Don Ramsey
NSRCA Judging Committee Chairman
AMA 6096 NSRCA 1506
email: donramsey@gmail.com
website: www.pages.suddenlink.net/donramsey

After long hard work the "Judge Evaluation Committee" came up with a method for evaluating judges. Here's team member Matt Kebabjian's explanation.

JUDGE EVALUATION METHOD
Ron Van Putte, Tony Stillman and Matt Kebabjian with Don Ramsey consulting
8/2003

I. Background:

Generally, judging is one of the most important and demanding tasks to be done at any contest, and all of us who perform it must take it most seriously. Scoring should be as fairly assessed as possible. It wouldn't be a stretch to say that "good" judging, (as in fairly assessed judging), leaves us as contestants feeling pretty good about the contest as a whole, where "poor" judging usually doesn't. Ideally the judges should know the book and apply it without bias, for the "name" contestant and unknown contestant alike. The ideal best judge sees no "halo", is unperturbed by the power, appearance or size of the model, and is very observant of the many downgrades, knows their cost and applies them without hesitation.

There are many among us who take a great deal of pride in doing this most demanding task very well. How do we determine who these people are and how do they measure up against all the other judges? We have not had the means to assess judging performance in the past, or to give judges enough feedback about their performances such that they could improve in key areas.

Ron Van Putte and Tony Stillman had been thinking of a way to assess judging ability by observing the scoring habits or patterns of judges, and assign them performance scores. These performance scores could then be used to assess each judge's ability in getting the pilots placed correctly, in the various major meets, such as Nats and Team Selection contests.

Ultimately, a Judge Ranking System could be devised such that judging performance would dictate how the judges compared to some absolute, unachievable best, and to each other. It would then become a fairly straightforward task for the NSRCA President to choose the top 3 judges and pass these names on to the AMA President. The final recommendation of judges that could attend the World Championship representing the US, is the AMA President's. It is desired that the US representative Judge to the WC should come from that list.

At the 2002 NSRCA Board meeting at the Nats, Ron Van Putte was empowered to proceed with a plan to establish a Judge Ranking Committee. The charter of the Committee would be to find a method that would use the reams of judging or scoring data generated at a contest like the Nats, to determine how the judges performed. It was at the 2002 Nats that Tony and Matt Kebabjian had an initial conversation in regard to using the raw scores themselves to paint a picture of judge performance. The initial ideas he and Ron had been considering were brought out and Matt began to understand the daunting task of developing the statistical analysis program that lay ahead. The committee was formed; it included Ron Van Putte, Tony Stillman and Matt Kebabjian. Judging Committee Chair, Don Ramsey, was also included as a consultant

II. The Goal: First, a couple of thoughts on what the assessment is not:

- It isn't a written test that one may take, earning a mark or grade.
- The system is not exclusive; every Pattern Competitor is a potential top judge.
- It doesn't matter how good a modeler or craftsman or Concours d'Elegance winner a person is or whether he or she is a modeler at all.
- Theoretically, Judge Ranking is not a political appointment.

Rather, we hope it is a quantitative measure of a judge's ability to understand the downgrades and their just application. Assessing the ability of judges over time, as well as ranking the judges according to that ability, is the desired key goal of this exercise. **THE ASSESSMENT SHALL BE BASED ON HOW A JUDGE PLACES THE CONTESTANTS AND HOW CLOSE HIS SCORES WERE TO THE AVERAGE SCORES EARNED BY THE CONTESTANTS.**

Developing a computer program, which could analyze the data and arrive at Judge Rank Values, was key in establishing judging performance.

III. Considerations: Initial thought process delved into determining the Placement Error that each judge was achieving with his scoring habits. We went through considerable discussions and deliberation on whether this one factor was adequate in determining the relative Rank of any judge. We agreed that, in addition to Placement Accuracy, a judge also needed to show Scoring Precision such that the highest Ranking could be earned.

Our initial approach was to display Rankings as small fractions or decimals, however we later agreed to change the Rank display and to show it as a base-1000 normalized number, per round. The two Factors, the **PLACEMENT ERROR FACTOR (PEF)** and the **SCORING ERROR FACTOR (SEF)**, each are based on 1000 normalized points and are averaged to show the Judge Rank for each round. These have been renamed **NORMALIZED PLACEMENT ERROR or NPE and NORMALIZED SCORING ERROR or NSE, respectively.**

As the many judges accrue Rank Value scores, and these are summed, naturally one judge is likely to have the best overall performance over time, evidenced by a maximum score sum. The committee decided that 8 best rounds are to be used to assess the overall ranking of the judges. Further, it was determined that the top 3 judges would then become the group the NSRCA President would recommend to the AMA President.

The list of judges and their scores would become public knowledge, and as such, the Chief Judge at the key events for Precision Aerobatics, would also use it to choose up to 80% of the judges for for Finals judging assignments. Whenever possible, room should always be allowed in these Judge Panels for new judges, who can learn from the veterans, and help establish their credentials as well.

Although scores from every contest in the country could potentially be used and a database containing hundreds of Pattern People could be established, we chose to include the scores from major meets only. The F3A and Masters Nats Finals, and Team Selection Preliminaries and Finals are used in the assessment.

