The Fairest of Them All: Using Variations of Beta-Binomial Distributions to Investigate Robust Scoring Methods

Mary Good - Bluffton University
Christopher Kinson - Albany State University
Karen Nielsen - University of Oklahoma
Malin Rapp-Olsson - University of Arizona
Mame Fatou Thiam - University of Wisconsin River Falls

Abstract

Contestants in subjective competitions often encounter the issue of unfair or inconsistent judging. Scores are awarded to contestants and winners are determined, but how do we know that the best contestants win? By assuming there exists a well-defined order of the contestants' abilities, we can explore which scoring method best captures the true order of the contestants. We use variations of beta-binomial distributions to model the contestants' abilities and the judges' scoring tendencies. We use Monte Carlo simulations to investigate seven scoring methods (mean, z-scores, median, raw rank, z rank, Olympic, and outliers) and determine which method, if any, works best for various judge panels. We apply our model to the scenario of a scholarship competition with 20 contestants, where the top three contestants receive equal monetary awards.

Keywords: beta-binomial, fairness, judge, Monte Carlo, rankings

1 Introduction

The question of unfair and inconsistent judging is always present in subjective contests. Before some competitions, judges receive training on how to score contestants consistently. These calibration exercises aim to promote fair contests. However, even with training, judges may still have personal biases and different scoring standards during events. As a result, scores from subjective events may not identify the strongest contestants. Assuming that contestants can be ordered from strongest to weakest, we want to identify scoring methods for which the strongest contestants are most likely to win.

Without a set order of best contestants, we can only attribute variations in judging to arbitrary preferences. Gordon and Truchon [4] explain the significance of assuming a situation

in which there exists a true ordering that judges may perceive differently. Several factors account for variability in scores. In addition to the true quality of the contestants, judges' past experiences [2] and judges' personal preferences [5] can add variability to scores.

To approach this problem, we use Monte Carlo methods to simulate subjective contests. We characterize the contestants by randomly sampling from a beta distribution and use these values to create variations of beta-binomial distributions for judges' scores. We simulate judges' scores from these distributions and compile them using seven different scoring methods. We compare the abilities of the scoring methods to select the strongest contestants, and we make recommendations.

2 Examples

Olympic figure skating competitions, scholarship awards, and the job employment process are just a few examples of subjective competitions.

Olympic figure skating competitions use a system that takes into consideration the elements of a specific figure skating performance. Nine judges are randomly selected from the twelve total judges and the highest and lowest scores are removed. The remaining seven scores are averaged.

The National Science Foundation (NSF) awards fellowships by reviewing applications in a subjective manner. Not all judges review all contestants. Instead, two or more judges score each application according to intellectual merit and potential societal impact. Recently, the NSF has standardized judges' scores by converting them to z-scores.

The Marine Corps Logistics Base of Albany, Georgia uses an interview panel for hiring purposes. The panel looks at job applications and gives candidates scores from 0 to 100. The panel sums the scores of each candidate, and then divides by the number of members on the panel. Only a certain number of candidates advance to the interviewing stage.

3 Modeling the Problem

To investigate some of the inherent problems in subjective judging, we develop the following model situation. We assume there are 20 contestants with predetermined ranks competing for three equal prizes. They are ordered with **contestant numbers** such that contestant 1 is assumed to be the best, contestant 2 is assumed to be the second best, and so on.

We assume there is a panel of 10 judges. Each contestant receives integer scores ranging from 0 to 40 from only 5 of the 10 judges. All 5 judges for each contestant are selected at random and without replacement. Each of the 10 judges evaluates $5 \times 20 \div 10 = 10$ different contestants. This represents a contest in which there are too many contestants for every judge to score each contestant.

4 Describing the Contestants

As mentioned previously, the twenty contestants have predetermined rankings. Naturally, the abilities of the contestants are not equally spaced. Thus we place a prior distribution on

the contestants in the following way. Let $X_1, X_2, \ldots, X_{20}$ be a random sample from a beta distribution with $\alpha = 4$ and $\beta = 1.3$. The probability density function of this distribution is shown in Figure 1. The mean of this distribution is $\mu = 0.755$. From this sample, we assign the largest order statistic to contestant 1, the second largest order statistic to contestant 2, and so on. This defines **contestant values** $C_1, C_2, \ldots, C_{20}$, where $C_i = X_{(21-i)}$. For example, $C_1 = X_{(20)}$, the largest order statistic. This defines random spacings between contestants. The parameters for the beta distribution are deliberately chosen so that the contestant values are left skewed. This is an effort to define a contestant population that is particularly skilled, because our competitors are probably above average. For example, in Cliff and King's study of wine judgment, the scores were skewed left, suggesting that the wines submitted to the contest were above average [2].

