Some observations on the MFTHBA's judge rating system *by Ken Kemp, Ph.D.*

When I read the anonymously written "***Attention in the barns …*** " article in the December, 2010 issue of the journal, I was surprised to see references made to my opinions and philosophy regarding judge ratings. I have been working on the development of a judge rating system for years. The details of what I have done regarding a procedure, data analyses, my opinions and philosophy regarding judges, judging, and judge rating, as well as a 12 year summary of recent MFTHBA Celebration judging results, are all on-line at: http://www.kenkemp.net/judging/. My approach is based on measuring agreement among judges which is different than the current MFTHBA rating system.

The "***Attention in the barns …*** " article incorrectly gives the reader the impression that I was involved with, approved of, or had influence on the recently adopted judge rating system. I do not understand why there was an attempt to link me to a rating procedure that I was not involved with in any way other than having had objections to it and concerns about it, which I clearly voiced to all who were involved with its approval process.

I was asked to be on the Judges Committee by President Jim Mann in February, 2010. I agreed to serve and we both thought at the time that I might be able to work with the Judges Committee in developing a judge rating system. However, that did not occur. It was clear from the beginning that the committee was not interested in my input. Part of the problem may have been that I live too far away from Ava to be able to attend meetings in person. When I tried attending the first meeting by telephone, I was ignored. After that I found communicating with the Judges Committee in any form to be a completely one-sided conversation. I communicated with them but they did not communicate with me. Donna Watson sent me meeting minutes occasionally and that was my only source of information. Why a committee that was interested in formulating a judge evaluation procedure chose to not work with a member of the committee who is a statistician and who had worked on rating judges for many years, is hard to understand.

When I found out, via the minutes, what the committee was considering using as a judge rating system, I immediately made my concerns known to the committee, to committee co-chairman Duane Scott, to President Jim Mann, Vice President Joyce Greaning, and Secretary-Treasurer Julie Moore. I sent email and I sent a 9 page letter detailing what I thought was wrong with the then proposed rating system and explained what I thought the committee needed to consider before adopting the proposed rating system. That included my opinion that the procedure should be well tested before ever being approved for use by the MFTHBA.

The committee adopted the rating system after Jerry Middleton proposed the idea to the committee at the April 6th, 2010 Judges Committee meeting. The minutes of that meeting make no mention of the details of the procedure, i.e. how it would work, how it would be used to identify outstanding judges, what the criteria would be for rater selection, etc. It was apparently a "Hey, let's try this!" kind of an idea. I wasn't at the meeting so I cannot say exactly what went on but there were no details in the meeting minutes as to how the scores were to be computed and there was no discussion as to what the committee thought a AAA rating should mean other than that judge candidates would have to score 80% or higher by the procedure. The system was to be based on having persons called raters who would reach a consensus of what the correct class placement would be for the top 10 horses in a class. There was no concern expressed as to

whether raters sitting in their seats would be able to see horses as well as the judges and, therefore, whether the raters' scores would be as or more reliable than those of the judges who would be much closer to the horses being judged in a class. It would seem that the criteria that would be used for selecting the raters required by the procedure would have been a key point of discussion given that their collective opinion (their composite ranking of a class) would be considered more important and more nearly correct than that of any of the judges when disagreements between the committee and a judge occurred. But no mention was made of any concerns about choosing raters in the meeting minutes. The most surprising, and most disappointing to a statistician, is that the cut-off values for A, AA, AAA ratings were apparently pulled from thin air. The rating system cut-off values that were suggested, which were subsequently adopted without question, were that 75% and below would qualify a person as an A rated judge, above 75% to 79% as an AA judge, and 80% or more would be an AAA rating. This was done with no apparent prior information regarding what typical judges could be expected to score when participating in the procedure. The cut-off values should have been determined after data had been collected from trial runs of the proposed procedure and the proportions of candidates scoring above various percentiles would have been available to estimate what appropriate cut-off or threshold percentiles should be. The committee should have established goals as to which groups of judge candidates they wanted to identify for the three rating classifications, e.g. the upper 10%, the upper quartile, etc., whatever they may have been. Appropriate cut-off values should have then been determined based on data but that was not what was done. Scoring above 80%, or any other arbitrarily chosen percentile, is meaningless because without data it is unknown what proportion of judge candidates would be included in the group scoring at, or above, a particular percentile. My opinion is that an AAA rating should be reserved for judges who are outstanding judges, ones who are at least in the upper 10% of all judges. Perhaps it should be for those in the upper 5%, if we want to indentify the really excellent judges in the breed. Suppose 75% of judge candidates on average score greater than 80% by the judge rating system. If that is the case, an AAA rating based on the arbitrary 80% requirement would be of very little help in identifying the outstanding judges. It would simply identify the judge candidates who are in the upper 75^{th} percentile of candidates. I tried to make the committee and other concerned parties aware of this problem but never got a response from anyone on this very important issue.

