MIKE STEINEKE

Just How Good Is IAC Judging?

A statistical look

BY DOUG LOVELL

The IAC is very fortunate to have a rich mine of aerobatic contest data. Thanks to Randy Owens, "Bwana" Bob Buckley, and our dozens of scorekeepers, we have just about every grade, from every judge, from every pilot, for every flight, in every IAC regional contest going back through 2005. That is almost 86,000 grades. More than is available from any other source.

With some prodding from a few other directors, including Klein Gilhousen, Tom Adams, and Wayne Roberts, I have compiled and processed this data in an attempt to get some meaningful information and measures of judging quality in the IAC. People talk more or less subjectively about whether the judging is any good. That usually goes along with their grades. If their grades are good, the judging is great! Here's an objective look at some numbers.

We came up with three different metrics with which to measure the performance of judges. All of them have to do with how closely the individual judge grading measures up against the collective, overall scoring result. The measures take into account two different comparisons of judge placement versus overall placement.

The first comparison is the actual score. We have the overall number of points achieved by a pilot versus the number of points given a pilot by the individual judge. The second comparison is the rank. A pilot's rank is the number of pilots who did better, plus one. The rank is commonly referred to as the placing. The first place pilot has rank one. The second place pilot, rank two, etc. We can compare the rank achieved by the pilot with the rank given by each individual judge.

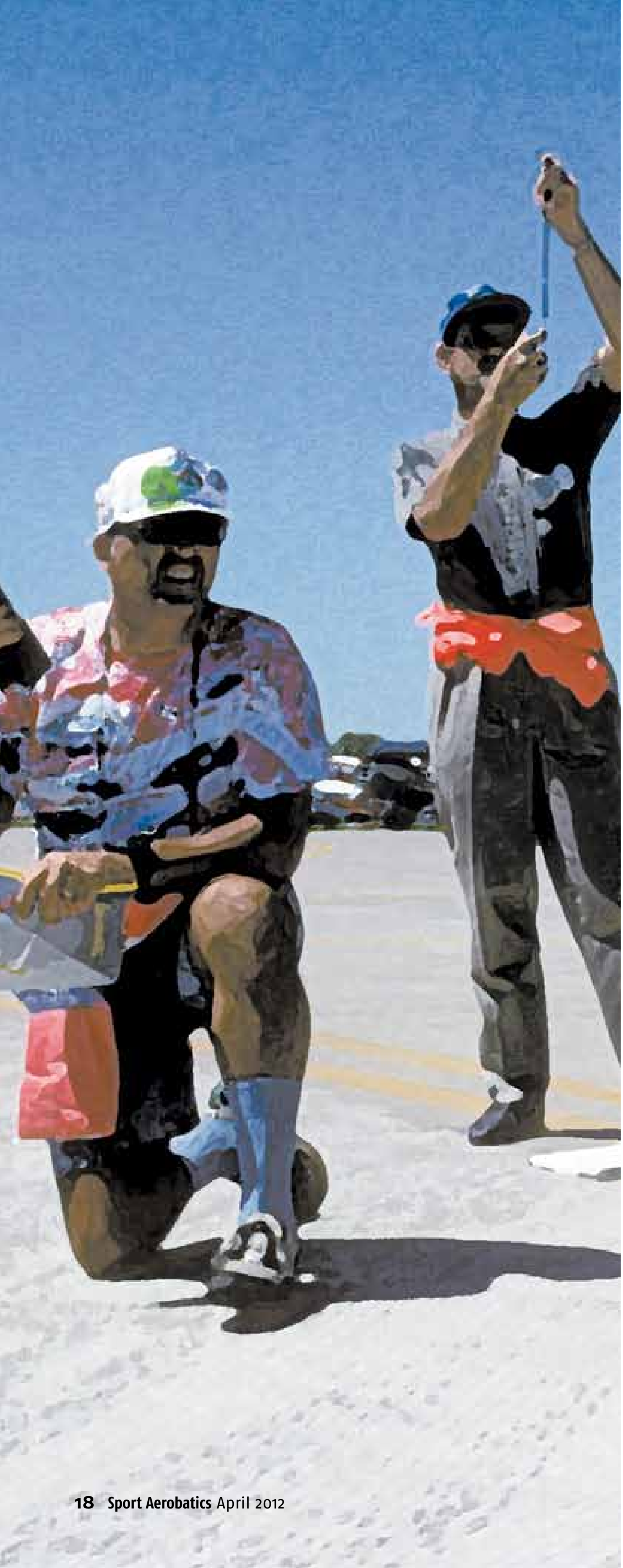
A major advantage of using rank is that rank strips away differences in scoring styles. A judge who gives generally lower grades might rank a pilot the same as a judge who gives generally higher grades. We ask judges to be consistent, and hope that each judge ranks the pilots fairly by applying consistent criteria in their grading.