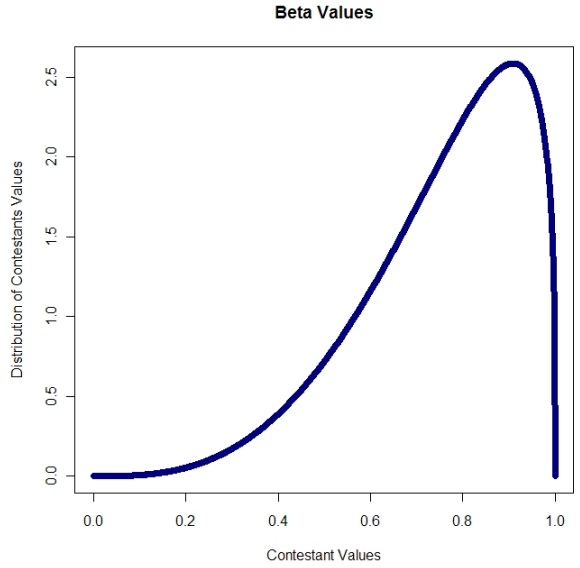

Figure 1 : Contestant values are selected from the Beta(4, 1.3) distribution to reflect an above average group of contestants.

5 Describing the Judge Profiles

We consider four types of judges that we describe as **fair**, **lenient**, **harsh**, and **moderate**. We assume that all judges attempt to order contestants correctly. The judges are modeled by various binomial distributions, with the parameter n usually being 40, the highest score that can be given, and p being a function of the contestant values. The score that a contestant is assigned depends on both the quality of the contestant and the type of judge assigned to the contestant.

We used variations of beta-binomial distributions to model judges' scores based on preliminary findings. Before coming to this decision, we considered some basic questions. First, would we use a discrete distribution or a continuous one? Would the scores have upper and lower limits? Knowing that each judge gives a score between 0 and 40, we chose a distribution that models the situation appropriately. Among the various distributions we explored, such as exponential, chi-square or Poisson, the binomial worked the best. Not only can it re-

strict the scores to be integers between 0 and 40, but it consistently assigns realistic scores to contestants. Scores from chi-square and exponential distributions proved to be too variable to be suitable to our situation, even after rounding and truncating. Although discrete like the binomial, the Poisson distribution generates less consistent scores. Also, Cliff and King [2] suggest that some judges appear to be using an "accept/reject" scoring system, which led us to believe that a binomial distribution best models such judges.

With the binomial distribution, we can model several different types of judges by adjusting the parameters. For example, the scoring behavior of a fair judge is represented by a binomial distribution where p is the contestant value for that particular contestant. Scores for a fair judge for the i^{th} best contestant come from Bin(40, C_i).

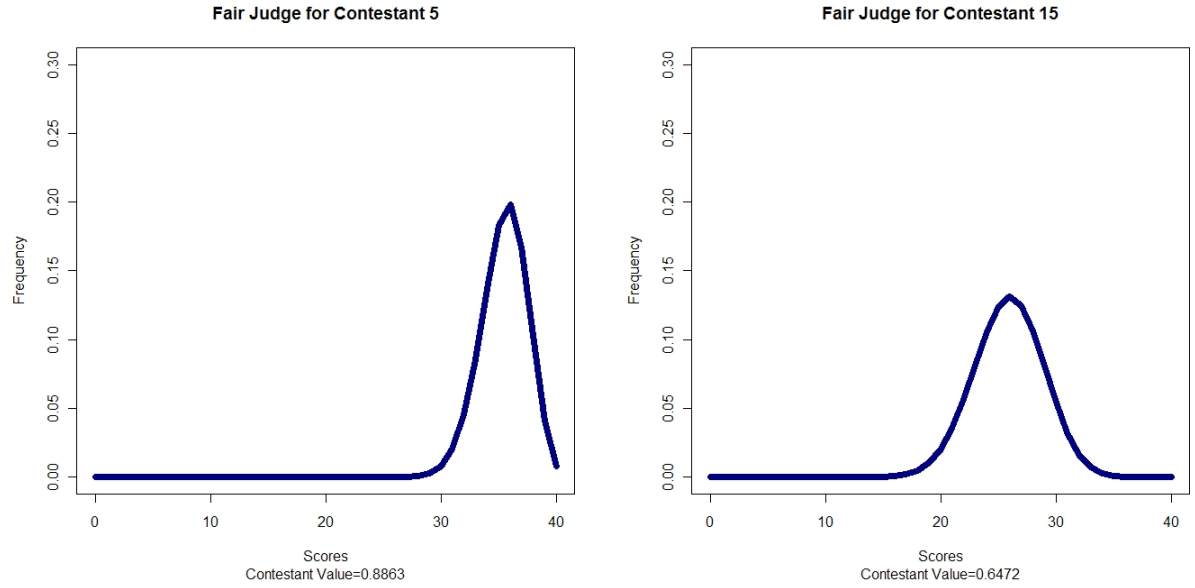

Figure 2 : Distributions of scores given by a fair judge to contestant 5 and contestant 15 when $c_5 = .8863$ and $c_{15} = .6472$.

