An Introduction to the Tarasov-Bauer-Long Process for Analysing Aerobatic Scores

NOTE: This article originally appeared in the British Aerobatic Association Newsletter and it is reprinted here since it is an excellent explanation of the Tarasov/Bauer/Long contest scoring system, a system that has been controversial, but which has served aerobatics well and faithfully since the late 1970's. IAC has been using TBL for a number of years and it has been in use at the World Championships since 1980. The British Aerobatic Association (BAeA), referred to in the article, had only recently begun using the system as their contest participation had grown enough to justify it. Thus, the article was intended for a British audience but has information useful to everyone involved in competition aerobatics. The author has given permission for the article to be used.

By Nick Buckenham, BAeA Judging Administrator

Introduction

A significant change for 1992 is the incorporation of the CIVA approved TBL process into the BAeA contest results software. In operation, the process applies proven statistical probability theory to the Judges' scores to resolve style differences and bias, and to avoid the inclusion of potentially faulty judgements in contest results. To understand just why we need TBL and how it works is of considerable importance to us all – for pilots because it is there to reduce the prospect of unsatisfactory judgements affecting your results, and for judges because not only will it introduce a completely new dimension of scrutiny into the sequence totals you work so hard to produce but it will also discreetly engage the attention of your friendly Chief Judge if your conclusion differs sufficiently from all those other folk on the panel.

Why do we need all this extra complication?

When people get together to judge how well a pre-defined competitive task is being tackled, the range of opinions expressed is often diverse (Figure 1). This is an entirely natural situation amongst humans, where the critique of any display of skill relies on the interpretation of rapidly changing visual cues.

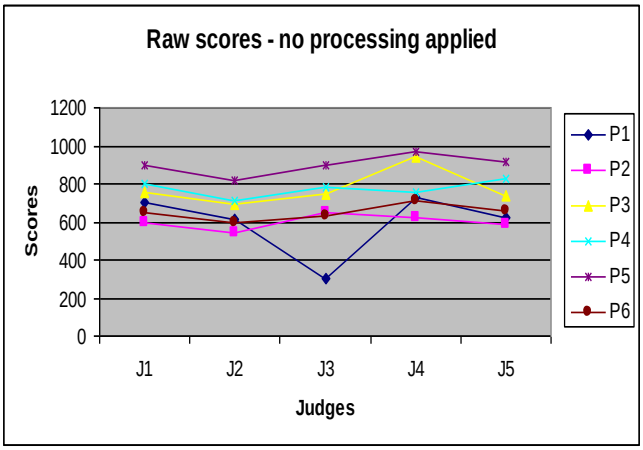

Figure 1: Raw scores – no processing required

In order to minimise the prospect of any way-out opinions having too much effect on the result, it is usual to average the accumulated scores to arrive at a final assessment which takes everybody's opinion into account.

Unfortunately this averaging approach can achieve the opposite of what we really want, which is to identify, and where needed, remove those "way-out opinions" because – if only we could see it – they are the ones most likely to be ill-judged and therefore should be discarded, leaving the rest to determine the more appropriate result.

In aerobatics the process of judging according to the rule-book normally leads to a series of generally similar personal views. However, one Judge's downgrading may be harsher or more lenient than the next (think of this as "style"), his personal feelings toward each competitor or aircraft type may predispose toward favour or dislike (this will lead to "bias"), and he will almost certainly miss or see things that other Judges do not. How then can we judge the Judges and so reach a conclusion which has a good probability of acceptance by all the concerned parties?

The key word is probability – the concept of a perceived level of confidence in collectively viewed judgements has entered the frame.

What we really mean is that we must be confident that opinions pitched outside some pre-defined level of reasonable acceptability will be identified as such and will not be used. This sort of situation is the daily bread and butter of well established probability theory which, when suitably applied, can produce a very clear-cut analysis of numerically expressed opinions provided that the appropriate criteria have been carefully established beforehand.

So how do we best approach the problem?

What has been developed through several previous editions is some arithmetic which addresses the Judge's raw scores in such a way that any which are probably unfair are discarded with an established level of confidence. To understand the process you need only accept some quite simple arithmetic procedures which are central to what is called "statistical probability".

The TBL system in effect does the following:

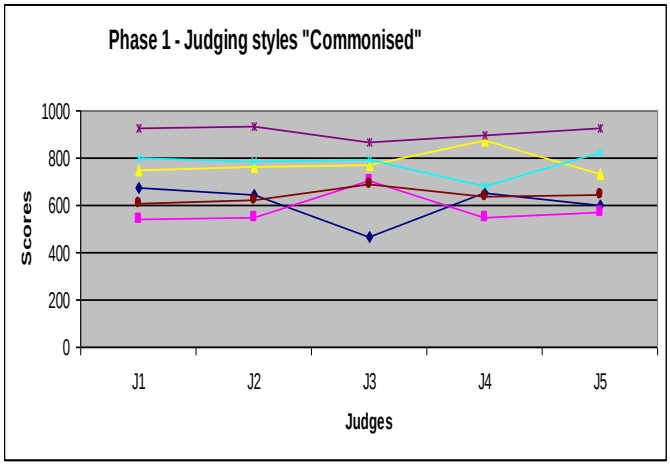

Figure 2: Phase 1: Judging styles 'commonised'

Phase 1: Commonising the judging styles

To make the comparisons that are required, the process must first re-model the scores to bring all the judging styles to a common format and remove any natural bias between the panel members. Following some calculations therefore, each Judge's set of scores is squeezed or stretched and moved en-bloc up or down so that the sets all show the same overall spread (style) and have identical averages (bias). Within each set, the pilot order and score progression must remain unaltered, but now valid score comparisons are possible between all the panel Judges on behalf of each pilot (Fig 2).

