

# A FAIRER INTERNATIONAL Scoring Method?

By Jackie Fie and Lance Crowley

In response to requests by members of the gymnastics technical community and to satisfy our own curiosity, a study was undertaken to determine if there was a way to improve on the current method used to determine the average score with a six-judge panel. I offer the following to the entire gymnastics community for consideration.

As you are aware, the method of arriving at the final score for a competition/event in which a six judge panel is used is to eliminate the high and low score and average the middle four scores (Avg Mid 4).

Since 1990 I have analyzed scores from numerous international competitions while working with, and developing, the *WTC Judges' Evaluation System*. During these studies I have often questioned the validity of the "Avg Mid 4" method. As an example of this questioning, consider the following "set" of six scores:

9.8 — 9.8 — 9.8 — 9.8 — 9.65 — 9.6

The average score from this "set", with the present method, would be 9.7625. Actually the final score would be 9.762, but that's a subject of another dissertation. When I study these scores, I think the final score should be 9.8. Four of the six judges, 2/3's of the panel, thought the score was 9.8. So why isn't the score 9.8?

To try to answer this question let's start with some basic thoughts on statistics. Statistics teaches us that, in a *subjective judgment by a number of equally qualified observers*, each of the judgments carries the same weight, has the

same validity. It is for this reason that I think the score in the above example should be 9.8. The fact that they (the judges) may not be equally qualified cannot be established at the moment of judgment; it takes the accumulation of a significant number of "sets" to make that determination. That being the case, why do we arbitrarily throw out the high and low score? On what scientific basis are the high and low scores deemed to be incorrect? The answer to these questions may go back to the days before computer scoring, it represented the easiest method of arriving at a final score, quickly.

I submit that the "Avg Mid 4" is arbitrary and is not based on logic or science. Given that our scoring computers can calculate scores instantly, why not consider other methods?

Over the years there have been several proposals for different methods to use the six judgments:

- It has been proposed that we use the average of the middle two scores. The concern with this method is that you lose the expert opinion of two judges. It also, by its nature, generates many more ties (a very real problem).
- Using all six scores has also been discussed. The concern with this method is that the judges are human beings and make honest mistakes. To affect the ranking of an athlete because of a mistake is wrong, in my opinion.
- The method that averages the Chair's score with the score from the "Avg Mid 4", the so-called "base score" has been used in special circumstances.

Given the events of the past Winter Olympic Games, it is appropriate to address the issue of cheating, particularly considering the shortcomings of the current method and the ease with which it can be changed, for the better. No one likes the term cheating, so allow

me to borrow a term coined by Dr. William Sands, "human engineered scores."

The "set" example used above, 9.8 — 9.8 — 9.8 — 9.8 — 9.65 — 9.6, can be arrived at by a panel of judges for one of several reasons. One, it could be their actual subjective scores for that particular exercise. Two, it could be that the judges with the 9.65 / 9.60 were working together to give that gymnast a lower score than she'd earned, i.e. "human engineered scores." What you may find is another "set" of scores where these same judges were the two high scores on some pre-selected, athlete. Again, while very few like to talk about it, two judges working together to "hit" one gymnast and "bump up" another gymnast is the most common method of "human engineered scores."

There have been proposals that increased the number of judges on a panel. That is not acceptable due to many considerations, not the least of which is the cost of the travel and housing for the judges.

Given all of the above, I offer the following scoring averaging method for consideration. The method/system is called the "Most Significant Method" (MSM). This averaging method was written into the *WTC Judges' Evaluation System* in early 1999, and has been used to analyze the results for a significant number of international competitions.

Using the six judges scores, the method starts the calculation by sorting them into descending order. The average of the middle two scores is calculated. It then uses these two scores (the middle two) plus the next two scores closest to this average. It then averages these four scores, to arrive at the final score. Thus, the two scores farthest from the average of the middle two scores are eliminated.