The more lenient judges are modeled slightly differently, although the inherent quality of the contestant still influences the score. One lenient judge gives scores from the distribution $\text{Bin}(20, C_i) + 20$, where C_i is the contestant value for the i^{th} best contestant. As seen in Figure 3, this judge never gives a score below 20 points. Another lenient judge's scores are modeled by adding 5 points to a fair judge's scoring distribution. We truncate if necessary so that no score is above 40. This judge gives scores from 5 to 40 and is likely to give each contestant a higher score than the fair judge would.

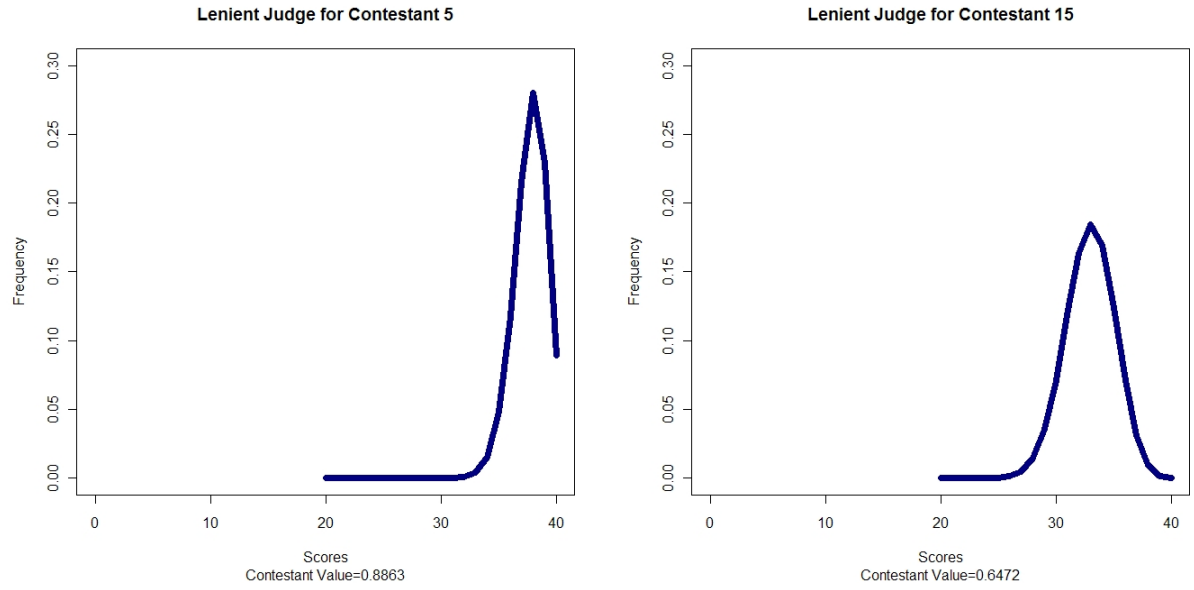

Figure 3 : Distributions of scores given by a lenient judge to contestant 5 and contestant 15 when $c_5 = .8863$ and $c_{15} = .6472$.

The harsher judges are modeled similarly. One type of harsh judge only gives scores up to 33 points. This is modeled by $\text{Bin}(33, C_i)$. Another harsh judge, with scores modeled by $\text{Bin}(40, 0.75 \times C_i)$, can still give scores up to 40 points, but with decreased likelihood.

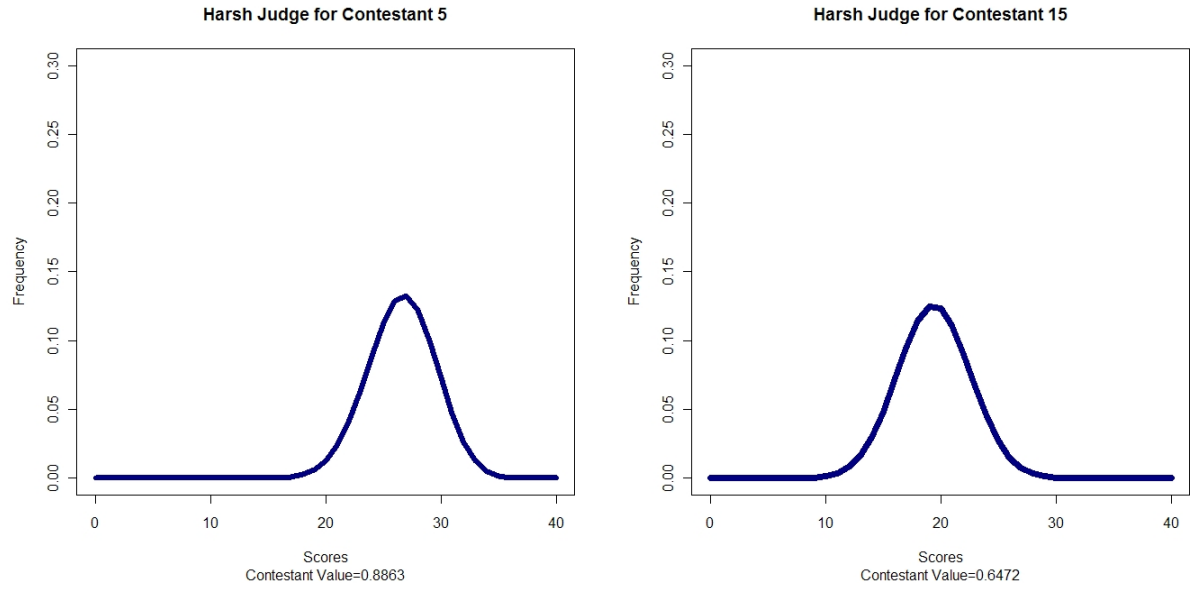

Figure 4 : Distributions of scores given by a harsh judge to contestant 5 and contestant 15 when $c_5 = .8863$ and $c_{15} = .6472$.

