Nationalistic Judging Bias in the 2000 Olympic Diving Competition

John W. Emerson and Silas Meredith

August 22, 2010

About the authors:

John W. Emerson is an Associate Professor of Statistics at Yale University.

Email: jayemerson@gmail.com

Silas Meredith is a math and statistics teacher at the Horace Mann School in New York, and is earning a masters degree in statistics at Columbia University.

Email: silas.meredith@gmail.com

The authors would like to thank Miki Seltzer and David Lin for their contributions to earlier work done on this problem, published in *the American Statistician* in 2009.

Put yourself in the shoes of Mexican diver Fernando Platas. You just lost the 2000 Olympic gold medal in the 3-meter springboard diving competition by an extremely narrow margin to Ni Xiong of China. In such a close competition, did Xiong's high scores from the Chinese judge Facheng Wang cost you the gold medal? Did Wang have a subconscious preference for the style of Chinese divers, or could he have intentionally tried to help Xiong? Table 1 presents scores for three of Xiong's dives from the semifinal round which count toward the final medal ranking. In all three dives, Wang awarded a higher score than the average of the seven judges.

Table 1: Scores for 3 of Xiong's Dives

NZL	CHN	GER	NOR	FRA	USA	PUR	Trimmed Mean
8.0	8.5	7.5	7.5	8.0	7.5	8.0	7.8
8.0	8.0	7.0	8.5	8.0	7.5	7.0	7.7
8.5	8.5	8.0	9.0	8.0	8.5	8.5	8.4

Facheng Wang awarded Ni Xiong a higher score than any other judge on the first dive. This high score was trimmed and didn't contribute directly to the trimmed mean (the average of the middle 5 scores); if it had been a half point lower, the trimmed mean would have still been 7.8. In contrast, the trimmed means would have been lower for the latter two dives if Wang's score had been a half point lower.

What do you think, Fernando Platas of Mexico? Is this proof of biased judging?

This article examines the 2000 Olympic diving scores for evidence of nationalistic judging bias. This competition is particularly well-suited for such an analysis in part because of data availability and transparency in scoring; the identity and nationality of each judge is provided with each score. Many other subjectively judged sports (for example, gymnastics and figure skating) partially or entirely dissociate the scores from the identities of the judges, making a similar study impossible. The data were obtained from official scoring sheets. The analysis was

conducted in the R Language and Environment for Statistical Computing. The data and the R code used for this analysis are available at http://www.stat.yale.edu/~jay/diving.

We will examine 3-meter springboard and 10-meter platform events for men and women. Each event consists of three rounds, and each round consists of 4-6 dives. Each dive is scored by a panel of seven judges of diverse nationalities. In the 2000 Olympic competition, there were 25 judges, 156 divers, and a total of 1541 dives performed in these four events. Scores can (and did) range from 0 to 10 in steps of 0.5, having an average of 6.83 and a median of 7. A judge may not judge a diver from his home country in the final round of the competition, but a total of 314 dives were scored by a judge whose nationality matched that of the particular diver. These "matching" or "matched" dives are the focus of our analysis.

Exploration

Fernando, how would you describe a biased judge in Olympic diving? You might say, "a biased judge is one who tends to award higher scores to her own countrymen than to divers from other countries." We will call this Bias Description 1 (BD1).

Did Wang award higher scores to Chinese divers than he did to divers from other countries? Let's examine the data. Wang's average score for Chinese divers (on his "matched" dives, in our terminology) was 8.45, but his average score for non-Chinese divers was 6.97. This certainly fits BD1: Wang did give substantially higher scores to Chinese divers than to non-Chinese divers. But there's a problem with BD1, and we hope that you can find it if you close this magazine and consider alternative explanations.

China is one of the top competitors in diving. Indeed, the average score from the whole competition, including all judges, was 8.16 for dives performed by Chinese divers and 6.72 for

all others. The problem with BD1 is simply that Chinese divers executed higher-quality dives than did non-Chinese divers. In this context, it doesn't seem at all biased that Wang scored Chinese divers about 1.5 points higher than non-Chinese divers; *any* judge, biased or unbiased, would also have scored Chinese divers higher than non-Chinese divers!