The first of the three major judging quality measures we examined is RI (said "are eye"). RI is a formula invented by a few people at CIVA for evaluating international judges. A zero value for RI means the judge ranked the pilots exactly the same as the overall ranking, regardless of how that judge graded the pilots. When a judge ranks a pilot differently than the overall ranking, RI penalizes the judge to an extent measured by the difference in the judge's score and the overall score. Higher RI is bad. Zero or lower RI is good. RI makes no penalty for strange grading unless the judge gets the ranking wrong. When a judge gets the ranking wrong, RI penalizes strongly for grading differences.

The second of the measures we examined is Rho (said "row" as in "row your boat"). Rho is a standard textbook statistical metric developed by Charles Spearman, now in use for over a century. It is a distance formula that measures how far an individual judge's ranking of the pilots differs from the overall ranking.

A Rho value of 100 means the judge ranked the pilots in perfect agreement. A Rho value of minus 100 means the judge was perfectly upside-down. A Rho value of zero means the judge was neither in agreement or upside-down.

The last of our measures is Gamma (as in "gamma ray"). Gamma is a second textbook metric developed by Leo Goodman and William Kruskal at the University of Chicago in the 1950s. Kruskal served terms as president



RI is a formula invented by a few people at CIVA for evaluating international judges

of both the Institute of Mathematical Statistics and the American Statistical Association.

Gamma looks at every possible pairing of pilots in a flight. If both the judge and the overall ranking place pilot A before pilot B or vice versa, that is a "concordant pair." If the judge puts pilot A before B while the overall ranking places pilot B before A, that is a "discordant pair." The Gamma measures the proportion of concordant and discordant pairs for each judge. The interpretation of Gamma is the same as for Rho; 100 is perfect, zero is bad. Negative values are worse down to minus 100, which means the judge's rankings were upside-down relative to the overall rankings.

You can view mathematical details in the notes pages at IACCDB.org: *www.IACCDB.org/pages/notes#metrics*.

It's important to note the metrics don't tell us which judge was right. It's entirely possible that four judges agreed on ranking an inferior performance first while a fifth judge correctly gave a first ranking to a superior pilot. The judge with the lowest metric might, in some rare circumstance, be the only judge who saw the flight correctly. The metrics tell us only which judges were in agreement with the overall result. The only way to measure actual correctness of the judging is to compare with the judgments of an expert. If we could all agree who the expert is, we could put the expert on the judging line and let him or her decide the contest.

For all of the experiments, we took the judge metric data from all of the flights in which there were nine or more pilots. With fewer than nine pilots the data tends to get "noisy." On two-pilot contests, for example, there are sometimes a couple of judges who have minus 100 and high RI because they ranked the two pilots opposite the overall result. The nine-pilot mark left us with almost 4,000 flights to look at. For a good statistical analysis, that is plenty.

