# Assessing Judging Bias: An Example From the 2000 Olympic Games

John W. EMERSON, Miki SELTZER, and David LIN

Judging bias is an inherent risk in subjectively judged sporting competitions, and recent controversies have spurred researchers to explore these biases wherever possible. Unfortunately, detailed judging results are usually unavailable to the public. For example, the international figure skating scoring system does not allow the study of nationalistic bias, because the scores are reported anonymously. Similarly, the National Basketball Association (NBA) blocked requests for underlying data after a recent academic study of racial bias of NBA referees. This article makes use of a rare case of fully available judging data, examining the diving competitions from the 2000 Summer Olympic Games. We discover strong evidence of nationalistic favoritism in the judging, including one case where the medal standings reasonably could have changed with unbiased judging. We offer a simple framework on which to base future studies of judging bias.

KEY WORDS: Competition; Rating; Regression; Subjective.

## 1. INTRODUCTION

The Summer and Winter Olympic Games currently boast 35 sports with many subdivisions, some of which are scored subjectively by judges. In subjectively judged events, suspicions of favoritism are unavoidable and occasionally erupt into public scandals. In the 2004 Summer Olympics, Russian gymnast Alexi Nemov was awarded a surprisingly low score on the high bar, which was booed loudly by the crowd. After 10 minutes of crowd heckling and judge conferencing, two judges from Malaysia and Canada raised their marks, increasing Nemov's final score from 9.725 to 9.762. Although the change did not boost Nemov onto the medals podium, this incident led to open speculation as to whether certain judges are biased against certain gymnasts or nationalities.

No sport is immune to controversy. Even when outcomes are objectively determined, the first person across cross the finish line is not necessarily the winner; consider Rosie Ruiz in the 1980 Boston Marathon and Ben Johnson in the 100-meter dash of the 1988 Summer Olympics. Ruiz cheated by joining the race near the end and then sprinting to the finish line, whereas Johnson tested positive for the anabolic steroid stanozolol and was stripped of his gold medal. Subjectively judged sports bear additional burdens beyond the control of the athletes; for example, judges may collude to influence outcomes, as allegedly occurred between French and Russian judges in the pairs skating of the 2002 Winter Olympics. More subtle biases also threaten the integrity of subjectively judged sports, however. Detailed data are difficult to obtain, because they are often hidden from public scrutiny. The recent studies by Price and Wolfers (2007) and Larsen, Price, and Wolfers (2008) examined racial biases among National Basketball Association (NBA) referees using data scraped from websites; the NBA subsequently blocked requests for official data records. We were surprised to discover the complete scoring sheets from the diving competition of the 2000 Olympic Games on the Web, and less surprised that these data have since been removed from their original location.

Previous studies have examined judging of various sports in different ways. Popović (2000) examined the rhythmic gymnastics results from the 2000 Summer Olympics and found that judges tended to favor gymnasts from their own countries but not in a consistent, statistically significant magnitude. He examined the differences between scores that a judge gave to a gymnast of his or her own country and the scores from the remaining judges on the panel, then used sign tests to determine whether there was a significant bias.

Lock and Lock (2003) examined figure skating judging in the 2002 Winter Olympics. At the time, the scoring system relied on judges' rankings of skaters' performances. Lock and Lock used a bootstrap technique designed to help identify significant inconsistencies among the judges, and identified one inconsistent judge. Their study of the rankings of the French judge who was implicated in the judging scandal in the pairs long program (that subsequently led to the creation of a new scoring system) is fascinating; the French judge actually had rankings that were among the most *highly* correlated with other rankings.

Zitzewitz (2006) explored nationalistic bias in winter sports and its relationship to organizational decision making. He identified significant judging bias in figure skating (under the 6-point scoring system used until the 2004 World Championships) and ski jumping in the 2002 Winter Olympics and other major international competitions. He found that judges scored athletes from their own countries higher than the other judges did, with the degree of bias varying according to the stakes of the competition. Zitzewitz focused primarily on possible explanations for the main source of bias (in favor of athletes from the judge's country), including judges' career concerns and incentives. Because of sparse data, he limited a study

John W. Emerson is Assistant Professor, Department of Statistics, Yale University, New Haven, CT 06511 (E-mail: *john.emerson@yale.edu*). Miki Seltzer works with Sony Electronics, San Diego, CA 92129. David Lin is an undergraduate, School of Mathematical Sciences, Peking University, Beijing, China.

of the complete set of interactions between judges and athlete nationalities to the top 10 judging countries in figure skating, where every judge from these 10 countries contributed scores for skaters from each of the countries. His discussion of this final model, including an analysis of possible block voting, was brief and did not provide summary statistics required to evaluate the strength of the evidence. But this model was not the focus of his article; he conducted a thorough analysis of primary judging bias, as well as an insightful exploration of the relationship between bias and score truncation based on dropping high and low scores from the panel.