Cliff and King [2] found that some judges avoid giving extreme scores. We call these moderate judges. One such judge is represented by the scoring distribution $\text{Bin}(30, C_i) + 5$. We represent another judge's scores by a binomial whose parameter $p = C_i + (1 - C_1)/1.01$.

We add the last term to each order statistic to increase the mean of the binomial distribution while keeping p below one. This differs only slightly from a fair judge because of the left-skewed beta distribution used.

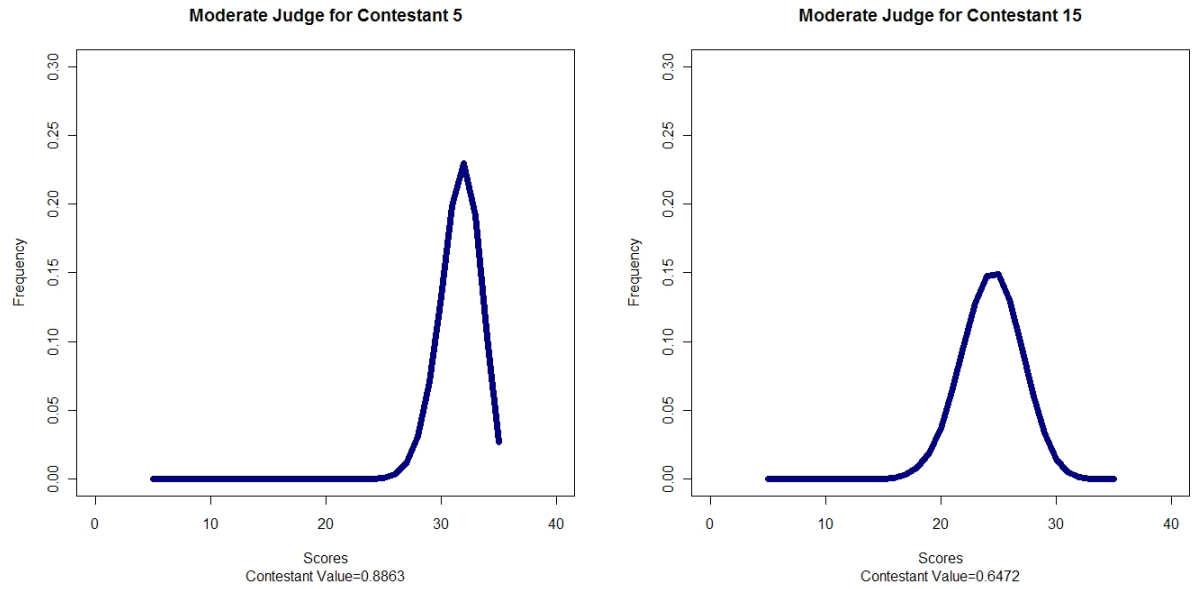

Figure 5 : Distributions of scores given by a moderate judge to contestant 5 and contestant 15 when $c_5 = .8863$ and $c_{15} = .6472$.

In each case, a variation of a beta-binomial distribution represents the judge's scoring behavior. Typically, a contestant receives a higher score from an easier judge. Also, on average, a judge awards higher scores to better candidates. This reflects our assumption that all judges attempt to order the contestants correctly, but simply have different standards when awarding numerical scores.

6 Scoring Methods

After each contestant receives scores, we use seven scoring methods to compile the scores into seven different sets of composite scores. For each of these, we can rank the contestants and compare this experimental rank to the predetermined rank. We call the seven scoring methods mean, z-scores, median, raw rank, z rank, Olympic, and outliers.

The **mean** and **median** scoring methods respectively take the mean and median of the scores for each contestant.

The **Olympic** and **outliers** methods report composite scores after removing specific types of scores. The Olympic method first removes each contestant's highest and lowest scores, and then returns the average of the remaining scores as the composite score. This mimics a scoring method used in Olympic figure skating events described earlier.

The outliers method first finds the mean score for each contestant. The method then removes any scores five or more points away from each contestant's mean score. After removing any such scores, the average of the remaining scores serves as the composite score

for the contestant.