Our analysis needs to consider the quality of dives by comparing the scores of a potentially biased judge (like Wang) to scores from the entire panel of judges on respective dives. Let's refine our description of a biased judge to be "a judge who tends to award scores higher than the panel average to his own countrymen." We'll call this Bias Description 2 (BD2), which seeks to account for the quality of dives; we will use the average of all seven judges' scores as a proxy for the unobservable, true "quality" of the dive.

Is Wang considered biased under this new definition? Again, let's look at the data. Wang often gave the *highest* score on the panel of 7 judges on his matching dives, 14 times in 22 opportunities. He also gave Chinese divers a score higher than the average panel score 16 out of 22 times.

Let's define a new term, "discrepancy," for the difference between a particular judge's score and the untrimmed mean of all 7 judges' scores. (We use the untrimmed mean because we are interested a judge's tendency to differ from other judges; we are not studying effects on the final calculated score, a topic for a different study.) A negative discrepancy indicates a score below the panel average, and suspicious positive discrepancies might be evidence of bias. Let's add the discrepancy to the first dive from Table 1, shown in Table 2.

Table 2: Xiong's scores and discrepancies for an example dive

	NZL	CHN	GER	NOR	FRA	USA	PUR	Untrimmed average
Score	8.0	8.5	7.5	7.5	8.0	7.5	8.0	7.86
Discrepancy	+0.14	+0.64	-0.36	-0.36	+0.14	-0.36	+0.14	

Wang's average discrepancy for his 22 matched dives is +0.17, so on average he scores Chinese divers more generously than the other judges. According to BD2, Wang does appear to be a biased judge because he gives Chinese divers this "nationalistic bump."

But there is a problem with BD2 and still reason to doubt that Wang is a biased judge. Again, we hope that you can identify a possible problem with BD2 if you close the magazine and think about other explanations.

We still haven't looked at how Wang treated non-Chinese divers. We might call Wang "enthusiastic" for Chinese divers, meaning that he tended to have positive discrepancies on matching dives. But BD2 is flawed because it doesn't account for the fact that Wang was "enthusiastic" about all divers. In fact, the average of his discrepancies for non-Chinese divers (on his non-matching dives) was also +0.17. He often gave the highest score on the panel (in 145 out of 335 times) for non-Chinese divers, and he awarded a score higher than the panel average in 212 out of 335 non-matching dives.

After two warm-up descriptions of a biased judge, we are ready for the less obvious but much more sensible Bias Description 3 (BD3): "a biased judge is one who awards higher scores than other judges to his own countrymen, *but fails to award higher scores to non-countrymen*."

Let's compare Wang with American judge Steve McFarland. Like Wang did for the Chinese, McFarland gave higher scores to Americans than to non-Americans (7.80 versus 6.70), but American divers were also very strong, so high scores are not enough to demonstrate bias. Like Wang, McFarland gave higher scores than other judges when judging countrymen (his

average discrepancy for matched dives was +0.20). However, unlike Wang, he was not equally enthusiastic for non-countrymen. His average discrepancy for non-matched dives was only +0.01.

Our analysis has led us to compare the average discrepancies for matched and unmatched dives. Let's formally define a statistic (a quantity calculated from the data): the difference of average discrepancies (DoAD). McFarland showed bias because his DoAD was +0.19 (calculated by subtracting +0.01 from +0.20). In contrast, Wang's DoAD was essentially 0 because his discrepancies were both about +0.17.

Do you feel better, Fernando? Unlike McFarland (and, as we will see, many other judges), Wang did not exhibit nationalistic bias measured in this way. (We have used Platas here only as an example of a diver who potentially stood to lose from biased judging; in real life, he accepted the results gracefully.)

Luck?