In this article we study the diving competition of the 2000 Summer Olympics. We address subtle difficulties with contrasts in the proposed linear model relating to the incomplete set of interactions between judges and diver nationalities. Section 2 describes the data and conducts exploratory data analysis, motivating the model proposed in Section 3. Section 4 presents the results of the study, in which we find striking evidence of nationalistic bias in the judging. We emphasize that the statistical term "bias" has particularly negative connotations in this study, however. We do not know whether the estimated "biases" are intentional, and they might reflect preferences for style (which could be correlated with nationality). A more comprehensive study of multiple competitions would be ideal. We conclude with a brief discussion on future studies of judging bias.

## 2. OLYMPIC DIVING DATA

The Olympic diving competition involves eight events: springboard and platform diving for men and women, with events for individuals as well as synchronized pairs of divers. The synchronized events have only one round of competition and fewer competitors than the individual events, which have preliminary, semifinal, and final rounds. In addition, in the synchronized events, the nationalities of the judges never coincide with those of the competitors. There is a final noteworthy difference between the individual and synchronized events: The synchronized event panels of nine judges are broken into groups of four and five, who evaluate the quality of execution and degree of synchronization, respectively, whereas the simpler individual competition panels of seven judges provide unified assessments of dive quality. Because of the sheer quantity and simplicity of scoring in the individual events, we drop the synchronized events from this study.

Table 1 summarizes the data from the individual competitions by round. The average number of dives per competitor was 9.87. There were no cases of judges scoring divers from their own countries (so-called "matching dives") in the final rounds, but 314 instances of matching dives among the 1277 dives in the preliminary and semifinal rounds. Judges assign scores ranging from 0 to 10 in increments of 0.5, and the trimmed mean of the middle five of seven judges' scores provides the basis for a dive's score. Although the size of the data set appears to be substantial, not all judges provided scores for all divers (only 58% of the judge/diver combinations occurred in the competition) or even for all nationalities (85% of the judge/diver nationality combinations occurred). Judges often scored only two or three dives of a particular diver, so information at the individual diver level is sparse.

Table 1. Sample sizes in the diving events of the 2000 Summer Olympic Games in Sydney, Australia.

| | Round | | | |
|---|---|---|---|---|
| | Preliminary | Semifinal | Final | Total |
| Dives | 948 | 329 | 264 | 1541 |
| Men's 3-meter springboard | 294 | 90 | 72 | 456 |
| Women's 3-meter springboard | 212 | 95 | 60 | 367 |
| Men's 10-meter platform | 243 | 72 | 72 | 387 |
| Women's 10-meter platform | 199 | 72 | 60 | 331 |
| Unique athlete countries | 42 | 22 | 15 | 42 |
| Unique athletes | 156 | 67 | 45 | 156 |
| Unique judge countries | 21 | 21 | 14 | 21 |
| Unique judges | 25 | 25 | 15 | 25 |
| "Matching dives" with common judge/diver countries | 201 | 113 | 0 | 314 |
| Scores (7 judges per dive) | 6,636 | 2303 | 1848 | 10,787 |

A preliminary examination of nationalistic bias might begin with a simple comparison of scores granted by judges to divers from their own country with scores granted to divers from other countries. The 314 matching scores have a mean of 7.46 ($\pm$ standard deviation 1.21), whereas the other 10,473 nonmatching scores have a mean of 6.81 ($\pm$ 1.48). The 95% confidence interval for the difference between the means of matching and nonmatching scores is 0.51–0.79. We might be tempted to declare a statistically significant nationalistic bias in the judging. Although the distribution of each set of scores is unimodal and only slightly skewed (acceptable for inference given the sample sizes), this approach is difficult to defend, for several reasons. First, the majority of judges come from countries that typically produce better divers, and so matching scores often correspond to better divers, who reasonably earn better scores on average. Second, this approach is inefficient, ignoring the paired structure of the data. Seven judges score each dive, and a more statistically efficient comparison should involve the six nonmatching scores. Third, some judges might provide scores that are on average higher or lower than their peers. Such a consistent bias might confound certain analyses of nationalistic bias, and a thorough analysis must account for this possibility.

A more thoughtful preliminary analysis might examine differences between a score and the mean of the other six scores of the panel, as advocated by Popović (2000); we call this a "difference-based analysis." As an example of this approach, Table 2 presents the scores for the first preliminary round dive of Ni Xiong, the eventual gold medal winner from China. The mean score of the non-Chinese judges was 8.25; a Chinese judge, Facheng Wang, awarded the dive a 7.5. The difference, $7.5 - 8.25 = -0.75$, might lead one believe that Wang is biased against Chinese divers. But such a difference also might be explained by a general tendency for Wang to award lower scores than other judges, independent of the nationality of the diver. In fact, the story is just the opposite when we examine all of Wang's scores; he appears to award scores that are higher on average than the other judges, and overall demonstrates no

Table 2. Judges' scores for a preliminary round 3-meter springboard dive by Ni Xiong of China.

| Judge | Score |
|---|---|
| Dennis Gear (NZL) | 8.5 |
| Facheng Wang (CHN) | 7.5 |
| Walter Alt (GER) | 8.5 |
| Bente Johnson (NOR) | 8.0 |
| Michel Boussard (FRA) | 8.5 |
| Steve McFarland (USA) | 8.0 |
| Felix Calderon (PUR) | 8.0 |

bias for or against Chinese divers. The mean difference between Wang's scores and the mean of the other judges on each panel is 0.20 for both Chinese and non-Chinese divers, and a 2-sample $t$-test comparing the two groups of scores results in a $p$ value of 0.9897.

