Interpreting Flight Results

Continuing our discussion of judging metrics

BY DOUG LOVELL

In the April issue article about judge metrics, we saw that IAC now has a number of measures of judge performance in addition to measures of pilot performance. We showed that:

• Judge metrics on flight results tell us which judges agreed about the pilot performances on that particular flight;

• No one can draw conclusions from one flight about whether a judge will agree with the judge panel on another flight; and

• No one can draw conclusions from one flight about the quality of a judge in general.

We promised to look more closely at the individual flight results and what they reveal.

In order to explore what flight results reveal, we'll look at one of the flights posted at *http://www.iaccdb.org*. The flight is ancient history; but, in order to keep peace, we're stripping away the names. Always remember one flight does not make or break a pilot or a judge. Consistent performance over time is the telling factor.

SPENCER THORNTON

Pilot	J1	J2	J3	J4	Score		PnIt	Result
P1	2498.50 (1)	2521.50 (1)	2243.00 (2)	2573.50 (1)	2459.13	74.75% (1)	30	2429.13 (1)
P2	2343.50 (2)	2391.00 (2)	2414.50 (1)	2520.50 (2)	2417.38	73.48% (2)	60	2357.38 (2)
P3	2246.50 (3)	2239.50 (3)	2017.50 (3)	2246.00 (5)	2187.38	66.49% (3)	120	2067.38 (3)
P4	1982.00 (5)	1935.00 (5)	1845.50 (4)	2354.50 (4)	2029.25	61.68% (5)	0	2029.25 (4)
P5	2160.50 (4)	1946.00 (4)	1838.50 (5)	2462.50 (3)	2101.88	63.89% (4)	90	2011.88 (5)

ρ = Rho
ri = RI
γ = Gamma

	ρ	ri	γ	Np	Npp	Nf	Fza	Ng	Ka	Mz	Mg
J1	100	0.00	100	5	10	1	5.00	50	32.90	0	0
J2	100	0.00	100	5	10	1	5.00	50	32.90	0	0
J3	80	9.20	60	5	10	1	5.00	50	32.90	0	0
J4	70	11.63	60	5	10	1	5.00	50	32.90	0	0

SCREEN 1

Let's look at the five-pilot, four-judge flight in Screen 1. The first thing to note is that the ordering of the pilots in agreement with the judge panel is 1-2-3-5-4 (See 'E'). That is because the fourth place pilot according to the judges took ninety penalty points ('F'). Because the fourth and fifth ranked pilots are close in points, ninety penalty points are sufficient to drop the pilot ranked fourth below the pilot ranked fifth by the judges. When we look at judge performance we compare each judge with the results before penalties.

Judges J1 and J2 ranked the pilots in agreement with the four panel. You'll see ('B') that they don't give necessarily higher scores or even produce much more of a range of scores. They aren't dominating the judge panel in any way. It's simply that, when you add up the scores from each judge, the result ranking agrees with the ranking given by J1 and J2.

Agreement on the fourth and fifth ranked pilots is poor. There is not unanimous, but at least majority, agreement on the three top-ranked pilots.

The Rho (ρ), RI (ri), and Gamma (γ) metrics show that judges J1 and J2 were in agreement with the ranking ('G'). They have perfect 100 Rho and Gamma, perfect zero RI. For judges J3 and J4 we see by Gamma they each got the same number of paired rankings correct ('J'). RI was tough on J4 ('I'). Rho shows also that J4 rankings had the least strong correlation with the panel ('H').

Looking at the rankings of judges J3 and J4 ('C') we see that J4 had the top two pilots ranked in agreement. Judge J3 thought P2 was the better pilot. Note that if pilot P2 looks only at scores ('A'), it appears judge J4 liked him better than judge J3. Judge J4 gave pilot P2 a higher score than any other judge. The rankings reveal that judge J3, not J4, is the judge who most favored pilot P2.

Judge J4 gets into trouble with the Rho and RI numbers by ranking the third place pilot last ('D'). Rho and RI pick that up. (Gamma picks up only that both judges J3 and J4 correctly ranked eight of the ten pilot pairings. Judge J3 has rank disagreements for P1 v. P2 and P4 v. P5. Judge J4 has rank disagreements for P3 v. P5 and P4 v. P5.)

DEKEVIN THORNTON

P3

F	J1	J2	J3	J4	Avg	K	PtsLoT
1	8.5 (1)	8.5 (1)	8.0 (1)	8.0 (3)	8.3 (1)	43	75.2
2	7.0 (5)	8.0 (2)	7.0 (3)	7.5 (4)	7.4 (4)	36	94.5
3	5.5 (5)	4.5 (5)	4.5 (5)	6.0 (5)	5.1 (5)	40	195.0
4	8.0 (1)	8.5 (1)	7.0 (1)	9.0 (1)	8.1 (1)	24	45.0
5	7.5 (3)	8.0 (1)	7.5 (1)	7.5 (3)	7.6 (1)	49	116.4
6	8.0 (1)	8.5 (1)	8.0 (1)	7.5 (1)	8.0 (1)	22	44.0
7	0.0 (4)	0.0 (4)	0.0 (4)	0.0 (4)	0.0 (4)	28	280.0
8	8.0 (1)	6.5 (3)	5.0 (3)	8.0 (2)	6.9 (2)	42	131.2
9	7.5 (4)	8.0 (1)	8.0 (3)	7.5 (5)	7.8 (4)	25	56.2
P	7.5 (3)	7.5 (3)	6.0 (3)	6.0 (5)	6.8 (3)	20	65.0
Points	2246.5 (3)	2239.5 (3)	2017.5 (3)	2246.0 (5)	2187.38 (3)		

Penalty	120
Points Earned	2067.38 (3)
Max Possible	3290.00
Percent of Possible	62.84

SCREEN 2

The next important question to ask is what judge J4 can learn from this. If you are judge J4, you can see your ranking for pilots P3 and P5 ('D') are the ones which disagree the most. Screen 2 shows the grades for pilot P3.