There are only three possible combinations of *rejected* scores; this system selects which of the three fit the statistical model for each "set":

1. The high and low scores (same as the current method)- OR
2. The two high scores — OR
3. The two low scores.

The actual calculation method is a bit more complex. The software must look up or down the sorted order based on the direction of the third score selected, in order to determine which score to use in the fourth position.

- If the third score is in the high direction: the system first looks in the low direction to determine if there is a score that is equal to or less than the next score in the high direction.
- If the third score selected is in the low direction: then the system looks to the high side to determine if that score is equal to or less than the next score in the low direction.

Again, the system will select which of the three options, listed above, fits each particular case.

Why is this system better than the current method? First, it has a scientific/mathematical base and more precisely reflects the statistically correct opinion of the judging panel. Using the above example, the final score would be 9.8. The other major advantage is that "human engineered scores" become a much more difficult task. It now requires three judges working together to affect a score. I submit that this is extremely difficult, for several reasons:

*(continued on page 10)*



PHOTOGRAPHY © JIM KELLY

(continued from page 7)

- There are very few judges who “human engineer scores.”
- By some remote chance, if three judges “human engineer scores” on the same event, the coordination required to affect the final score would be difficult.

It is hoped that once the judges, coaches and national officials understand the difficulty of “human engineering scores” using this system, the judges will simply submit their honest judgment. We will then be rid of a problem that has plagued us for years.

I submit that this method –

- Has a sound scientific/mathematical basis
- It is easy to implement, no changes are required to the structure of existing judging panels
- Costs very little to do, no more than an hour or two of software programming
- Will make scoring much fairer to our athletes, and
- May prevent the type of publicity that has hit figure skating recently.

Analyses of international competitions show significant differences in results. While analysis is interesting, there is no “gold standard” on which to base the *absolute correct* score/rank for a given routine. For this reason, the *scientific and logical merits* of this system must be the deciding factor as to whether or not it is better than the current method. I submit that it is, in fact, a significant improvement.

Anyone who would like to discuss additional details of this method, please contact Lance Crowley via email at [ipcrowley@worldnet.att.net](mailto:ipcrowley@worldnet.att.net).

**Note:** The above method is now referred to as MSM6. A similar method for 4 judge panels MSM4, has been submitted to FIG, USAG and the USECA for publication.



## An Improved Method for SCORE AVERAGING WITH A FOUR-JUDGE PANEL?

### INTRODUCTION

In response to requests by members of the gymnastics technical community and to satisfy our own curiosity, a study was undertaken to determine if there was a way to improve on the current method used to determine the average score with a four-judge panel. It was thought that the same general principles discussed on page 6, 7, and 10 for an improved system for six judge panels could be applied to four judge panels. The proposed change to the six judge panel score averaging method is referred to as the “Most Significant Method” (MSM). For purpose of this discussion that system will be renamed MSM6 and this newly proposed four judge panel system, MSM4.

It must be clearly understood that there is no system/method that will ever allow a four-judge panel to be as effective as a six-judge panel, regardless of the score averaging technique used. That said, this study indicates that significant improvements can be made to the four-judge standard method. (STD-4).

In addition to the standard method (STD-4) of averaging the middle two scores, these alternatives were considered:

- Average all 4 scores
- Average the closest 3 scores (using the MSM method)
- Preferred MSM4-1 (explained below)
- Non-preferred MSM4-2

The WTC Judges’ Analysis was amended to generate results for each of the above score averaging scenarios. The statistical analysis was done in Excel. Well over 3000 panel judgments were used from 3 World Championships and 2 Olympic Games. However, the primary studies were done using the C-I scores from all four events from the 2001 World Championships. The middle four “sorted” scores were used to calculate results for the STD-4 and four proposed methods. The following statistics were tabulated:

- Correlation coefficients against the MSM6 scores.
- Summary statistics (average, standard deviation, maximum, minimum, etc.).
- The number of times scores were unchanged, higher, or lower.
- Number of tied scores.

#### EXPLANATION OF THE PREFERRED MSM4-1 METHOD

The calculation process starts by sorting the four scores into descending order. To make this explanation easier to understand, these scores are labeled: **H** = the highest of the four, **HM** = the high middle of the four, **LM** = the low middle of the four and **L** = the lowest of the four.