First, we looked at the metrics themselves to compare them. Do they measure the same thing or something different? Figure 1 shows an x-y plot of the Rho and Gamma metrics. Each point has the value of Rho and the value of Gamma for one judge on one flight. It's clear that if the value of Rho is high (good), the value of Gamma is also high (good). Rho and Gamma are what statisticians call "highly correlated." They are comparable measures. If you know the value of one, then you can fairly predict the value of the other.

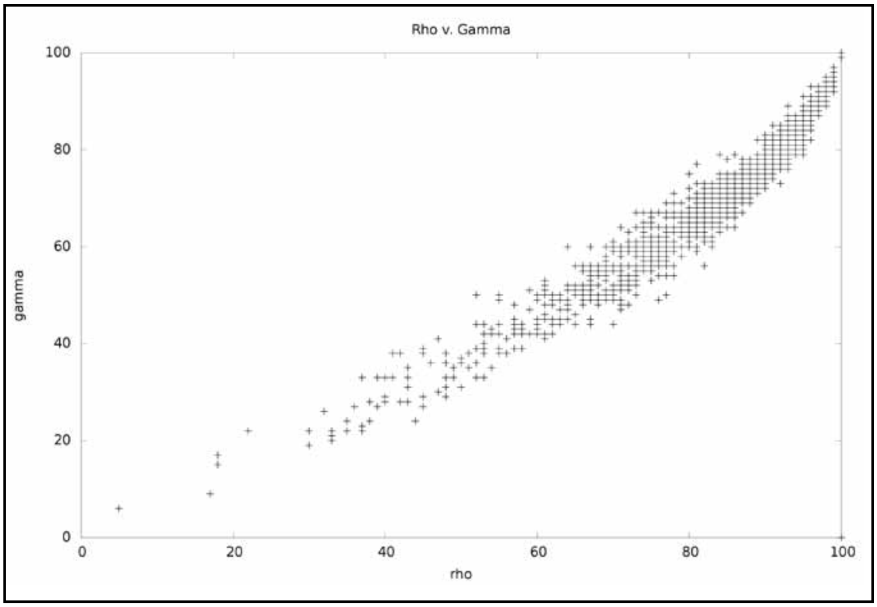


FIGURE 1

Next we looked at Rho together with RI. Do they measure the same thing or something different? Figure 2 shows an x-y plot of the Rho and RI metrics. Each point has the value of Rho and the value of RI for one judge on one flight. When Rho is high (good), RI tends to be low (good), but spreads in a range about five to seven points wide. As Rho gets lower, the RI spread becomes rapidly more pronounced.

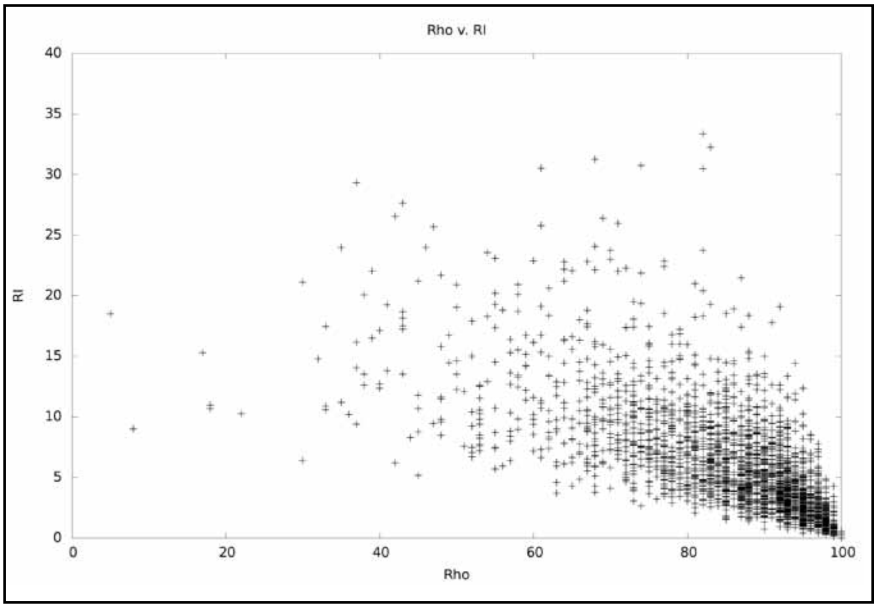


FIGURE 2

You cannot very accurately predict the value of RI given Rho as Rho gets lower, nor can you predict the value of Rho given RI. Whatever RI is measuring, it isn't exactly the same as what Rho (and by inference, gamma) is measuring. You can tell that very good Rho will share a corner with very good RI—sort of.