This difference-based analysis is a huge improvement over the naïve preliminary analysis, and we repeated it for each of the judges to assess for nationalistic preferences. This approach could even be extended to assess differences in judges' preferences across all countries. There remains room for improvement, however. This analysis uses a single score repeatedly for many different comparisons and relies on the mean of six "other" judges as a proxy for dive quality, failing to account for potential biases of these other six judges. Its simplicity is attractive, however, and we provide the results of this difference-based analysis of primary bias of the judges for divers from their home countries in Table 3. We find strong preliminary evidence of judging bias, typically on the order of 0.35 points, with the notable exception of Facheng Wang of China.

## 3. A MODEL OF SCORING

The difference-based analysis is simple and intuitively appealing, but fails to simultaneously incorporate the preferences of all the judges. Recognizing this, we propose a simple model for studying the nationalistic preferences of judges reflected in the scoring of Olympic diving,

$$s_{j,i} = \lambda_i + \mu_j + \beta_{j,C(i)} + \varepsilon_{j,i},$$

where we observe score $s_{j,i}$ for judge $j$ on some dive $i$ and $\lambda_i$ is the unobserved quality of dive $i$, $\mu_j$ is the systematic tendency of judge $j$ to award higher or lower scores than the other judges, and $\beta_{j,C(i)}$ is the interaction term corresponding to judge $j$'s preference for divers of nationality $C(i)$. Contrasts are used for the main effect constraint,

$$\sum_j \mu_j = 0,$$

as well as for the interaction terms,

$$\sum_c \beta_{j,c} = 0$$

for each judge $j$ and

$$\sum_j \beta_{j,c} = 0$$

for each country $c$. For statistical inference, the error terms, $\varepsilon_{j,i}$, are assumed to be iid $\mathrm{N}(0, \sigma)$. This might seem odd, given the discrete nature of the judges' scores, but data analysis is rarely an exact science; we examine this assumption later. Finally, we must omit the dives performed by divers from a small number of countries whose competitors were all eliminated in the preliminary round. The inclusion of these dives introduces colinearities; judges' preferences between these countries are not estimable.

We do not estimate main effects for diver countries. Some divers may be more accomplished than others, and some countries may specialize in diving. But our interest lies in judging bias, assessed purely through a comparison of judges' scores to

Table 3. Results of an exploratory difference-based analysis of the bias of judges in favor of divers from their home countries.

| | Mean of matching differences ($SD, n$) | Mean of nonmatching differences ($SD, n$) | $p$ value |
|---|---|---|---|
| **Barnett (AUS)** | **0.2127 (0.5216, 38)** | **−0.1242 (0.5238, 623)** | **0.0004** |
| Huber (AUT) | 0.3646 (0.4475, 8) | 0.0045 (0.4707, 374) | 0.0575 |
| Boys (CAN) | 0.3141 (0.3571, 13) | 0.0657 (0.4224, 398) | 0.0289 |
| Seaman (CAN) | 0.1771 (0.4419, 16) | −0.0050 (0.4970, 265) | 0.1301 |
| Wang (CHN) | 0.2007 (0.4215, 22) | 0.2020 (0.5123, 335) | 0.9897 |
| **Xu (CHN)** | **0.3519 (0.3722, 18)** | **0.0437 (0.4207, 263)** | **0.0030** |
| Ruiz-Pedreguera (CUB) | 0.3333 (0.3855, 11) | 0.0080 (0.3936, 470) | 0.0191 |
| **Cruz (ESP)** | **0.3333 (0.2814, 11)** | **−0.0218 (0.4306, 475)** | **0.0018** |
| Boussard (FRA) | 0.0000 (0.2324, 10) | −0.1276 (0.4543, 692) | 0.1217 |
| Boothroyd (GBR) | 0.3750 (0.5037, 16) | 0.0453 (0.4953, 395) | 0.0205 |
| **Alt (GER)** | **0.3600 (0.3763, 25)** | **−0.0958 (0.4996, 473)** | **<0.0001** |
| **Burk (GER)** | **0.4333 (0.5209, 10)** | **−0.1074 (0.4663, 149)** | **0.0095** |
| **Mena (MEX)** | **0.2887 (0.4447, 28)** | **−0.0667 (0.4592, 828)** | **0.0003** |
| Calderon (PUR) | 0.2667 (0.3302, 5) | −0.0839 (0.5144, 712) | 0.0760 |
| **Zaitsev (RUS)** | **0.3092 (0.3801, 38)** | **−0.0229 (0.4571, 557)** | **<0.0001** |
| Geissbuhler (SUI) | 0.7778 (0.2679, 3) | −0.0149 (0.4428, 398) | 0.0337 |
| **McFarland (USA)** | **0.2341 (0.3410, 42)** | **0.0122 (0.4654, 615)** | **0.0002** |

NOTE: The $p$ values result from two-sample $t$-tests comparing the matching and nonmatching differences between the judges' scores to the consensus of the other six judges on each dive.

the estimated quality of the dives; the proposed model and contrasts provide the easiest way to address this question of bias. Adding main effects would be an unhelpful change in the parameterization with a statistically equivalent fit.