Phase 2: TBL processing

Now TBL looks at the high and low scores in each pilot's set, and throws out any that are too "far out" to be fair. This is done by subtracting the average for the set from each one and dividing the result by the "sample standard deviation" – if the result of this sum is greater than 1.645 then according to statistical probability theory we can be at least 90% confident that it is unfair, so the score is discarded. This calculation and the mathematically derived 1.645 criteria is the key to the correctness of the TBL process, and is based on many years of experience by CIVA with contest scores at all levels.

Of course, for a pilot, the discarding of any scores changes the average and standard deviation of his remaining results, and so the whole process must be repeated. After several cycles any "unfair" scores will have gone, and those that remain will all satisfy the essential 90% confidence criteria.

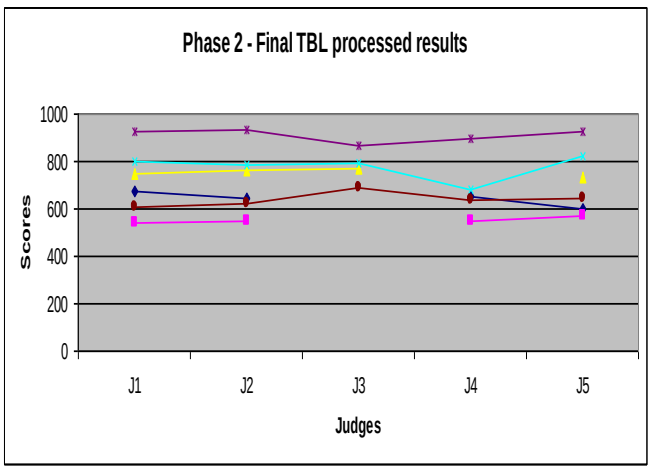

Figure 3: Phase 2: Final TBL processed results

The final published results (Figure 3)

As is usual, these are derived from an average of each pilot's scores. The final TBL iteration therefore has any appropriate penalty marks applied (for box or minimum height infringements etc), and the total scores are then sorted in descending order to rank the pilots first to last.

Educating and improving the Judges

One obviously useful by-product of this process is that it provides all of the evidence to assess how each Judge has performed by comparison with the overall judging panel average and when seen against the 90% level of confidence criteria. Up until now, our Judges have been secure in the position of having their own performance assessed only by the occasional irate pilot.

The TBL system as installed in our computers will now produce as a matter of course an analysis for each (anonymous) Judge showing the percentage of scores accepted as "OK", and a comparison with the panel style (spread of scores) and bias (average). Our intention here is to make it a good thing for a Chief Judge to promote genuinely positive and educational discussions with each member of the panel whilst the judging process is still fresh in their minds, and so to unravel and resolve any problems or difficulties that were experienced.

What will TBL do and what won't it do?

This is of course a "dumb" system which – although functioning to an impeccable set of theoretical rules – almost by definition brings with it a 10% possibility of upsetting an honest Judge's day. The trade-off is that we expect not only to achieve a set of results with at least 90% confidence that they are "fair" every time, but the system also provides us with a wonderful tool to address our judging standards. Be assured that our panel of Chief Judges is looking forward with interest to a season with a series of briskly analysed judging results. Who knows, we might even have a trophy or two for demonstrably "good" standards of judging. Well whatever next!

What does TBL do? In essence, it ensures that every Judge's opinion has equal weight, and that each "sequence score" by each Judge is accepted only if it lies within an acceptable margin (90% probably "ok") from the panel average – this average is thus the arbiter of "correctness".

What doesn't TBL do? If anything I suppose the strength could also be a weakness – that is, it of necessity takes the dominant judging panel view as the "correct" one. It certainly can't make right scores out of wrong ones – if six out of eight Judges are distracted and make a 'pig's ear' of one pilot's efforts, then for TBL this becomes the controlling assessment of his performance, and the other two diligent souls who got it right will see their scores unceremoniously zapped. In practice this would be extremely unusual. From the Judging line, it is almost impossible to deliberately upset the final results without collusion among the majority of Judges, and if that starts to happen then someone is definitely on the wrong planet.

The step taken by the BAeA is to permanently engage the TBL process at every event for all levels above Beginners. Our tests have been thorough, and interestingly also show that below a certain number of Judges and/or contestants (a block around five by five) the TBL +/- 10% limits become sufficiently broad to exert very little effect at all – further proof that the more judges there are the more arguments you'll get … you know how it goes! We are certainly expecting to see a real step forward in our ability to monitor and improve the standard of judging at our contests.