In Screen 2, you're looking for figures graded lower than the other judges. In particular, lower than judges J1 and J2 since they were the judges in agreement. Looking at the grades ('C'), we find nothing. Your grades are for the most part equal to or higher than the grades of the other judges. To get any information, you have to look at the ranks.

You can see by the rankings for figures one, nine, and presentation ('A', 'E', 'F') you (judge J4) gave relatively worse scores to pilot P3 than did the other judges. The error on figure one ('A') causes the most damage, because that is the figure with the highest K value ('D'). Your (judge J4) eight was lower than the score you gave two other pilots, unlike judge J3's eight ('B'). Judge J3's eight is the highest grade given by J3 on that figure.

As judge J4, you can conclude that you might be missing something on the first figure of the flight—something the other judges saw when grading the other pilots. To confirm, look at the pilot P5 grades in Screen 3. The P5 grades clarify the picture. As judge J4 you gave Pilot P5 your best score on the first figure ('B') while the other judges gave pilot P5 their worst score ('A'). Clearly, you missed something on the first figure that the three other judges caught. It's probably too late to ask them exactly what they saw for that pilot on that day; but, you could ask them what they look for on that particular figure.

> Numbers, numbers, numbers. As one person said when they looked at these, "That's a lot of numbers."

After consultation you might conclude that the other three judges had something funny for breakfast; that you had it right. Remember that the numbers tell you where you differed and whether you were in the majority or the minority. They don't tell whether you were right or wrong. If you are one out of four, however; the likelihood is high that you missed something.

P5

F	J1	J2	J3	J4	Avg	K	PtsLoT
1	7.0 (5)	5.0 (5)	6.0 (5)	8.5 (1)	6.6 (5)	43	145.1
2	7.5 (4)	8.0 (2)	4.0 (5)	9.5 (1)	7.2 (5)	36	99.0
3	8.0 (1)	8.5 (1)	7.5 (1)	9.0 (1)	8.2 (1)	40	70.0
4	7.0 (3)	7.5 (3)	7.0 (1)	8.5 (4)	7.5 (3)	24	60.0
5	7.5 (3)	7.5 (3)	6.0 (3)	8.0 (2)	7.3 (3)	49	134.7
6	6.0 (2)	7.0 (2)	6.0 (3)	7.5 (1)	6.6 (3)	22	74.2
7	0.0 (4)	0.0 (4)	0.0 (4)	0.0 (4)	0.0 (4)	28	280.0
8	6.0 (4)	2.0 (5)	5.0 (3)	7.0 (4)	5.0 (5)	42	210.0
9	8.0 (1)	7.5 (4)	8.5 (1)	8.0 (2)	8.0 (2)	25	50.0
P	7.5 (3)	6.5 (5)	6.0 (3)	7.0 (3)	6.8 (3)	20	65.0
Points	2160.5 (4)	1946.0 (4)	1838.5 (5)	2462.5 (3)	2101.88 (4)		

Penalty	90
Points Earned	2011.88 (5)
Max Possible	3290.00
Percent of Possible	61.15

SCREEN 3

We must take care as judges that we don't try to get good match numbers. We don't want to go out with the attitude of trying to game our statistics, perhaps by restricting the range of our grades or trying to guess which pilots the other judges will favor. The best judges are the judges who judge frequently, and who fairly apply consistent criteria. The statistics are a device for finding insight, not an end in themselves. The judge "leader board" posted at http://iaccdb.org/leaders/judges/ primarily highlights judges who are most actively judging-- judges who judge a lot.

Let's switch gears now and look at what Screen 3 can tell the pilot. You are pilot P5 looking at your grades for the flight. First, note that the seven from judge J3 on figure four ('K') is not the same as the seven from judge J1 ('J'). Judge J1's seven is a third place seven. Judge J3's seven is a first place seven. Likewise, on the same figure four, the eight-five from judge J4 ('L') is not the most favorable grade. It is in fact nearly the worst grade from that judge for that figure. That is merely to point out that raw grades don't tell the whole story, or even much of the story. To get the real picture you have to look at ranks. So as pilot P5, how can you improve your dismal performance? Where will you gain the most benefit from practice and improvement? First, look at the good stuff. All of the judges thought figure three was flown better than any other pilot flew figure three ('I'). Almost the same for figure nine ('M'). Whatever you're doing on figures three and nine, keep it up. Next look for figures ranked especially low. Figures one, two, and eight were ranked last ('D', 'C', 'N'). Of those, figures one and eight have the highest K value ('E', 'O'). Check the comments on figures one and eight. Work on those. At last, look at the "PtsLoT" column. That's "points left on the table." Obviously it never helps to zero a figure ('G'), even though one other pilot did the same on this known flight. Going beyond the zero, it's evident that improvement on figure eight has the most to gain in terms of getting points back ('H'). Figure one is a close second ('F'). Points is what it's all about. More points win. Figure-out what you need to improve to score better on figures one and eight. Work on that in practice. Get that right and the next outing will have a better outcome.

Numbers, numbers, numbers. As one person said when they looked at these, "That's a lot of numbers." In the movie "Moneyball" with Brad Pitt (based on the documentary book by Michael Lewis) the manager of the Oakland A's baseball team gets to the World Series in 2002, with a very uncompetitive player budget, by using the numbers to select a winning team. If you made it through this article, you've picked up some tools to improve your results. As a pilot, you can get more value from gallons of fuel invested. As a judge, well, you didn't have much of anything to go on before. Now you do. *IAC*