I also strongly disagreed with the use of all of the top 10 placings to rate judges, especially with the placings being equally weighted. I contend two things in this regard. The first is that it is the top 3 places that exhibitors care about most. I think it is much more important to exhibitors as to whether they are 1^{st} rather than 2^{nd}, or 2^{nd} rather than 3^{rd}, than it is whether they are 7^{th} rather than 8^{th}, or 9^{th} rather than 8^{th}, etc. The current rating system treats all 10 positions as if they are of equal relevance. My other contention is that it is not humanly possible for anyone to place the top 10 horses in classes according to the breed standard, do it with consistency, and do it all day long. Judges have to concentrate mostly on the top few positions and have to pay less attention to how the individual horses in the bottom of a class are ranked because, first, they know which positions matter most and that it is more important to get them right, and second, there is limited time to get a class placed. In my research on this issue I have always computed judge ratings based on the top 3 positions, the top 5 positions, and all 10 positions. My research has shown that the most effective way to identify judges who do the best job of getting the top of a class placed correctly is to concentrate on the top 3 or 5 positions only. I sent detailed examples to the committee of how using all 10 positions makes it less likely the judges who do the best job of placing the top of a class will be identified, and, as a result I thought it to be counter productive

to consider all 10 places. The reason I recommended rating only the top of a class was so judges who happen to get the bottom placings of classes correct, or nearly correct, by skill or by luck, would not be confused with those who consistently get the top placings in classes nearly correct. But again, there was no response from the committee. No changes were made in the procedure. Thus, all placings, up to 10 when available, are used to score judge performance. I give examples of how this effects judge scores later in my discussion.

The committee approved the rating system by unanimous vote at the 4/6/2010 meeting and the MFTHBA board approved it at their April meeting on April 13, 2010. My objections and concerns were not enough to even slow the process down, let alone change it or have it be tested before being approved.

The rating system was first used at the 2010 Futurity/Spring Show. There were 8 raters used but there were no criteria specified regarding what the qualifications were for those who were selected. There were 6 judges. The two groups combined turned in 135 individual scores. Only 13 of those scores were less than 80% which resulted in 91% of the individual scores exceeding the AAA performance threshold. The criterion as to whether a judge is rated AAA, or not, is based on the average of the scores turned in for a particular judge for the classes he/she judged and this ranged from 4 to 13 classes per person at the Spring Show. All 6 judges and all 8 raters had average scores higher than 80%. All 14, i.e. 100%, would have qualified to be AAA judges. When I indicated that I thought that 80% was too low of a value to use because everyone who was rated by it had passed with an AAA level of performance, I was told that it was because all of the participants were really good judges. The results in the examples given below show that excellent judging skills may not have been the reason why everyone passed at the AAA level. In fact they show one does not have to be a good judge at all to exceed the 80% threshold and be rated a AAA judge by the MFHBA.

Because the average scores of the raters are the basis for what is considered the correct placing of a class, it is crucial that the raters be very competent, experienced, unbiased judges who not only know the breed standard but judge by it and only it, always. They must be persons with true judging expertise. If individuals are chosen to be raters simply because they are readily available, because they have been around and are known (and liked) by those making the selections, because they are assumed to be a person who knows a good horse when they see one, or simply because they are on the Judges Committee or are a carded judge, the strategy of using the average scores of such a group as the basis for evaluating judge performance will be invalid. If raters simply give their opinions based on what they prefer, or what they personally think looks nice or is exciting, rather than give evaluations of how well each horse compares to the MFTHBA breed standard, which they must know inside and out in order to be an expert judge, the judging will be based on what is popular, on politics, on name recognition, or other criteria that will likely be different than the MFTHBA breed standard in the MFTHBA rule book. There was strong evidence that this happened at the 2010 Celebration resulting in a very controversial choice. Not being very selective and very specific regarding who is qualified to be a rater will be the undoing of the current system even if its other weaknesses are dealt with. The average score of a group of persons who lack expertise will never be an adequate substitute for the expertise the group lacks, no matter how many such opinions are averaged over. The importance of who the raters are in the current system seems to have been overlooked by the Judges Committee as there are no qualifying criteria mentioned anywhere, as far as I have been able to determine. The details of how the rating system works, the qualifications necessary for a person to qualify as a

potential rater, the method by which both raters and judges are chosen for a particular show, and the meaning behind what being a A, an AA, and an AAA judge is, should all be completely spelled out and posted on the MFTHBA website. A list of current judges and their ratings should also be kept up to date and online.