Both our data and the data used by Zitzewitz (2006) suffer from an incomplete set of judging results; not every judge scores dives for divers of every nationality. Zitzewitz addressed this issue by limiting that part of his analysis to the top 10 judging countries for which he could estimate the complete set of interactions. In this case the proposed linear model can be estimated via least squares and has a clear advantage over the difference-based approach; it recognizes that bias is expressed with respect to the unobserved quality of the dive and that quality is not simply the consensus of the panel scores. In practical applications, however, estimation of the model with data from a large competition poses a challenge. If direct least squares estimation is computationally infeasible, then the estimation may proceed in an iterative manner, alternating between estimation of dive quality and interaction (and main effect) parameters. This may be viewed either as an iterative algorithm for maximum likelihood or as an application of the EM algorithm (Dempster, Laird, and Rubin 1977) in this instance, the difference is purely philosophical.

With empty cells in the two-way contingency table of scores by judge and diver nationality, statistical software packages can fail to estimate the model interaction terms with the desired contrasts. Some environments (e.g., the R language) might be able to estimate the model using treatment contrasts in which a baseline level is omitted and coefficients are estimated for subsequent levels of the factor with respect to the baseline level, as long as the baseline level has a complete set of interactions. But even when this estimation is successful, these coefficients are not convenient for this problem, and the standard errors are not appropriate for centered estimates of the coefficients (i.e., the desired contrasts for the proposed model discussed earlier).

With the desired contrasts (e.g., called "contr.sum" in the R language), any empty cell in the two-way contingency table of scores by judge and diver nationality results in a singularity in the model design matrix and useless coefficient estimates. A simple, yet not ideal way, to avoid this difficulty is to introduce a single imaginary score for each case in which a judge fails to evaluate dives from a particular country. In the course of iteratively estimating the proposed model, we must make gradual adjustments for the artificial scores so that the corresponding estimates $\hat{\beta}_{j,C(i)}$ to converge to 0, preventing them from influencing other coefficients through the restrictions imposed by the contrasts. The resulting $p$ values for these coefficients equal 1, reinforcing our sense of poor knowledge of judges' preferences for divers from these countries. The standard error of the residual will be correct, but the $R^2$ value will be slightly higher than it should be, and the standard errors of the coefficients will be inflated because of the near-singularity in the model design matrix.

Our solution to this problem involves manual creation of the model design matrix. Each judge scores dives for divers representing some subset of the diver nationalities, and contrasts must be customized for each judge to provide the desired constraint,

$$\sum_{c \in N(j)} \beta_{j,c} = 0,$$

where $N(j)$ denotes the set of nationalities scored by judge $j$. At the same time, the design matrix must satisfy the constraint

$$\sum_{j \in J(c)} \beta_{j,c} = 0,$$

where $J(c)$ denotes the set of judges scoring dives of divers from country $c$. On our website we provide an R function for constructing the portion of the model design matrix corresponding to the interaction terms with the desired constraints. This approach produces the same coefficient estimates as the simpler approach described previously, but with correct statistics for inference based on the linear model when estimated directly via least squares. We return to a brief discussion of inference in Section 5.

## 4. RESULTS

We first omitted dives of the competitors from eight countries whose divers failed to reach the semifinal of an event. Using this reduced data set, we estimated the model using 10,423 scores by 25 judges from 21 countries of 1489 dives of 147 divers from 34 countries. In total, we estimated 2297 quantities; of these, 1489 were the estimated dive qualities. There were 25 main effects for the judges and 783 interaction terms between judges and diver nationalities.

The main judge effects were dispersed closely around 0, with only a few exceptions. Chinese judge Facheng Wang scored dives 0.22 points higher than the other judges on average, whereas Australian judge Madeleine Barnett and French judge Michel Boussard were generally critical of dive quality ($-0.11$ and $-0.15$ points below average, respectively); the $p$ values for these coefficients are $<10^{-6}$, indicating striking statistically significant differences from the average judgment. None of the other judges exhibited a strong, systematic tendency to award higher or lower scores.