IV. Method: The calculations, it turned out, are quite straightforward. Using Excel Spreadsheet format, with it's extensive built-in functions, the calculations seldom required more than basic statistics. The details of the method are as follows:

NORMALIZED PLACEMENT ERROR, NPE, (aka, Placement Error Factor, PEF):

1- Average normalized scores from all judges, (the consensus), determine the actual placement of the contestants for the round, placing the contestants in order 1,2,3, etc.

2- Each judge's scoring practice also "places" each contestant for the round, 1,2,3 etc

3- The difference between the consensus placement and the individual judge's placement for any one contestant, is the placement error for that contestant. For example, consensus places one contestant 1st; the judge placed the same contestant 2nd; difference is 1, assigned to the judge for the contestant for the round. Same is done for all contestants, all judges for each round scored.

4- All placement errors are summed for each judge.

5- The Maximum possible Deviation is determined. That calculation is based on the number of contestants and it is basically a max number derived from getting every position exactly reversed. (The mid position is the only one that is not wrong in this calculation.) Mathematically, it is the number of contestants multiplied by ½ the number of contestants, and the result rounded down:

eg.- for 7 Finals contestants, 7 x 3.5= 24.5; rounded down= 24

Note that any Finals may have anywhere from 5 to 12 contestants. Since it becomes much more difficult to correctly place a greater number of contestants, the errors for larger contests could be large. This, in turn, would drive the **NPE** to be small and generally of useless value to the judge. It was important to find the means by which contests with varying pilot entries could be assessed more fairly and be comparable to one another.

Note that the Max Deviation increases quickly as the number of Finalists increases. As the Max Dev increases due to pilot number, so does the relative Placement error from each judge. However the ratio of the two numbers, or fractions of Placement Errors, should be reasonably similar. This results in considerably less dependence on contestant number and allows for direct comparison between contests with dissimilar contestant entries.

6- The Error Fraction is determined by dividing the Placement Error Sum for any judge by the Maximum Deviation possible

7- The Error Fraction is subtracted from 1 and the difference is multiplied by 1000. That is the **Normalized Placement Error, NPE,** shown on the evaluation**.**

eg.- for the above number of contestants, Judge 1 got 5 of the 7 pilots correct but missed 2 placements by a total of 3 places. 3/24=0.125…..(1-0.125)x1000=875 normalized points = **NPE**

NORMALIZED SCORING ERROR, NSE, (aka, Scoring Error Factor, SEF): The raw scores are used for this Factor

1- The average RAW SCORES for all judges (per contestant), establish the JUDGE CONSENSUS scores for each contestant.

2- The GRAND CONSENSUS SCORE is then determined; it is the average of all the individual consensus scores

3- A judge's actual raw score for each contestant, is then compared to the consensus for the individual contestant. The DIFFERENCE between judge raw score and consensus raw score is determined

4- The scoring differences per judge for all contestants, are averaged. This factor makes dependence on the number of contestants much less important compared to other methods we tried

5- A SCORING ERROR FRACTION is calculated by dividing the average scoring differences for the judges, by the grand consensus score.

6- The scoring error fraction is subtracted from 1and the difference is multiplied by 1000. This is the **NORMALIZED SCORING ERROR or NSE** shown on the evaluation.

7- We found that the fractions tended to be small when calculated this way, so we multiplied the fractions by a Kfactor to effect a normalized value that was more in line with the **NPE** values determined above. The KFactor in the calculations is 3. This move was necessary such that neither the **NPE** nor the **NSE** would dominate the average Rank Value

eg. - The Grand Consensus Score is 500
The Average Score Difference for Judge 1 is 25 points
The Scoring Fraction is 25/500= .05
NSE = (1-(3x.05)) x1000= 850 normalized points, (3= KFactor)

V. Final Judge Rank:

1- Take the average of the two Factors

eg.- (**NPE** + **NSE**)/2; (875+850)/2=862.5

2- This becomes the reported Rank for the judge for the round. The maximum possible is 1000 for any round, but it is essentially almost impossible to achieve. Both the NPE and the NSE error fractions would need to be 0 (zero), which would drive the **NPE** and the **NSE** to 1000 normalized points.

3- There is no further normalizing done of the individual **NPE**'s and **NSE**'s.

VI. Epilogue: That's the system we derived. It should be considered as a tool to be used to advance our art. We believe it achieves the Judge Ranking as desired. It also shows the area in which any judge may need to improve upon to get higher scores. A long time coming but worth the wait.

We actually saw some rather weird situations surface as we started to analyze the data. One would expect that there is good correlation between PLACEMENT and SCORING error factors. As **NPE (PEF)** becomes large, so would the **NSE (SEF),** for example. In reality, there is lower correlation between the two Factors than one would expect. Just because a judge gets all the contestants placed correctly it didn't mean that his scoring error would be very small also. The so called "hi" or "low" judge, could in fact place the contestants correctly but his scoring is off substantially from the average.

Another judge scores the contestants very close to average resulting in very good **NSE (SEF),** yet his placement was poor earning him a poor **NPE (PEF).** The reason this happens is that a judge will earn a whole point of error in placement, even if his score is only off by a fraction of a point from average. The error in placement contributes significantly to his **NPE** where the fraction of point difference in his scoring doesn't contribute significantly to his **NSE**. These types of findings were abundant throughout the analysis. It turned out to be very intriguing overall, and basically indicated that a judge had to get both areas very close or suffer a poor Rank score.

One final thought: the method we derived is by no means a static thing, never to be touched again. Amendments should occur periodically as more knowledge into its function is obtained.