The **z-score** method expresses each raw score in terms of its distance above or below the mean. We determine the sample mean and standard deviation for each of the judges by looking at the complete set of scores a particular judge gives. Then we calculate the z-scores by using the formula $(y_{i,n} - \bar{y}_n) \div s_n$, where $\bar{y}_n$ and s_n are the sample mean and sample standard deviation respectively of the n^{th} judge's scores. Also, $y_{i,n}$ is the i^{th} score given by the n^{th} judge. For example, a z-score of 1.5 represents a score 1.5 standard deviations above a judge's mean.

The **z rank** scoring method expands on the z-scores method by going another step. After converting all raw scores into z-scores, the z rank method assigns a ranking to the 100 scores given in the contest, where 100 denotes the best possible rank. Upon ranking all of the z-scores, the z rank method returns the average of each contestant's rank scores as the composite score.

A similar method, **raw rank**, orders the original scores rather than z-scores. Again, each score receives a rank from 1 to 100 with 100 being the best. Alvo and Cabilio [1] point out the need to account for potential ties in research design. In instances of ties, the rank score becomes the mean of the ranks for which the scores tie. The composite score for the raw rank method uses the average of each contestant's ranked scores.

7 Our approach

We use the program R to run Monte Carlo simulations with a large number of trials. We begin by defining the characteristics of the panel of 10 judges. For example, we may select five fair judges, two lenient judges, two harsh judges, and one moderate judge from the various judge profiles of these types. Then we define the contestants in terms of their contestant values, selected from Beta(4, 1.3).

Next, we randomly assign judges to contestants. Recall that only five of the ten judges score each contestant. We then generate the scores given in each judge-contestant pairing. Keep in mind that these scores depend on the harshness of the judge, the strength of the contestant, and the unique beta-binomial distribution created for each judge-contestant pair. After all scores have been assigned, we note which of the seven scoring methods are successful.

We consider a scoring method successful if it correctly places the predetermined best three candidates (contestants 1, 2, and 3) in the top three positions, in any order. This models scholarship competitions where several of the best candidates receive the same award. A tie between the 3^{rd} and 4^{th} place contestants also counts as a success.

Our use of the Monte Carlo method demands that we run many simulations in order to generate meaningful results. While keeping the same judge assignments and the same contestant values, we generate scores multiple times and count the successes for each scoring method. After 1000 repetitions of score generation, we hope our success counts will inform us about the best scoring methods.

So far, we have described simulating the same contest 1000 times. Next, we create a new contest with the same judge profiles by selecting new contestant values from the Beta(4, 1.3) distribution and assigning judges to contestants again. We again generate scores 1000 times and count the successes for this contest. One time through this cycle is called an **outer trial**. An outer trial is illustrated in Figure 6. In order to discover the best scoring methods,

we repeat the outer trial 1000 times and compile the results.

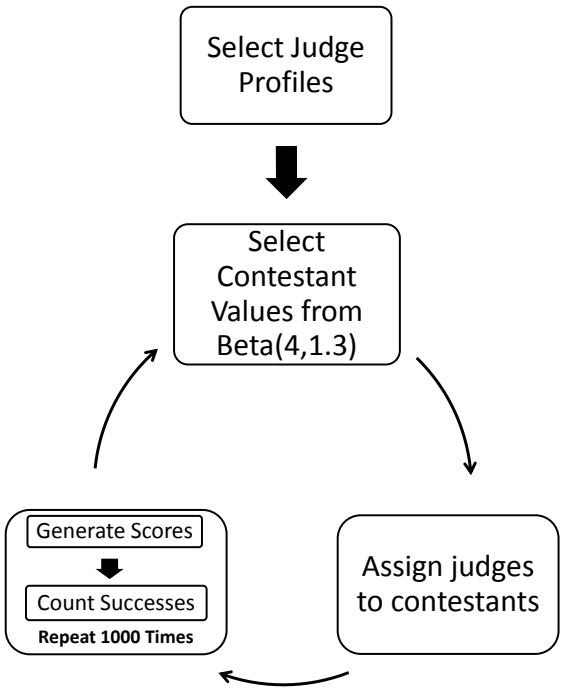

Figure 6 : Flow chart of simulation. The circular path represents one outer trial. Within each outer trial, scores are generated 1000 times. Outer trials are also repeated 1000 times.

8 Evaluating the Scoring Methods

One way to measure how well each scoring method performs is to compute the percentage of times the method is a success over the entire simulation. However, whether a scoring method is deemed successsful is relative rather than absolute. For example, if the first method is successful only 200 out of 1000 times but the other methods are successful only 100 out of 1000 times, then the first method clearly is the best for that outer trial.

With this in mind, we also compare the methods by ranking them from 1 (fewest successes) to 7 (most successes) for each outer trial and looking at the averages of these rankings across the total number of outer trials. This manner of analysis awards the most value to the best method of each outer trial, but still gives the second best method a high value as well.