One more experiment. We decided to do more with Gamma, although we could just as readily have chosen Rho, and with RI, because RI is what CIVA uses. We decided to look at Gamma and RI for each individual judge. The question is whether Gamma or RI measure anything about a judge from which you can draw any inference about the ability of that judge. Does a good RI number mean you are a good judge? And how about with Gamma?

There was a typesetting error in the original article. This omitted text added for clarity.

To answer this question we plotted, for each judge, all of their Gamma and RI values. Figure 3 shows the plot for RI. Figure 4 shows the plot for Gamma.

First, for any given judge, the values do not cluster around any particular value. This means that the value of RI or the value of Gamma on one flight

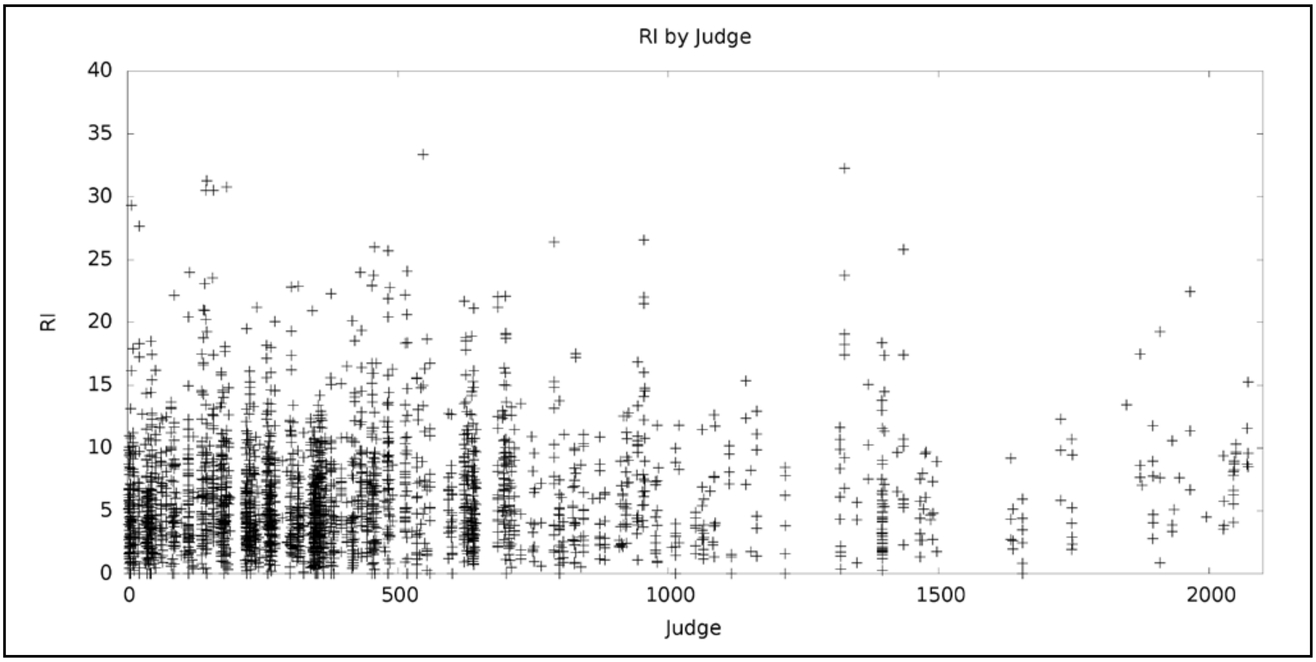


FIGURE 3

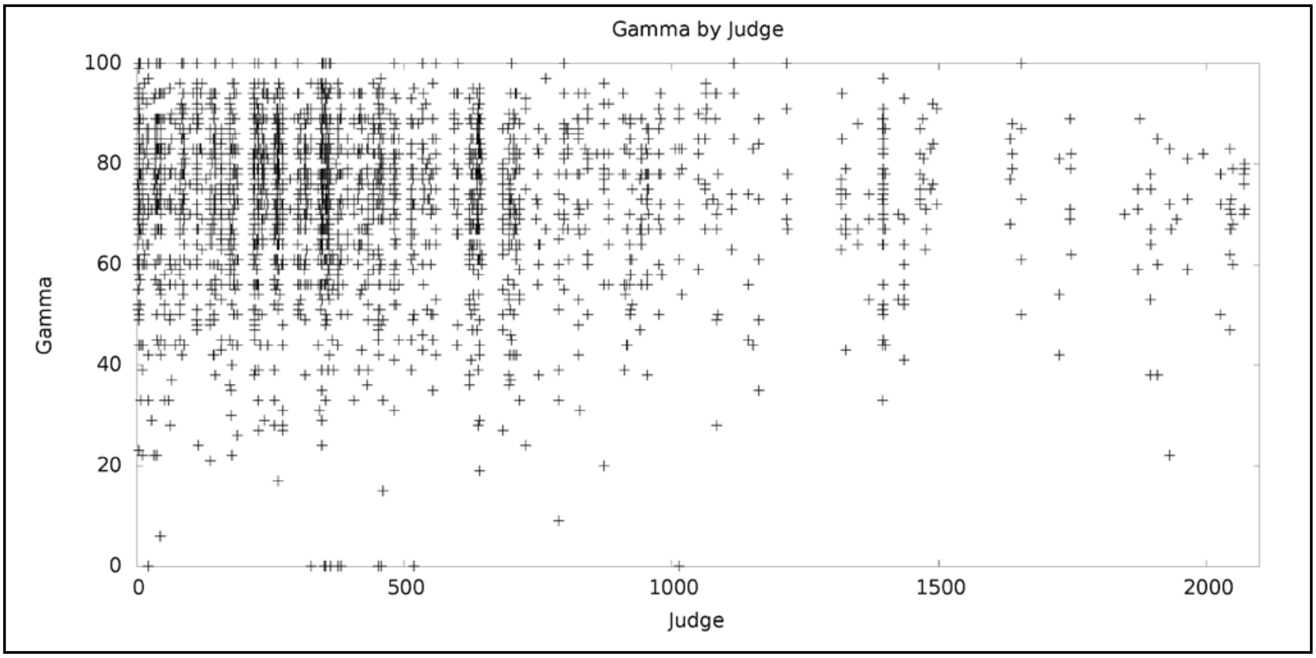


FIGURE 4

does not give any indication of how a judge will perform on the next flight. That a judge looks out of whack on one flight doesn't tell you they are a bad judge. Nor do zero RI and 100 Gamma tell you they're the best judge in the world. If they did that consistently on every flight judged they would be the best judge in the world. Doing it on one flight is good-great—for that flight.

Second, the values of RI and Gamma fall into about the same range for every judge. There is really good news in this. For all of the judges, most of the values are in the 55 to 100 range for Gamma and below 15 for RI. The histogram in Figure 5 shows the distribution of Gamma values assigned all judges on all of the flights. It confirms that the agreement of the judges is pretty good most of the time. We are very fortunate in the IAC to have, with occasional exceptions, a panel of judges who agree on the pilot rankings. The IAC can train judges, place them on the line, and get very good results.