One of the central questions that a statistician asks is "Could this observed effect be due to chance?" Like any judge, American judge Steve McFarland was at times above and at times below the panel mean. Isn't it possible that he just happened to have higher discrepancies when Americans were diving, purely by chance? To demonstrate bias, we need to show that he awarded high discrepancies to Americans in a manner that couldn't reasonably be explained by random variation. But how much evidence is "enough?"

If American judge McFarland had been unbiased and his discrepancies had been randomly distributed across all divers, how likely is it that his DoAD would have been +0.19 or higher? To answer this question, we employ a statistical technique called a permutation test. We

take all of McFarland's discrepancies from the actual competition and randomly assign them across all the dives scored by McFarland. Equivalently, we can randomly permute the nationality labels for these dives: "American" or "non-American." After doing this, we can recalculate our test statistic, the DoAD, corresponding to this re-ordering of discrepancies (or permutation of nationality labels). In this way we can study the values of the DoAD that would typically arise if, in fact, scoring were independent of diver and judge nationality.

The result of repeating this exercise twice is shown in Table 3. The first column contains the original discrepancies for the first six dives scored by McFarland; the second column shows the original nationalities of the divers. Columns three and four contain random permutations of the nationality labels. Because only the first six dives are shown, some of the "American" labels were shuffled into lower rows of the table by the permutation and are not visible here. Similarly, the DoAD values in the bottom row of the table are based on the full columns of discrepancies and can't be calculated from the values in these six rows.

If we repeat this process many times we obtain an estimate of the *sampling distribution* for the DoAD test statistic under the null hypothesis that McFarland is an unbiased judge. Using this sampling distribution we are able to estimate how unlikely it would be for McFarland to have a DoAD of +0.19 or higher by chance alone.

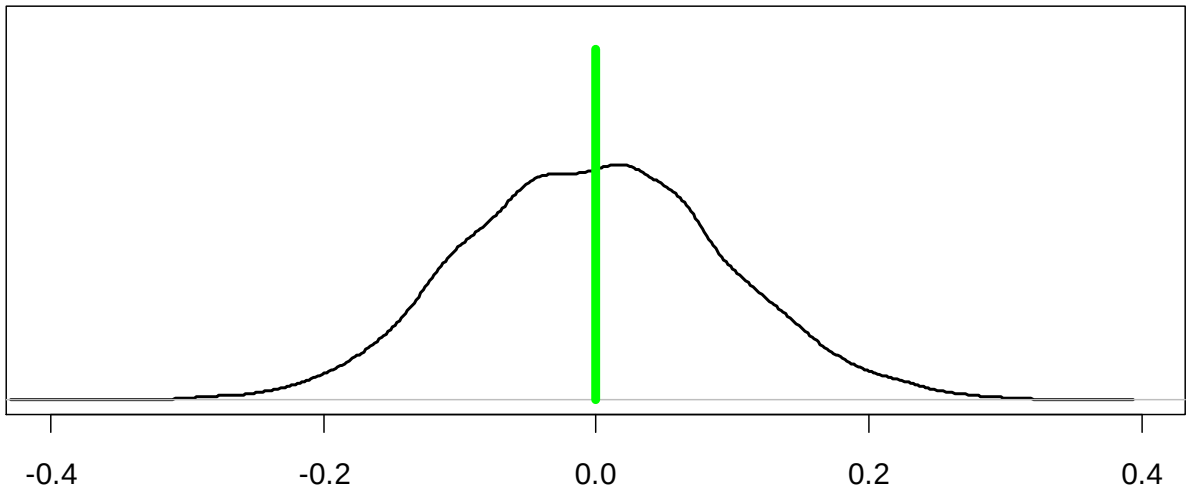
Table 3: McFarland's scores with permuted nationality labels

Discrepancy	Nationality	Permuted Labels	Permuted Labels
-0.21429	non-American	non-American	non-American
-0.07143	non-American	American	non-American
+0.28571	non-American	non-American	non-American
+0.42857	American	non-American	non-American
+0.14286	American	non-American	non-American
...	...	...	...
	DoAD = +0.19	DoAD = +0.06	DoAD = -0.02