In order to analyze the scoring methods more deeply, we consider how far each method's results stray from the ideal. Since we are interested in getting the three predetermined best contestants in the top three placements, we compute the average of the contestant numbers that each method places in the top three. Ideally, a method places the top three in the top three in any of the orders: {1,2,3},{1,3,2},{2,1,3},{2,3,1},{3,1,2} or {3,2,1}. In each of these cases, a contestant can be tied with one or both of the other contestants in the top three. The corresponding average in any of these cases is 2.0 and thus represents the best

possible average. Suppose that one simulation places contestants in the order: 2, 1, and has 3 and 9 tied for the third place. Then the corresponding average for the top placements is $(2 + 1 + 3 + 9) \div 4 = 3.75$.

We compare the scoring methods by looking at these averages across the total number of outer trials. The larger the average of the top three contestant numbers, the less accurate the method. For example, if a method determines contestants 1, 2, and 4 (average = 2.33) to be the winners, it is more accurate than a method that determines contestants 2, 3, and 4 (average = 3) to be the winners. This allows us to gauge which method reports scores with the greatest accuracy, whether the trial constitutes a success or not.

Likewise, we consider how far the best three contestants are from winning with each scoring method. Specifically, we take the mean of the placements of contestants 1, 2, and 3. If contestants 1 and 2 are placed first and second respectively, and contestant 3 is placed fifth, then the average placement of the three strongest contestants is $(1 + 2 + 5) \div 3 = 2.67$. Again, we want methods with an average placement of the best contestants as close to 2.0 as possible. This analysis measures how close a method comes to getting the most deserving contestants in the top three placements, whereas the previous analysis deals with the predetermined rank of the winning contestants.

9 Results

To be able to determine which scoring methods are superior, we must examine how they compare when there are several different combinations of judges on the panel. We use six different panels of judges for our investigations (see Table A in Appendix for details). We use ten fair judges first because this is what competitions strive for when they use calibration techniques. Likewise, we investigate other situations where all judges are similar in nature, such as panels of ten lenient judges and ten harsh judges. The next situations are perhaps more realistic because they feature a variety of judge types. The first has two lenient, two moderate, one harsh, and five fair judges. Next we want a panel with more harsh judges than lenient ones, so we choose one lenient, two moderate, two harsh, and five fair judges. Conversely, a situation with fewer harsh judges than lenient ones has two lenient, two moderate, one harsh, and five fair judges.

For each panel of judges, we begin with the first comparison method listed in Section 8, percentage of successful trials. Table 1 shows the percentage of successes for each judge panel.

Percentages of Successful Trials for Various Judge Panels

	10 Fair	10 L	10 H	2L/2M/2H	1L/2M/2H	2L/2M/1H
Mean	***63.63	*45.96	*36.88	28.85	38.60	32.13
Z-Scores	49.97	31.24	31.91	**42.87	*46.44	**45.63
Median	*62.79	***55.10	***39.13	*38.91	***48.17	*42.15
Raw Rank	59.87	41.87	32.23	26.79	35.64	28.84
Z Rank	50.88	31.95	32.57	***43.96	**47.14	***46.74
Olympic	**63.20	**46.80	**38.48	32.93	42.19	35.98
Outliers	62.40	44.78	34.17	29.37	35.85	32.77

Table 1 : Percentage of successes of methods for six panels of various combinations of Lenient (L), Moderate (M), and Harsh (H) judges.
Any remaining judges out of 10 are fair.
***=highest percentage **= second highest percentage * = third highest percentage

Note that the percentages of success are highest when all judges are fair. In all three cases where the judges have the same type of profile (10 fair, 10 lenient, 10 harsh), the methods with the most successes are mean, median, and Olympic, and the methods with the fewest successes are z-scores, z rank, and raw rank.

However, when the panel consists of different types of judges, the z-scores, z rank, and median scoring methods have the most successes. The scoring methods with the fewest successes are raw rank, outliers, and mean.

Recall our second way of comparing the scoring methods, in which we average the rank of each method. The results from this comparison method are in Table 2.

Average Rank of Methods for Various Judge Panels

	10 Fair	10 L	10 H	2L/2M/2H	1L/2M/2H	2L/2M/1H
Mean	***5.89	***5.39	*4.75	3.57	3.79	3.59
Z-Scores	2.42	2.74	3.31	**4.69	*4.37	**4.68
Median	*4.43	**5.36	***5.51	*4.62	***5.03	*4.65
Raw Rank	3.09	2.82	2.22	2.73	2.77	2.57
Z Rank	2.59	2.91	3.63	***5.05	**4.66	***5.02
Olympic	**5.21	*4.66	**5.48	3.97	4.26	4.03
Outliers	4.37	4.12	3.10	3.36	3.13	3.46

Table 2 : Average rank of methods for six panels of various combinations of Lenient (L), Moderate (M), and Harsh (H) judges.
Any remaining judges out of 10 are fair.
***=highest average rank **= second highest average rank * = third highest average rank

Again, the mean, median, and Olympic methods rank the best for the panels of ten similar judges. When there is a variety of judges on the panel, z-scores, z rank, and median still rank the best.

To further analyze our scoring methods, we compare the average contestant number placed in the top three and the average placement of contestants 1, 2, and 3 by each method, as seen in Table 3 and Table 4. Recall that low values are desired in these analyses.