A simple demonstration of how TBL does its job

What follows here is a very simple example of the TBL process applied to a limited field of pilots and Judges. The raw scores have deliberately been set to depict a pretty dreadful standard of Judging in order to negate the worst aspects of the "small numbers" syndrome, and also to better illustrate for you the way the system works. Some graphical presentations of each phase have been included to make it easier to see how the "commonising" process and the 90% confidence criteria ease out the unacceptable scores. For simplicity here penalties have been ignored.

In this example the Judging Panel behaved generally as follows:

Judge 1 – had a normal spread of scores without significant bias.

Judge 2 – had a normal spread of scores but with a relatively low scoring bias by comparison with the panel average.

Judge 3 – had an average bias but unfortunately also had an outstanding dislike of Pilot 1.

Judge 4 – was unusually impressed by pilot 3 and fairly cold to Pilot 4.

Judge 5 – similar to Judge 1 – had a normal spread of scores without significant bias.

Table of raw scores after all pilots have flown. *'Traditional'*
The maximum possible score is 1000 marks per pilot. *Results sheet*

	J1	J2	J3	J4	J5	Rank	Score
P1	700	610	300	725	620	6th – P1	591.0
P2	600	545	650	625	590	5th – P2	602.0
P3	760	695	750	945	740	2nd – P3	778.0
P4	800	710	785	755	825	3rd – P4	775.0
P5	900	815	895	965	920	1st – P5	899.0
P6	650	595	630	710	660	4th – P6	649.0

The first step is to calculate raw Mean and Sample Standard Deviation data for all Judges.

	J1	J2	J3	J4	J5
Mean	735.0	661.7	668.3	787.5	725.8
St. D.	108.4	97.7	204.5	136.9	127.8

Figures for the whole panel

Mean for all Judges = 715.7

St. D. for all Judges = 138.4

Phase 1 – Commonise the judging styles of the raw scores using step 1 results to expand or contract the variances and to standardise all of the averages.

Now calculate the TBL 90% confidence limits of score acceptability for each pilot (Mean +/- 1.645 St D)

	J1	J2	J3	J4	J5	Mean	St D	Low	High
P1	700	610	300	725	620	606.7	82.5	470.9	742.4
P2	600	545	650	625	590	583.3	67.7	472.0	694.7
P3	760	695	750	945	740	777.5	56.6	684.4	870.6
P4	800	710	785	755	825	776.7	54.4	687.2	866.1
P5	900	815	895	965	920	909.9	27.1	865.3	954.6
P6	650	595	630	710	660	639.9	31.4	588.3	691.5
Mean	715.7	715.7	715.7	715.7	715.7	For each Judge the Mean and the			
St D	138.4	138.4	138.4	138.4	138.4	Standard Deviation is now the same.			

Phase 2 – Disregard scores falling outside the 90% confidence limits and using only the remaining scores re-calculate the statistical data and then repeat the process until all the results lie within their appropriate set of high/low limits.

	J1	J2	J3	J4	J5	Mean	St D	Low	High
P1	671.0	642.4	(466.4)	652.5	601.1	641.7	29.6	593.1	690.4
P2	543.3	550.3	(703.3)	551.3	568.6	553.4	10.8	535.7	571.2
P3	757.6	762.9	770.9	(874.9)	731.0	753.1	17.6	724.1	782.1
P4	798.7	784.2	794.6	(682.8)	823.0	800.1	16.4	773.1	827.2
P5	926.4	933.0	869.1	895.2	925.9	909.9	27.1	865.3	954.6
P6	607.1	621.2	689.1	637.3	644.4	639.9	31.4	588.3	691.5

The scores in brackets (…) above lie outside TBL limits and have been discarded.

	Rank	Score
	1st – P5	909.9
	2nd – P4	800.1
Final TBL	3rd – P3	753.1
Results Sheet	4th – P1	641.7
	5th – P6	639.3
	6th – P2	553.4

Assessment of Judges' Performance

Measurement	J1	J2	J3	J4	J5
Scores accepted %	100	100	66	66	100
Judge/Panel Mean %	103	93	93	110	101
Judge/Panel St. D. %	78	71	148	99	92

So – what has happened?

From the first step you can see that the bias of Judges 2 & 3 (low) and 4 (high) was quickly resolved, and an early indication of the style problems shown by Judge 3 is clearly identified.

Look now at the box of phase 2 scores and you can see that both Judge 3 & 4 had their scores for the pilots to whom they were "unfair" identified and discarded, and also Judge 3 to pilot 2 who (by comparison) he seriously over scored. Would you have noticed that?

The TBL ranking as you can see has modified the order that the "traditional" results sheet would have given, and also provided the material for the Chief Judge to identify the causes and chat to the two Judges concerned.

If you're really honest you have probably been all of these Judges at some time or other. You might have admitted that to yourself, but to others? In my short experience the real truth of the game is to fly at one level and judge the rest. Are those other guys on the panel as good to you as you are to them?

Note: If a pilot, for various reasons, cannot make his flight or if his model does not satisfy the noise test, the flight is scored zero. This score must not be taken into account in the TBL process.