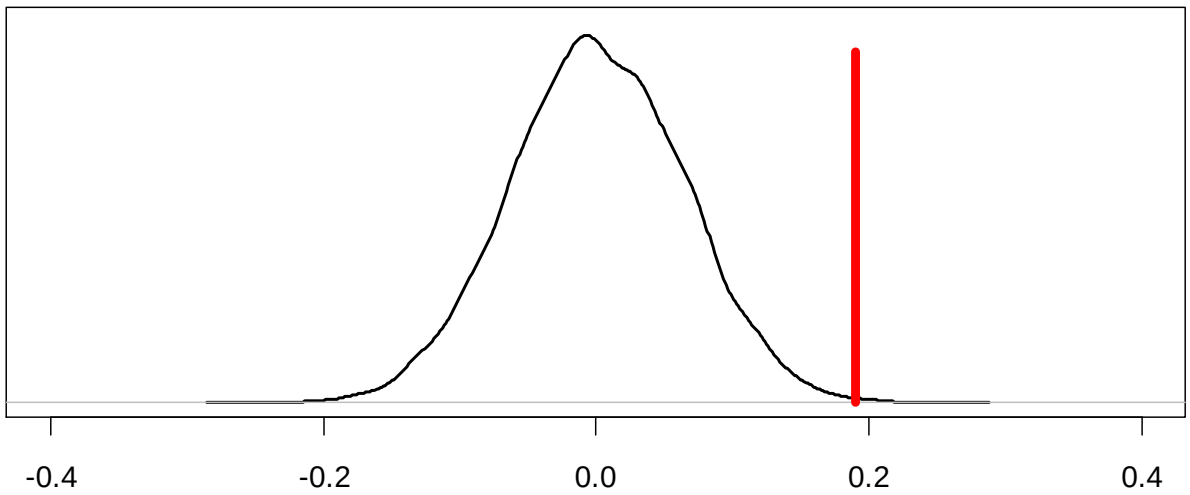
The result of the permutation test provides strong evidence that something other than chance resulted in disproportionately many positive discrepancies in McFarland's scores of American divers. In 100,000 permutations of the labels "American" and "non-American," there was a simulated DoAD of +0.19 or higher only 0.128% of the time. In statistical jargon, we have conducted a hypothesis test and obtained an estimated p-value of 0.00128 (the probability, assuming McFarland had been unbiased, of observing a random DoAD as or more extreme as the one actually observed in the competition, +0.19).

The graphs in Figure 1 show the estimated sampling distributions of our test statistics for Wang and McFarland under the assumption that discrepancies occur independently of nationality. The vertical lines show the observed values of the DoAD in the actual competition data. The areas under the curves to the right of the observed DoADs correspond to the estimated p-values for Wang and McFarland. Wang's observed DoAD was near 0 (the green line) and his p-value near 0.5, representing a 50% chance of a fair judge having a DoAD larger than 0. McFarland's DoAD was +0.19 (the red line) and his p-value 0.00128, representing the small likelihood of observing a DoAD as or more extreme as +0.19 if, in fact, McFarland's discrepancies were independent of the nationality of the divers (thus the tiny area under the curve to the right of the red line).

Figure 1: Estimated sampling distributions of the DoAD



Difference of Average Discrepancy (DoAD) for Judge Wang



Difference of Average Discrepancy (DoAD) for Judge McFarland

Nationalistic bias?

When we perform permutation tests for each of the judges, we see that McFarland's level of bias was extremely common (see Table 4, below). Most judges gave some type of nationalistic bump to their countrymen without giving a similar bump to non-countrymen. For almost all of the judges, the DoAD was more than +0.20, with several higher than +0.40. In the end, Wang was apparently the least biased judge because he gave slightly higher scores to Chinese and non-Chinese divers alike.