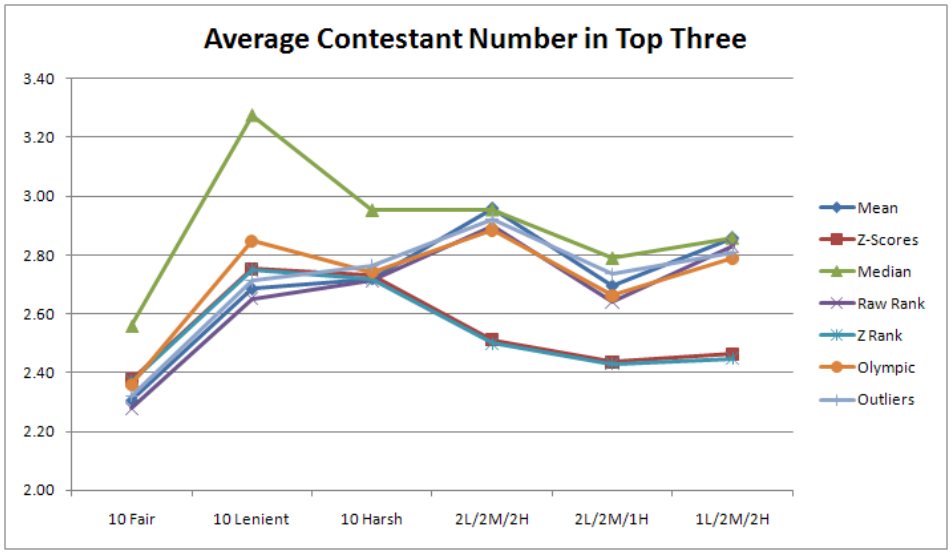

Figure 7 : Data from Table 3 with Lenient (L), Moderate (M), and Harsh (H) judges. Any remaining judges out of 10 are fair. Note that the most accurate scoring method changes when judge panels are diversified.

Average Contestant Number in Top Three for Various Judge Panels

	10 Fair	10 L	10 H	2L/2M/2H	1L/2M/2H	2L/2M/1H
Mean	2.31	2.68	2.72	2.96	2.69	2.86
Z-Scores	2.38	2.75	2.73	2.51	2.44	2.46
Median	2.56	3.27	2.95	2.95	2.79	2.86
Raw Rank	*2.28	*2.65	*2.71	2.90	2.64	2.83
Z Rank	2.37	2.75	2.72	*2.50	*2.43	*2.45
Olympic	2.36	2.85	2.74	2.89	2.66	2.79
Outliers	2.32	2.71	2.76	2.92	2.73	2.81

Table 3 : Average of contestant numbers in top three for six panels of various combinations of Lenient (L), Moderate (M), and Harsh (H) judges.Any remaining judges out of 10 are fair.
*=lowest average contestant number

Average Placement of the Best Three Contestants for Various Judge Panels

	10 Fair	10 L	10 H	2L/2M/2H	1L/2M/2H	2L/2M/1H
Mean	*2.27	*2.59	*2.69	2.92	2.64	2.83
Z-Scores	2.37	2.74	2.72	2.51	2.43	2.46
Median	2.39	3.13	2.84	2.90	2.63	2.80
Raw Rank	*2.27	2.61	2.71	2.88	2.62	2.82
Z Rank	2.37	2.73	2.71	*2.49	*2.42	*2.44
Olympic	2.30	2.74	2.71	2.85	2.60	2.76
Outliers	2.28	2.62	2.77	2.99	2.70	2.87

Table 4 : Average placements of the best three contestants for six panels of various combinations of Lenient (L), Moderate (M), and Harsh (H) judges. Any remaining judges out of 10 are fair.
*=lowest average rank of the best three contestants

These analyses measure method accuracy by recording the average contestant number of candidates placed in the top three and the average placement of the best three contestants. For example, when a panel consists of ten harsh judges, the median has the most successes (Table 1). However, the average contestant number placed in the top three by the raw rank is the lowest (Table 3), and thus the best by this criterion. This suggests that for ten harsh judges, even when the raw rank method does not succeed, it is overall closer to getting the top three in the top three placements than the other methods. When the placements of the best three contestants are averaged, the mean scoring method produces the lowest results.

Out of the three methods (mean, median, and Olympic) that produce the most successes and have the highest ranks when all judges have similar scoring behavior, the mean method has the lowest (best) average contestant number (Table 3) and lowest average placement of the best three contestants (Table 4).

Of the three methods (z-scores, z rank, and median) that give the highest number of successes and have the highest ranks when the judges have different scoring behaviors, the z rank method has the lowest (best) average contestant number (Table 3) and lowest average placement of the best three contestants (Table 4).