Table 4 compares the nationalistic bias for the 17 judges who scored divers from their own countries. The results of the exploratory analysis of differences, discussed in Section 2, generally match the results of the linear model and are included for completeness. Only judges from Canada, China, France, and Puerto Rico might avoid further scrutiny based on these data. The other judges all exhibited varying degrees of bias in favor of divers of the same nationality, with 10 of the 17 bias coefficients having $p$ values $< 0.01$. Judges from Australia, Germany, Mexico, Russia, and Switzerland appear most guilty of nationalistic bias of about 0.3 points or more ($p$ value $< 0.0002$), and most biases are estimated to be at least 0.25 points in magnitude, with 11 of the 17 $p$ values $< 0.05$. Several particular cases are worth noting, highlighted with an asterisk in Table 4. The full analysis uncovered the notable biases of Swiss judge Michael Geissbuhler, British judge Sydney Boothroyd, and Austrian judge Peter Huber, which were

Table 4. Comparing estimates of primary judging bias using the exploratory difference-based analysis and the full linear regression analysis.

| | Bias (difference analysis) | $p$ value | Bias (linear model) | $p$ value |
|---|---|---|---|---|
| **Barnett (AUS)** | **0.3370** | **0.0004** | **0.2827** | **0.0001** |
| **Huber (AUT)*** | **0.3601** | **0.0575** | **0.4361** | **0.0090** |
| Boys (CAN) | 0.2484 | 0.0289 | 0.2210 | 0.0812 |
| Seaman (CAN) | 0.1821 | 0.1301 | 0.1747 | 0.1356 |
| Wang (CHN) | −0.0012 | 0.9897 | −0.0430 | 0.6632 |
| Xu (CHN)* | 0.3081 | 0.0030 | 0.2148 | 0.0507 |
| Ruiz-Pedreguera (CUB) | 0.3254 | 0.0191 | 0.3103 | 0.0251 |
| **Cruz (ESP)** | **0.3551** | **0.0018** | **0.3632** | **0.0093** |
| Boussard (FRA) | 0.1276 | 0.1217 | 0.1763 | 0.2162 |
| **Boothroyd (GBR)*** | **0.3296** | **0.0205** | **0.3036** | **0.0087** |
| **Alt (GER)** | **0.4558** | **<0.0001** | **0.4477** | **0.0019** |
| **Burk (GER)** | **0.5407** | **0.0095** | **0.4797** | **0.0001** |
| **Mena (MEX)** | **0.3554** | **0.0003** | **0.3175** | **0.0002** |
| Calderon (PUR) | 0.3506 | 0.0760 | 0.2075 | 0.3256 |
| **Zaitsev (RUS)** | **0.3321** | **<0.0001** | **0.3025** | **<0.0001** |
| **Geissbuhler (SUI)*** | **0.7926** | **0.0337** | **1.3016** | **<0.0001** |
| **McFarland (USA)** | **0.2219** | **0.0002** | **0.2014** | **0.0047** |

NOTE: * denotes interesting cases discussed in Section 4, while bold denotes cases whith $p$ values $< 0.01$.

far less pronounced in the exploratory difference-based analysis, and reduced suspicions of the nationalistic bias of Chinese judge Yiming Xu.

Table 5 shows the most statistically significant preferences that judges exhibited for any of the diver nationalities. Some of these estimates are based on very few dives; a cursory examination of the Swiss dives involving the implicated judges reveals typically poor-quality dives, where the high scores probably reflect judge sympathy and the result simply did not matter for the competition. Boothroyd's apparent support for Chinese divers and Wang's penalization of Australian divers seem far more worrisome, however.

Table 5. The most significant biases identified by the full regression model.

| Judge (country) | Diver country | Dives | Coefficient | $p$ value |
|---|---|---|---|---|
| Barnett (AUS) | AUS | 38 | 0.28 | 0.000169 |
| Barnett (AUS) | BRA | 14 | 0.45 | 0.000235 |
| Wang (CHN) | AUS | 22 | −0.38 | 0.000127 |
| Wang (CHN) | KAZ | 6 | 0.62 | 0.000511 |
| Wang (CHN) | TPE | 3 | 1.18 | 0.000005 |
| Ruiz-Pedreguera (CUB) | GER | 11 | −0.28 | 0.000617 |
| Cermakova (CZE) | TPE | 4 | −0.78 | 0.000564 |
| Hassan (EGY) | CAN | 30 | 0.32 | 0.000373 |
| Hassan (EGY) | GER | 21 | 0.36 | 0.000334 |
| Boothroyd (GBR) | CHN | 32 | 0.28 | 0.000999 |
| ALT (GER) | AUT | 11 | 0.51 | 0.000345 |
| ALT (GER) | GER | 25 | 0.35 | 0.000143 |
| Mena (MEX) | MEX | 28 | 0.32 | 0.000257 |
| Hood (NZL) | USA | 42 | 0.26 | 0.000471 |
| Gear (NZL) | INA | 2 | −1.06 | 0.000705 |
| Calderon (PUR) | CUB | 20 | 0.36 | 0.000940 |
| Zaitsev (RUS) | RUS | 38 | 0.30 | 0.000045 |
| Geissbuhler (SUI) | SUI | 3 | 1.30 | 0.000098 |

Rather that present an exhaustive table of estimated coefficients, standard errors, and $p$ values, we offer a graphical display of the results. Consider U.S. judge Steve MacFarland, who appears to be biased in favor of U.S. divers, with an estimated coefficient $\hat{\beta}_{McFarland,USA} = 0.20$ (standard error, 0.07; $p$ value $= 0.0047$). Figure 1 displays the positive values with circles and negative coefficient estimates with squares; the area of each symbol is proportional to the magnitude of the coefficient estimate. Black circles highlight primary positive nationalistic biases. In addition, filled symbols denote $p$ values $<$ 0.01, and coefficient estimates with $p$ values $> 0.10$ are omitted from the display. The largest shaded circle represents the statistically significant bias of Swiss judge Michael Geissbuhler for Swiss divers (with the $p$ value of the regression coefficient, 1.30, $<10^{-4}$). Rows and columns are sorted alphabetically by country.