After I read the "**Attention in the barns …**" article and found myself linked to it in some way, I decided to take a close look at just how the scoring system itself behaves. Here is what I found it does, assuming there are 10 or more horses in a class. First, an individual value is calculated for each of a judge's top 10 placings in a class according to the following calculation: value = (1 - absval(raters' consensus placement - judge's placement)/10) x 100 (where absval stands for the absolute value or the positive difference), for example suppose a judge placed the raters' 1st place horse 5th, then his value for that placing would be value = (1 - absval(1 - 5)/10) x 100 = (1 - absval(-4)/10) x 100 = (1 - 0.4) x 100 = 0.6 x 100= 60%. Values like these are computed for places 1-10 in each class for each judge and they are averaged for the particular class-judge combinations so that each judge has an average score, based on 10 values, for each of the classes he/she judges. Specifically an individual judge's score for a given class would be:
score = (value1 + value2 + ... + value10) / 10, assuming there are at least 10 horses per class. If a judge places a class correctly, his score is 100%. The criterion for being rated AAA is that the average of all of a judge's scores for the classes he is evaluated on must be 80%, or more.

Consider the examples below and you will see the judge rating system lacks sensitivity to say the least.

Correct placement:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Score**
Judge's placement 1:	1	2	3	4	5	6	7	8	9	10	100%
Judge's placement 2:	10	9	8	7	6	5	4	3	2	1	50%
Judge's placement 3:	5	4	3	2	1	10	9	8	7	6	76%
Judge's placement 4:	5	4	3	2	1	6	10	8	7	9	82%
Judge's placement 5:	1	2	3	4	5	10	9	8	7	6	88%
Judge's placement 6:	5	4	3	2	1	6	7	8	9	10	88%
Judge's placement 7:	4	5	7	3	1	6	2	8	9	10	80%
Judge's placement 8:	3	7	6	1	4	2	5	8	9	10	80%
										Average	***80.8%***

Suppose the scores above represent outcomes from 8 individual classes for a given judge. The judge's final score would be based on the average of the class scores given to the right of the table. The highest possible score is 100% (class1) and the lowest possible score is 50% (class2). I generated 10,000 random permutations of the numbers 1 to10 and computed scores using the random permutations as placings in the formula given above to determine the median score for someone who guesses at all 10 positions for all classes judged. The median score for complete guess work is 66%. This means that half of the random guesses were below 66% and half were above 66%. Sixty-six percent is what we could expect someone who knows nothing about horses to average by just filling out a judge's cards with their guesses. This means that for a person to score 80% they have to know just enough to get some of the better horses in the upper half of their placings and some of the lower ranked horses in the lower half of their placings. They need enough of an idea of what they are doing to score 14 percentage points better than someone who would simply guess at every position. We can see by looking at the scores for classes 4 through 8

above that the placement of a class certainly does not have to be very close to the correct placing in order to produce a score of 80% or better. In light of the above examples, we have to ask, what value does the AAA rating has? It clearly is not an indication that a person is an outstanding judge. If the above results typify what an AAA judge's performance looks like, it is hard to imagine what just an average or a poor judge's performance must be.

In the above examples, the judge placed the top 3 horses correctly in only 2 of 8 classes yet he/she would have scored enough (80.8%) to be rated an AAA judge. In my opinion, only one class was placed well, the first one, one was placed fairly well (class 5) and the rest were completely blown, yet this very weak performance would be rated AAA, the highest rating possible for a judge's performance. Note that classes 5 and 6 both have scores of 88%. In my opinion class 5 is placed much better than class 6. It is only because the bottom of class 6 was placed correctly that the score is 88%. This is why I think it is a bad idea to consider all 10 positions when comparing judges. Doing so makes it less likely that judges who place the top of classes well will be differentiated from those who do not, as the comparison of class 5 to class 6 clearly demonstrates. In this example if the two 88% scores were scores from two different judges, the conclusion would be that they are equally qualified. But which judge would you want to show under if the scores represented the judges' average performances?

To make matters worse, it was recently approved that judges will no longer have to renew their judge's card by taking and passing the usual exam. All they have to do now is place the classes specified at a Futurity/Spring Show or at a Celebration and get at least an 80% average score. They will then be a qualified, AAA judge who presumably will be qualified to judge any show in the breed.