10 Conclusion

After exploring our results, we find two patterns for the six judging panels based on whether we have judges all of the same type or a variety of judges. When the panel consists of just one type of judge, we recommend the mean as the best scoring method. Part of this recommendation comes from the fact that in all three cases (10 fair, 10 lenient and 10 harsh), the mean is consistently among the top three scoring methods in terms of the percentages of success (Table 1) and rank averages (Table 2). Although the same can be said about the median and Olympic methods, the average of the top three contestant numbers (Table 3) and the average placement of the three strongest contestants (Table 4) are both lower for the mean, providing further evidence that it is better than the other two methods.

In the situations where the panel consists of a variety of judge types, we recommend the z rank as the best scoring method. The z rank method is consistently in the top three scoring methods for percentages of success (Table 1) and rank averages (Table 2). The z-scores and median scoring methods are close behind, but the average of the top three contestant numbers (Table 3) and the average rank of the three strongest contestants (Table 4) for the z rank method have values closest to 2.0. Therefore, the z rank scoring method is recommended based on all of our criteria.

As our recommendations depend on the types of judges in a competition, it is important to be able to identify the scoring behavior of judges in real contests. According to Fenwick and Chatterjee, [3] "if judges... apply consistent judging criteria, we would expect highly correlated scorings." Therefore, it is important to examine the correlation between judges after scores have been recorded. Fenwick and Chatterjee [3] demonstrate how this is done by determining the correlation between judges at the 1980 Olympics. If judges appear to be giving scores that are consistent with all other judges, we can simply use the mean scoring method. On the other hand, low intercorrelations between judges are evidence of inconsistent and

unreliable scoring [5]. Weakly correlated scores indicate that the judges have different scoring standards, and thus we recommend using the z rank scoring method based on our results.

Although we examined several different judge panels, there are many more combinations of lenient, harsh, moderate, and fair judges. With more time, we could examine ten-judge panels where more of the judges are lenient, moderate, or harsh. Also, we only applied seven different scoring methods, but certainly there are more. The issues of score analyses apply to countless real world situations and we hope to have shed light on the judging of subjective contests.

11 Acknowledgements

Firstly, we are grateful to The National Science Foundation, The National Security Agency, and Miami University for funding this opportunity. We would like to thank SUMSRI co-directors Dr. Patrick Dowling and Dr. Reza Akhtar as well as the rest of the SUMSRI staff. Particularly, we would like to extend our gratitude to Dr. Emily Murphree, our incredible seminar director, for her patience, assistance, and instruction throughout our research. We would also like to thank our inspiring Graduate Assistant, John Lewis, for his encouragement and friendship. We would like to offer our thanks to Dr. Tom Farmer for his insightful information on technical writing. Lastly, we would like to thank our fellow SUMSRI students for the wonderful memories and we wish them the best in the future.

References

[1] Mayer Alvo and Paul Cabilio. Average rank correlation statistics in the presence of ties. *Comm. Statist. A—Theory Methods*, 14(9):2095–2108, 1985.

[2] Margaret Cliff A. and Marjorie King C. A proposed approach for evaluating expert wine judge performance using decriptive statistics. *Journal of Wine Research*, 7(2):83–90, 1996.

[3] Ian Fenwick and Sangit Chatterjee. Perception, preference, and patriotism: An exploratory analysis of the 1980 winter olympics. *The American Statistician*, 35(3):170–173, 1981.

[4] Stephen Gordon and Michel Truchon. Social choice, optimal inference and figure skating. *Social Choice Welfare*, 30(2):265 – 284, 2008.

[5] Joseph M. Madden. A note on olympic judging. *Professional Psychology: Research and Practice*, 6(2):111 – 113, 1975.

Appendix

Table A

Panel Name	Judges' Scoring Distributions	Times Used	Judge Types
10 F	Bin(40, C_i)	10	Fair
10 L	Bin(20, C_i) + 20	5	Lenient
	Bin(40, C_i) + 5	5	Lenient
10 H	Bin(33, C_i)	5	Harsh
	Bin(40, $0.75 \times C_i$)	5	Harsh
2L/2M/2H	Bin(20, C_i) + 20	1	Lenient
	Bin(40, C_i) + 5	1	Lenient
	Bin(30, C_i) + 5	1	Moderate
	Bin(40, $C_i + (1 - C_1)/1.01$)	1	Moderate
	Bin(33, C_i)	1	Harsh
	Bin(40, $0.75 \times C_i$)	1	Harsh
	Bin(40, C_i)	4	Fair
1L/2M/2H	Bin(40, C_i) + 5	1	Lenient
	Bin(30, C_i) + 5	1	Moderate
	Bin(40, $C_i + (1 - C_1)/1.01$)	1	Moderate
	Bin(33, C_i)	1	Harsh
	Bin(40, $0.75 \times C_i$)	1	Harsh
	Bin(40, C_i)	5	Fair
2L/2M/1H	Bin(20, C_i) + 20	1	Lenient
	Bin(40, C_i) + 5	1	Lenient
	Bin(30, C_i) + 5	1	Moderate
	Bin(40, $C_i + (1 - C_1)/1.01$)	1	Moderate
	Bin(33, C_i)	1	Harsh
	Bin(40, C_i)	5	Fair