The minimum and maximum residuals are $-2.64$ and 1.58, respectively, with a residual standard error of 0.42 and $R^2 =$ 0.9972 (unsurprising given the estimation of each dive quality). Figure 2 shows a normal quantile plot of the residuals, as well as a plot of the residuals by the estimated dive qualities. A small number of unusually low scores might be called into question, one in particular: the third final-round 3-meter springboard dive of Chinese diver Mingxia Fu, the eventual gold medalist. The surprisingly low score of 5.5 awarded by New Zealander Robin Hood was perhaps an attempt to assist divers from Germany, Russia, or the U.S., who trailed Fu in the competition; our model estimates a dive quality of around 8. The other low residuals are less interesting, associated with poor dives less critical to the outcome of the competition. We see less variability in the residuals for top-quality dives (i.e., those above an estimated quality of 8 or 8.5). Neither of these plots provides a basis for doubting the finding of many statistically significant biases, however.

We now consider the practical significance of these results. In the men's 3-meter springboard competition, for example, the top 18 divers from the preliminary round qualified for the semifinal round. The aggregated results of the preliminary and semifinal rounds were used for the selection of the top 12 divers for the finals. The medal standings were determined by aggregating the results of the semifinal and final dives, ignoring the results of the preliminary round. The competition was close throughout; the eventual medalists placed first, second, and third after each of the three rounds. Only the semifinal and final round dives determined the medal standings, however. Ni Xiong (China) won the gold, barely edging out Fernando Platas (Mexico) by 0.3 points, with Dmitri Saoutine (Russia) winning the bronze, 5 points behind. There gap between the third-place and fourth-place competitors (about 32 points) was much more substantial that that separating the top three competitors. Using the estimated dive qualities, we can easily calculate "unbiased dive scores," essentially the expectation of the scores in the absence of any judging bias or systematic tendencies of judges to award higher or lower scores. Here the results of the competition would not have changed, and Ni Xiong actually would have won the gold by a much greater margin, about 6 points. It appears that the judges in aggregate provided a slight upward bias helping the Mexican diver while severely penalizing the Chinese diver, almost costing him the competition.

