An objective approach for evaluating judging skills

12/19/2009

Prepared for MFTHBA,

By
Dr. Kenneth Kemp, Ph.D.
Kansas State University

Background

There are many problems associated with judged events regardless of the endeavor or the organization involved. From scandals at the Olympics to members of various clubs and organizations crying foul, there is, and always has been, widespread dissatisfaction with the outcomes of many judged events. Fair and consistent judging are the key factors to minimizing the level of discontentment that participants feel when being judged. Judging is an acquired skill and both good training and a lot of practice are fundamental prerequisites for becoming a competent judge in addition to one having the necessary capabilities required to keep track of many facts about many entries without becoming confused.

To assure that judges are competent, organizations need methods in place that allow them to objectively evaluate how well a person can judge before they are allowed to judge important events. In order for judge training to be improved, there must be a way to assess the effectiveness of existing training programs. The best way to determine the effectiveness of a training procedure is to objectively quantify the amount of improvement associated with it.

We are all aware of the discontentment regarding the judging within the MFTHBA. Of course such dissatisfaction is not unique to the MFTHBA. However, clearly those who spend so much effort, time, and money to compete in shows deserve to be judged as fairly and competently as possible. Several years ago, as a professional statistician, I began to think of ways to evaluate judging performance and about eight years ago I was able to interest Dr. Paul Nelson, a colleague of mine, in the problem as well. Over the past few years we have been able to develop techniques that allow us to objectively quantify differences in the performances of judges. Our work has resulted in three publications related to this area of research: “Small Sample Estimation of a Baseline Ranking” which appeared in *Communications in Statistics - Theory and Methods, Vol. 29, No. 1, 2000, pp 19-43,* “Testing for the Presence of a Maverick Judge” which appeared in *Communications in Statistics - Theory and Methods, Vol. 32, No. 4, 2003, pp 806-826,* and “*Performance Based Bayesian Inference for Distance Models on Partial Rankings*” which appeared in the *Journal of Statistics and Applications, Vol. 1, No. 1, 2006, pp 91-111.*

Proposal

Our technique for objectively evaluating judging performance relies on using a panel of highly experienced, proven experts to provide “Official” placings of video taped classes. Such a panel should consist of five to eight of the very best judges in the organization. The panel will be asked to review video taped classes and to subsequently arrive at a placing for each class. Each panelist would be asked to place the classes independently and any class where other than minor disagreements occur would not be included for further use, i.e. only classes where there is strong agreement among the experts would be used to establish a basis for videoed classes that would be used for testing judge candidates. The consensus placing of the panel will be considered to be the “correct” placing for the respective classes and should reflect and conform to the current judging standards of the breed. We need 10, or more, such classes with 10 horses per class to attain optimal performance of our technique but more horses could be used. The classes should include horses that are different enough that it would be reasonable to expect the panel of expert judges to agree on the placings in the classes particularly with regard to the top 3 places. Classes with very close pairings of horses, where it is not clear to the expert eye that one horse is better than another should be eliminated because strong agreement among the expert judges is of primary importance in establishing the “correct” or “Official” placement for each class.

Once the “Official” placement for each videoed class is obtained, prospective judge candidates would be asked to review the classes and provide their own placings. Each candidate's placings would be summarized by our techniques to quantify how closely the candidates’ placings agree with those of the expert panel. Those who come closer to duplicating the “Official” results will have demonstrated they are among the better judge prospects according to our criteria. Our procedure puts emphasis on correctly identifying the top 3 places in each class. Once we have data from many candidates we will be able to establish criteria that will allow us to classify a particular prospect in the upper, middle, or lower performance range. The identity of those participating will, of course, be kept in complete confidence.

The information provided by our measures of agreement will be useful to candidates in determining their level of performance as a MFTHBA judge. Sub par performers may be given an opportunity to take further training and then be subsequently reevaluated, should they so chose. They would have to demonstrate sufficient improvement before they could become a carded judge.

Identifying individuals who perform in the upper range according to our statistics would allow the organization to establish a pool of judges who are qualified to judge important shows such as the Three Old Futurity and the Celebration. Only those who perform at the upper level of performance should be eligible to judge such shows.

Our procedure can be used to evaluate the effectiveness of training procedures as well as to evaluate individual judging performance. Serious candidates, who do not do well in placing the videoed classes on their first attempt, should seek further training and experience as a means of improving their judging ability. After further training they would be reevaluated. The change in their performance would be quantified by computing the difference in their before and after training performance levels according to our statistics. If an individual fails to improve adequately after taking the prescribed training, it would be likely that either the training they received was ineffective or that the particular individual lacked the necessary talent and/or commitment to become a qualified judge. However, if the majority of those receiving additional training fail to respond sufficiently, it would likely be an indication that ineffective training methods were being used and revised training procedures would need to be developed that would more effectively convey the important aspects of judging which would help candidates recognize the important differences among the horses they judge in a show and/or on a video.

Conclusion

The MFTHBA would be taking a major step forward in being able to identify individuals who have superior ability to judge major events by adopting an evaluation method similar to that described above. The testing procedure could be the foundation of formal judge certification program. Unrest with regard to judging should be greatly reduced as a result of the judges being better trained and qualified and the contestants knowing that the individuals judging their horses have objectively proven that they are among the most competent judges available in recognizing which horses meet current breed standards. If those who score best using our statistics are also those chosen to be judges at the important shows, there can be no question as to why they were chosen and, thus, no accusations of foul play or selection bias could be justified. Everyone would be more confident of fair treatment knowing judges were chosen based on an objective criterion.