The first two scores used to calculate the final average score are the two middle scores, **HM** and **LM**. The third score selected, is the one closest to average of the **HM** and **LM**. If the **H** and **L** are equidistant from the middle average, the final score is simply the average of the two middle scores (the STD-4 method). If the **H** score is selected, that score is averaged with the **LM** score, and then that average is averaged with the **HM** score. If the **L** score is selected, then it is averaged with the **HM** score, this result is then averaged with the **LM** score.

Consider the following judge's scores and averages (STD-4) from a recent International competition that used four-judge panels:

	J1	J2	J3	J4	AVG
Gymnast #1	<del>9.6</del>	<del>9.0</del>	9.35	9.35	=9.35
Gymnast #2	9.3	<del>8.8</del>	<del>9.5</del>	9.4	=9.35

Using the preferred MSM4 method, the Final Scores would be:

**Gymnast #1      9.6      9.0      9.35      9.35**

- Sorted into descending order: 9.6 9.35 9.35 ~~9.0~~
- Average of middle two scores = 9.35
- 9.6 – 9.35 = 0.25 and 9.35 – 9.0 = 0.35, thus the 9.6 is closest to the average of the middle two scores and will be used in the calculation, the 9.0 will be dropped.
- $(9.6 + 9.35) / 2 = 9.475$
- $(9.475 + 9.35) / 2 = 9.4125$ , the Final Score

**Gymnast #2      9.3      8.8      9.5      9.4**

- Sorted into descending order: 9.5 9.4 9.3 ~~8.8~~
- Average of the middle two scores = 9.35
- 9.5 – 9.35 = 0.15 and 9.35 – 8.8 = 0.55, thus the 9.5 is closest to the average of the middle two scores and will be used in the calculation, the 8.8 will be dropped.
- $(9.5 + 9.3) / 2 = 9.4$
- $(9.4 + 9.4) / 2 = 9.4$ , the Final Score

The above formula is:  $((H + LM) / 2 + HM) / 2$ . If the H score is discarded, the formula is:  $((L + HM) / 2 + LM) / 2$ .

#### CONCLUSION

As pointed out in the previous paper, there is no “gold standard” for a correct score. Gymnastics is subjectively judged, thus is subject to all the nuances associated with that fact. Since the MSM6 has been shown to generate the best possible results for a six-judge panel, the MSM4 study used results from MSM6 as the basis for the analysis and comparison.

#### PREFERRED MSM4-1 VS. NON-PREFERRED MSM4-2

Of the four possible methods studied, the two MSM4 methods are the best and very similar in overall effectiveness. However, the final decision to use the MSM4-1 method was based on a significantly reduced number of ties and a correlation coefficient that was only slightly lower than the MSM4-2. There is no significant difference between the average score of the MSM4-2 method and the STD-4 method, thus there would be no significant changes to the average score of a competition expected.

The major benefits derived from the preferred MSM4 method, over the STD-4 method are:

- Significantly reduced number of ties, thus the method is more discerning. Check this with Excel!  
Using as an example the UB from the 01 WC – 151 judgments, the STD-4 method generated 60 ties, nearly 40% of all the scores. The MSM4-1 method generated 30 ties, nearly a twofold reduction.
- High correlation to MSM6 with final rank placement of the competitors
- Reduced possibility of “human engineered scores” since the methodology makes it more difficult to guess which will be the counting scores.

We offer this method to the gymnastics community for their consideration; we believe it is a major improvement over the current method. Anyone who would like to discuss additional details of this method, please contact Lance Crowley via email at [ipcrowley@worldnet.att.net](mailto:ipcrowley@worldnet.att.net).

#### NOTE

*Paper emailed to FIG President Bruno Grandi, the FIG Executive Committee, and the FIG Women's and Men's Technical Committees on July 6 and 7, 2002, titled, “A Fairer International Scoring Method?”*

*An edited version was published in the November issue of International Gymnast magazine, titled “The Best Average” (page 34).*