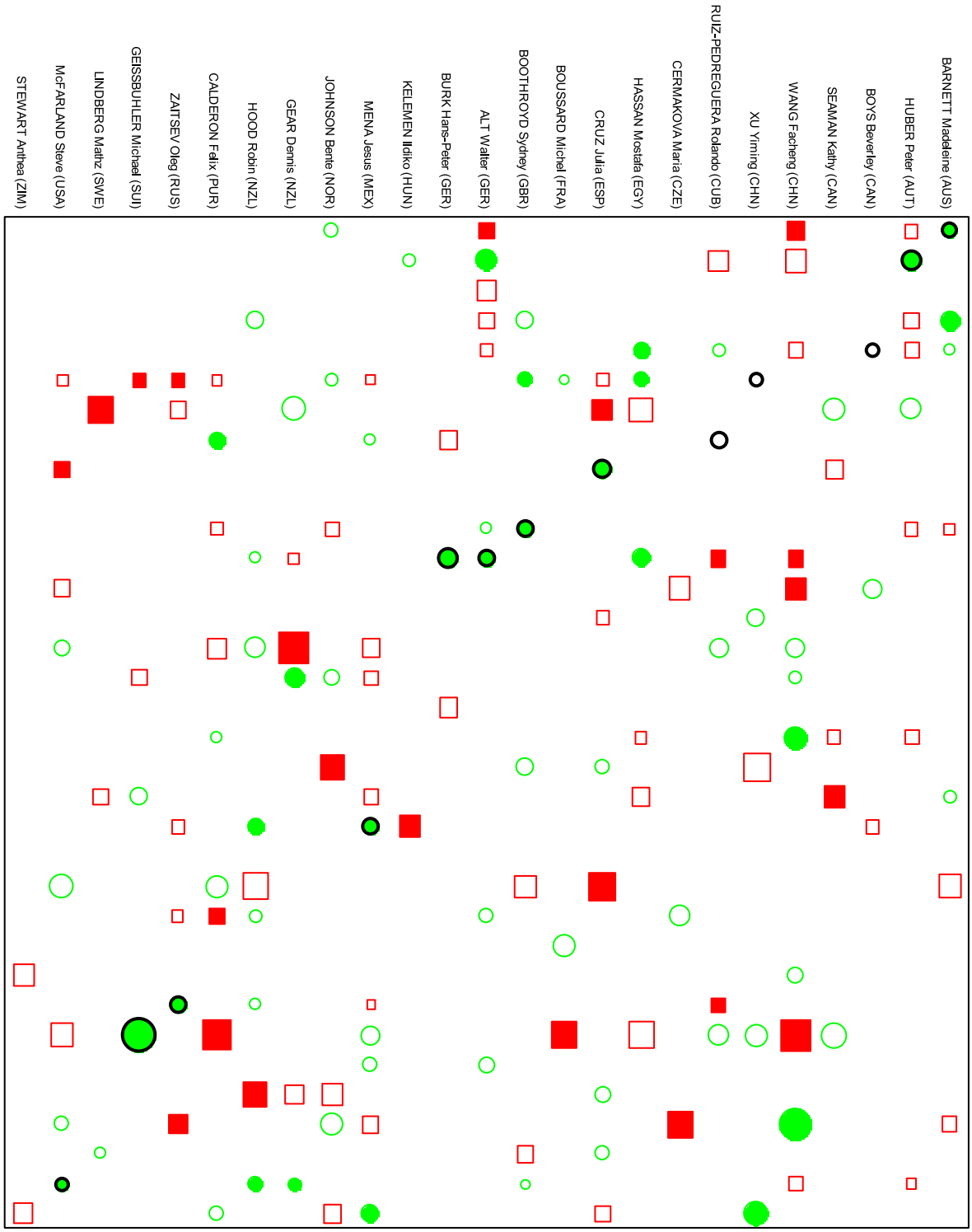


Figure 1. The area of each symbol is proportional to the size of the regression coefficient, depicting the magnitude of a positive (circle) or negative (square) bias. Only those with $p$ values $< 0.10$ are shown, and those with $p$ values $< 0.01$ are shaded. Black circles denote primary positive biases of judges for divers from their own countries.

In the women's 10-meter platform competition, the removal of judging bias might have changed the medal standing. U.S. diver Laura Wilkinson finished ahead of Chinese diver Li Na by 0.74 points. After removing the effect of judging preferences, the dive qualities indicate that Li could have won the gold medal in expectation by a margin of 0.36 points. Although Wilkinson's "unbiased score total" was virtually identical to her actual total, the Chinese diver was penalized slightly by the judges. Again, we emphasize that this implication ignores the natural and unavoidable variability in scoring, and the particular outcome in a close contest such as this is essentially a coin toss.

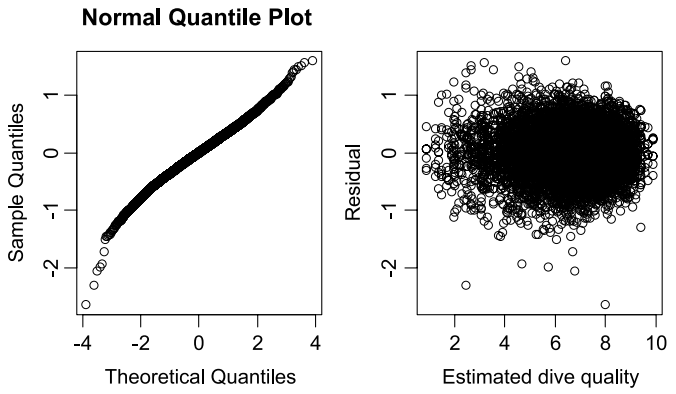


Figure 2. Residual plots for the full regression model. A few unusually low scores are apparent in each plot. The smallest residual (corresponding to the estimated dive quality of about 8, at right) represents a suspiciously low score (5.5) given by New Zealand judge Robin Hood to Chinese gold medal winner Mingxia Fu.

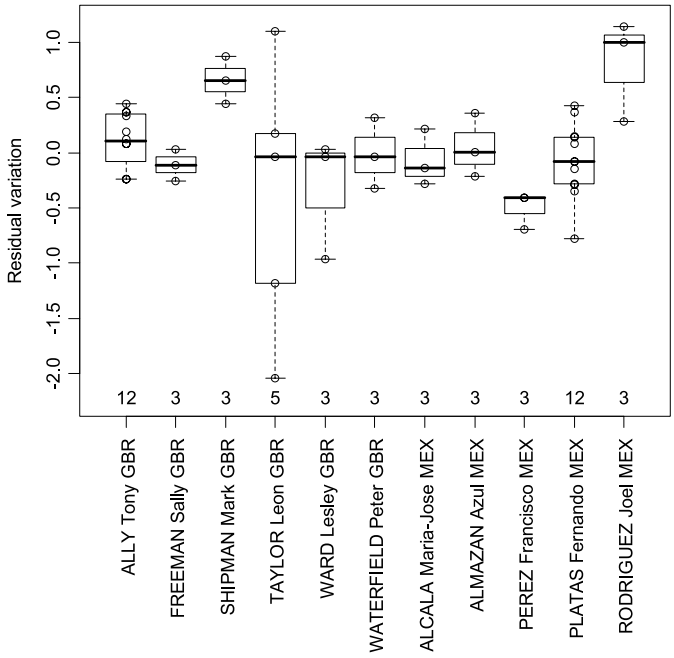


Figure 3. Puerto Rican judge Felix Calderon exhibited some limited evidence of preferences for individual divers from Great Britain and Mexico beyond what was captured by the estimates of his nationalistic preferences. The plot includes both the points and boxplots, with the counts of numbers of dives scored appearing at the bottom.

