way that all judges show an identical level of variability for each data set. There is no need to change the individual judges grades so that all judges have the same average.

5.3. Normalisation relies on comparing the Standard Deviation of each single judge in the data set with the overall Standard Deviation of all judges in the data set. This involves the derivation of a Standard Deviation Ratio (SDR). Two methods are possible for this, and are discussed in the reports by the two professional statisticians. The method finally adopted, setting the SDR using the mean of the judges Standard Deviations, has been selected because it is more robust and conservative as the number of judges is reduced.

6. Determination of Fitted Values

6.1. Where HZ or A grades have to be replaced by numerical values, these should be determined by an analysis of the data group which takes full account of the overall scores of both the pilots and the judges and not just by simple averaging of the other judges’ grades for the given pilot. The proposed method is a pragmatic, but statistically sound, approach which can be easily understood and simply implemented in practice.

7. Anomalous Figure Grades

7.1. To be fair to all pilots, it is wise to examine the normalised grades to see if any judge has been unduly partial, or probably erroneous, in his assessment of a figure. The new system employs a two-way analysis of variance over judges and pilots to achieve this. The method compares the normalised values with the fitted values, using the technique of Standardised Residuals to determine the degree of reliability of each normalised grade.

7.2. Any such grades found to be anomalous, at a predefined confidence threshold, must be identified and subsequently treated as though they were missing from the data set.

8. The Threshold for Anomalies

8.1. The threshold set for determining the anomalous nature of a grade is subjective and can be varied. CIVA currently adopts a cut-off at 5% confidence (2*RMS) and the new system will initially conform to this historic policy.

8.2. Any software implementation of the methodology, however, can be written such that the threshold can be reset with a single variable. In time, and after analysis of the operation of the system over a number of competitions, it might be judged advisable to change the datum. The most likely cause for such a change would be the increase in accuracy and consistency of the judging.

8.3. In domestic competition, when the number of judges is lower than the CIVA minimum, a more conservative anomaly threshold might be more appropriate. With 4 or 5 judges, the threshold might be set at 2½% (2.3*RMS); with 2 or 3 judges at 1¼% (2.5*RMS). With these lower thresholds, more data points would be conserved but a partial judge would have a greater chance of influencing the outcome.

9. Replacement of Unconfirmed HZ, A and Anomalous Figure Grades

9.1. Figure grades determined to be anomalous at the agreed threshold are set to ‘Missing’. At this point, the actual number of valid data points in a group may be slightly reduced and therefore it is appropriate to re-normalise the remaining numerical grades and determine a second set of fitted values.

9.2. The final figure grades are created by using these second normalised grades with the unconfirmed HZ, A and Missing grades replaced by the new fitted values.

9.3. Whenever an anomalous grade or HZ is replaced, this should be registered by the software as an increment to the sum of Low Score, High Score or Hard Zero errors attributable to the judge concerned.

10. Re-Constitution of Grades by Pilot

10.1. During the figure analysis, data has been grouped on the basis of each individual figure (compulsory programmes) or by associating K-factors (free programmes).

10.2. Following the figure analysis, the data must be re-sorted into a figure/judge matrix for each pilot. At this point, a further score sheet should be printed for each pilot showing how the grades have changed through the normalisation and replacement processes.

11. Analysis of Sequence Total Scores

11.1. After the completion of the figure analysis, sequence totals by each judge for each pilot must be calculated. This will produce a single data matrix as follows:

Pilot	Judge 1	Judge 2	Judge 3	…	Judge j
1	Total(1,1)	Total(1,2)	…	…	Total(1,j)
2	Total(2,1)	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
p	Total(p,1)	…	…	…	Total(p,j)

11.2. Provided that the number of pilots exceeds ten (10), it is appropriate to make a further similar analysis of the sequence totals to assess the reliability of each judges total for each pilot. This analysis is carried out to detect repeated small degrees of partiality, by a judge.

11.3. Firstly the sequence totals are normalised as before and then a similar two-way analysis of variance is carried out to seek anomalous scores. Again, an arbitrary anomaly threshold must be established and the new system will initially adopt a 5% confidence level as is currently the CIVA tradition.

11.4. After this analysis final sequence totals will have been derived using the normalised totals with any anomalous totals replaced by fitted values. These final sequence totals will the results of the programme, subject to the subtraction of penalty points as per current CIVA Regulations.

11.5. Once the final results (before penalties) have been determined, it is possible to compare each judges rank for each pilot (before figure processing) with the overall rank (after all processing) for each pilot to determine each judge's Ranking Index.

11.6. If a judge has a pilot total replaced during this second analysis, it should be recorded by the processing software. Such anomalies should give rise to a Sequence Anomaly Index for each judge, to be included in the Judges' Performance Index. I recommend that CIVA Regulations be amended to include such a Sequence Anomaly Index in the JPI system and that it should be included mathematically in the overall computation of the JPI.

12. Software Implementation

12.1. As part of the management of the review, I have produced a number of logic flow diagrams which should form the basis of any software implementation of this analytical process. These should be published by CIVA to encourage conformity of any third party software implementation that might occur in individual countries.

12.2. If the necessary changes can be made to the existing CIVA scoring software, and tested, in time for the WAC 2005 in Burgos, then I recommend that this should be done. This would give early warning to all affiliated nations of the methods used and would provide actual data for detailed analysis and further review before the 2005 CIVA plenary meeting in November.

13. Further Implications for Judges Performance Indices

13.1. Once sufficient actual contest data has been processed, typical and extreme values of the various components of the JPI can be considered. These may then be weighted and combined on a summation basis. This will give better discrimination between good and less good judging than the current system of aggregating rankings for each separate index.

14. Recommendations

14.1. I recommend that:

14.1.1. The statistical process developed by Dr. Pike be adopted by CIVA as the new standard for FAI aerobatic contests.

14.1.2. The current minimum of seven judges for CIVA contests be retained.

14.1.3. The initial anomaly threshold be set at 95% (2*RMS) pending future system evaluation based on actual contest results.

14.1.4. The Discrimination Index be removed from the aggregated JPI.

14.1.5. A Sequence Anomaly Index (SAI) be included in the aggregated JPI.

14.1.6. This new system should be used at WAC 2005 if it can be implemented and tested in good time. Draft amendments to CIVA Regulations, including Appendix 2, should precede this use.

14.1.7. The system's adoption should be ratified by the CIVA plenary meeting in November 2005.

14.1.8. In due course, a weighted summation of JPI elements replace the current ranking aggregation.

Alan Cassidy
Maidenhead
England

29 March 2005