Table 4: Results of the permutation tests

Judge	Number of Matched Dives	Average Discrepancy for Matched Dives	Number of Non-Matched Dives	Average Discrepancy for Non-Matched Dives	Difference of Average Discrepancies (DoAD)	**p-value**
Alt, Walter (GER)	25	+0.31	473	-0.08	0.39	**<0.0001**
Barnett, Madeleine (AUS)	38	+0.18	623	-0.11	0.29	**<0.0001**
Boothroyd, Sydney (GBR)	16	+0.32	395	+0.04	0.28	**0.0042**
Boussard, Michel (FRA)	10	0.00	692	-0.11	0.11	**0.1918**
Boys, Beverley (CAN)	13	+0.27	398	+0.06	0.21	**0.0202**
Burk, Hans-Peter (GER)	10	+0.37	149	-0.09	0.46	**0.004**
Calderon, Felix (PUR)	5	+0.23	712	-0.07	0.30	**0.0633**
Cruz, Julia (ESP)	11	+0.29	475	-0.02	0.30	**0.003**
Geissguhler, Michael (SUI)	3	+0.67	398	-0.01	0.68	**0.0015**
Huber, Peter (AUT)	8	+0.31	374	0.00	0.31	**0.0162**
McFarland, Steve (USA)	42	+0.20	615	+0.01	0.19	**0.0013**
Mena, Jesus (MEX)	28	+0.25	828	-0.06	0.30	**<0.0001**
Ruiz-Pedreguera, Rolando (CUB)	11	+0.29	470	+0.01	0.28	**0.0033**
Seamen, Kathy (CAN)	16	+0.15	265	-0.00	0.16	**0.0730**
Wang, Facheng (CHN)	22	+0.17	335	+0.17	0.00	**0.5039**
Xu, Yiming (CHN)	18	+0.30	263	+0.04	0.26	**0.0017**
Zaitsev, Oleg (RUS)	38	+0.27	557	-0.02	0.28	**<0.0001**

These results provide very strong evidence that nationalistic bias was prevalent in the 2000 Olympic diving competition. Most of the p-values are below 0.10 and the majority of them are less than 0.01. Quite simply, if judges' scores were awarded independently of diver nationality, you wouldn't see a data set like this one. Consciously or subconsciously, most of these judges did in fact favor their countrymen.

But what is effect of this level of bias? Roughly speaking, having a DoAD of +0.25 is comparable to giving a half-point bump on about half of the matching dives. This is not an enormous difference, and in many cases would not be enough to change the trimmed mean for that dive. But it is enough of a difference to potentially change the outcome of a close contest. As for Xiong and Platas, taking nationalistic bias out of the equation would most likely not have changed the outcome of the competition: Wang had positive discrepancies for Xiong as well as Platas. Wang was one the least biased judges according to BD3.

There is an argument to be made that in at least one of the events (the women's 10-meter platform) enough bias was demonstrated that the medal standings could reasonably have changed with unbiased judging. For a more complete explanation of this claim about the medal standings, and a fuller statistical analysis that accounts for judges' demonstrated bias for or against all judged countries, see Emerson, Seltzer, and Lin (2009).

Though we used the example of McFarland in order to introduce the reader to an individual case, the intention was not to single him out; we have tried to show that his level of bias was very typical of judging during this competition. Though we have seen evidence that judges favored their countrymen, this is not evidence of cheating or malicious intent. It is not unreasonable to think that stylistic preferences could play a role, and there may be many other factors. It is beyond the scope of this paper (and beyond the capability of this data set) to

determine why this bias exists. But because bias is a relevant topic for any sport with inherent subjectivity, we hope that such an analysis encourages fairness in judges, referees, and organizations that oversee the competitions.

Further Reading and Data

A different approach to studying bias is presented in Emerson, Seltzer, and Lin (2009), "Assessing Judging Bias: an Example from the 2000 Olympic Games" appearing in *the American Statistician* 63(2): 124-131. The data and an R script for parts of the analysis are available at http://www.stat.yale.edu/~jay/diving/.