We see every IAC judge without exception out of whack with the judging line once in a while, spot on ranking the pilots nearly perfectly once in a while, and most frequently ranking about three-quarters of the pairs in agreement with the result. With a 75 percent confidence of one judge having any pair-wise ranking correct, there is an 84 percent confidence that a three judge agreement is correct, 90 percent confidence that a five judge agreement is correct, and 93 percent confidence that a seven judge agreement is correct. The more judges who agree, the better our confidence in the result, and that's why we go to the trouble of fielding as many well-trained and competent judges as we can muster at a contest.

We can work with our training programs to improve the 75 percent number. We can monitor the number to verify improvement. Keep in mind that number is very good. On a 12 pilot flight there are 66 pair-wise rankings. Judges are getting about 50 of those in agreement with the panel.

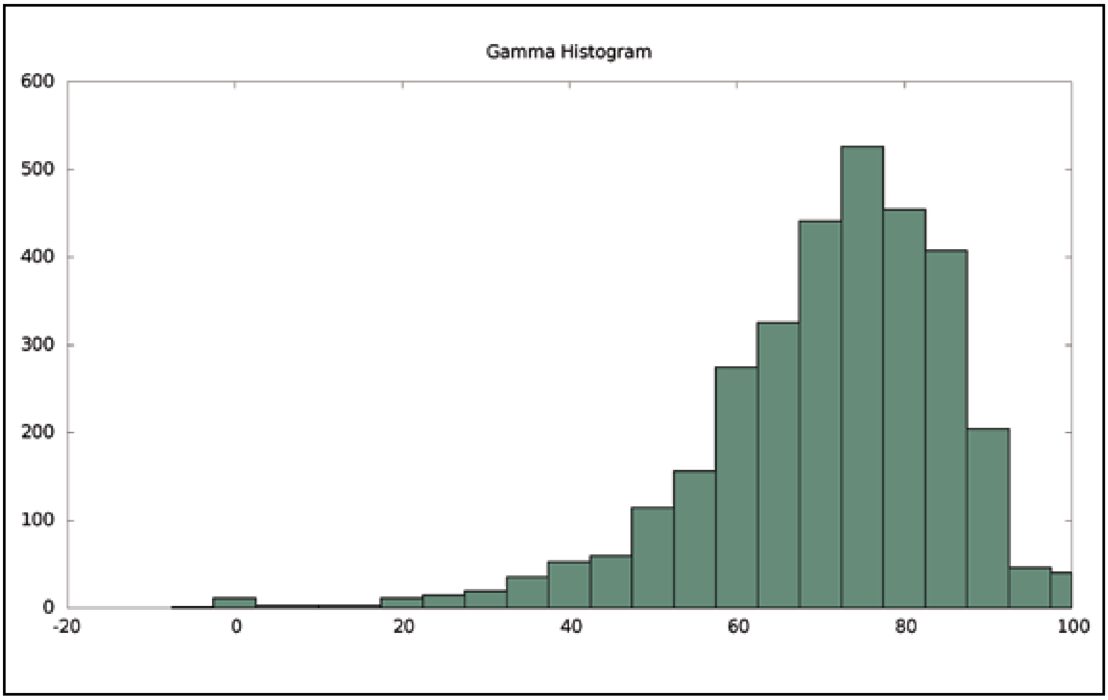


FIGURE 5

We'll look more in depth in another article at individual flight results and what they can tell us. The conclusions to draw from this article are these:

- We now have three metrics for every judge, on every flight, in every category, at every contest in the IAC. The Rho and Gamma metrics have a strong correlation, showing that they consistently measure something similar.
- The 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 how good the judge will be in general, or about whether a judge will agree with the judge panel on another flight.
- Looking at thousands of flights, the judge metrics show that, in the IAC, every judge will agree more closely on some flights, not so closely on others, and acceptably well just about all of the time. In general, we have very good judging panels in the IAC.

My thanks to Tom Myers, Wayne Roberts, Tom Adams, Klein Gilhousen, and Don Peterson for their reviews of this article. The article benefited greatly from their questions, suggestions, and observations. IAC