We next provide an informal examination of bias within the preliminary and semifinal rounds, focusing on countries contributing divers to the finals of the events. We used the same methodology to estimate the model twice, but using only the preliminary and semifinal round dives. The smaller samples produced far less pronounced results, as expected, but the results were consistent for the most part. We note, however, that Russian judge Oleg Zaitsev tended to penalize Chinese and Mexican divers, top competitors of the Russian divers, in the semifinals but not in the preliminary rounds. This makes strategic sense, because only the semifinal scores contribute to the final standings. Similarly, Madeleine Barnett of Australia was neutral toward U.S. divers in the preliminary round but strongly biased against them in the semifinals. U.S. and Australian divers finished next to each other in the final standings of three of the four events, and in the last event, a U.S. diver failed to advance to the finals (edged out by an Australian). Walter Alt of Germany similarly penalized U.S. divers in the semifinals, after being somewhat helpful toward the Americans in the preliminary rounds.

We concluded our analysis by considering whether judges exhibited preferences for individual divers, perhaps on the basis of style (which is not measured but might be observed by the judges). A residual from our model represents the difference between judge $j$'s score and the predicted score of judge $j$ for dive $i$, given the nationalities of the judge and the diver. But there might still be preferences of judges for divers from a given country that our model has not captured; these preferences should be evident in these residual differences. To examine this issue, we conducted an analysis of variance predicting these residuals using the individual divers as the explanatory factor. The interpretation of the model coefficients is simple: They represent the preferences of the judge for individual divers (for unobserved reasons), after controlling for the nationalistic preferences.

The result of this final analysis of judge preferences for individual divers is overwhelmingly underwhelming. Only one of the judges exhibited differences in their preferences for individual divers that might be of statistical significance (without a correction for multiple testing), after controlling for nationalistic preferences and the estimated dive qualities. Given the small number of scores available for most of the judge–diver combinations, this finding is not surprising. The only judge exhibiting marginally interesting preferences for individual divers was Felix Calderon of Puerto Rico ($p$ value $= 0.02$ from ANOVA). Figure 3 summarizes the residual variation of Calderon for divers from Great Britain and Mexico, the two countries containing the most significant (yet hardly convincing) evidence of differences in Calderon's preferences across divers. Calderon may have liked the styles of Rodriguez and Shipman, scoring their dives more highly than predicted compared with most of their teammates from Mexico and Great Britain, and likewise scoring some of these teammates lower than predicted by the model. But Calderon judged only three dives of most of these divers, including Shipman and Rodriguez. Given the number of judges analyzed and the small amount of information available for study at this level, we lack the power to identify differences in preference for individual divers, if they exist.

## 5. CONCLUSION

The data set studied in this work is available on the Web at *http://www.stat.yale.edu/~jay/diving/*. This website also provides an R function for constructing that part of the model design matrix corresponding to the interaction terms using the desired contrasts.

We used the manual construction of contrasts to estimate a complete model for judging bias. With an incomplete set of interactions between judges and diver countries, statistical software packages may produce inexplicable results, particularly

when moving beyond the typical treatment contrasts. Our approach could be used in similar studies of judging bias with any statistical software package having at least a rudimentary scripting language. When direct estimation of the full model via least squares is feasible, the resulting statistical inference is appropriate. When estimation proceeds iteratively over different sets of the coefficients, the resulting standard errors will underestimate the true uncertainty. In the 2000 Olympic diving competition, the evidence of primary judging bias (i.e., judges in favor of divers from their own countries) is particularly noteworthy, but many other biases are evident as well. Other hypotheses about judging behavior could be explored using these data, although results from other competitions would be helpful.

*[Received June 2008. Revised October 2008.]*

## REFERENCES

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 185–197.

Larsen, T., Price, J., and Wolfers, J. (2008), "Racial Bias in the NBA: Implications in Betting Markets," *Journal of Quantitative Analysis in Sports*, 4 (2), article 7.

Lock, R., and Lock, K. F. (2003), "The Statistical Sports Fan: Judging Figure Skating Judges," *STATS*, 36, 20–24.

Popović, R. (2000), "International Bias Detected in Judging Rhythmic Gymnastics Competition at Sydney-2000 Olympic Games," *Physical Education and Sport*, 1, 1–13.

Price, J., and Wolfers, J. (2007), "Racial Discrimination Among NBA Referees," working paper, National Bureau of Economic Research, Cambridge.

Zitzewitz, E. (2006), "Nationalism in Winter Sports Judging and Its Lessons for Organizational Decision Making," *Journal of Economics and Management Strategy*, 15, 67–99.