Presentation

Satu Helin & Xuechen Hu

1 April 2020

**Revealing the Secrets of Figure Skating Scoring**

BuzzFeed Story <https://www.buzzfeednews.com/article/johntemplon/the-edge>

BuzzFeed data + notebooks <https://github.com/BuzzFeedNews/2018-02-figure-skating-analysis>

Scoring explanation video <https://www.youtube.com/watch?v=wheQNIdGZns>

Description of ISU algorithm <https://www.documentcloud.org/documents/4367532-2098-Officials-Evaluation-Figure-Skating-2017-20>

Interviews: Reporter John Templon (by email), figure skating amateur trainer Eeva Ketvel (Messenger)

**Why figure skating was under the radar**

The story examines the home-country preference bias in the scoring system of the International Skating Union (ISU). It was published at the eve of Winter Olympics 2018; as a discipline based on the evaluation, the scoring system has always been under discussion. BuzzFeed wanted to find out if the bias could be proven.

The problem in figure skating has been the possible pressure from judges’ home-country federations: the judges are chosen by their federations. To protect the judges from that pressure, in 2002, the ISU changed the scoring system to anonymous judging. Yet, at the same time, transparency was taken away from the system. This meant that nobody would notice if the judge just wanted to favor certain skaters. In the 2016-2017 season, the ISU canceled the anonymity and made it possible to trace the bias in the system again. This meant that there finally was data available for investigation.

**BuzzFeed methodology in short**

Nowadays, ISU uses a specific algorithm to find judges that might be giving the points unfairly. In practice, the algorithm detects judges whose points differ from others.

BuzzFeed recreated the ISU’s bias detection algorithm, using the [description](https://www.documentcloud.org/documents/4367532-2098-Officials-Evaluation-Figure-Skating-2017-2018.html) they had on ISU webpage, and built their own bias evaluation system. They gathered more than a year’s scoring data from international figure skating races using documents released after the competition. The example of the data source [looks like this.](http://www.isuresults.com/results/season1718/gpf1718/gpf2017_protocol.pdf?) Altogether they scraped 17 pdf documents like the example and created three csv files based on them.

In data analysis, they got help from three statisticians, from which two had worked with figure skating data and one with sports data. In the [very carefully documented analysis](https://github.com/BuzzFeedNews/2018-02-figure-skating-analysis), BuzzFeed [compared](https://github.com/BuzzFeedNews/2018-02-figure-skating-analysis/blob/master/notebooks/home-country-preference-analysis.ipynb) the scores given by judges to skaters from their own country to those they give skaters from other countries.

From the pdf-documents, BuzzFeed extracts 1,726 performances, 23,932 aspects, and 214,531 scores. BuzzFeed then removes the junior performances from the dataset and keeps 1632 senior performances. Then they calculate the total points each judge gives to the performance, subtract the points that are deducted by the technical panel and get the overall score. To calculate the home country preference, each judge’s score is compared to the average score to account for their overall tendency to give scores higher or lower than the average. To examine if the overall home-country preference exists in the scoring process, a t-test is run and the p value is as small as 5.1762242757333122e-225, which proves that the phenomenon is statistically significant.

Then the analysis digs into individual judges’ preferences. Because the p-value could not be trusted alone (some judges only refer in a few competitions) they wanted to make sure that the outcome was not only a coincidence. To do this, they assigned all of the judge's score differences to home-country or non-home-country skaters at random. For each judge, the simulation was repeated 1 million times. The point was to check how often the simulated scoring preference equals or exceeds the actual home-country preference. This provided an estimated probability of home-country preference occurring by chance alone. The test resulted that the odds for this kind of chance were less than 1:100000.

The main finding was that The ISU’s algorithm system was not effective enough. It failed to flag any of the biased cases that were detected by the BuzzFeed’s algorithm.

**Why the ISU’s algorithm (and system) is problematic?**

BuzzFeed recognized five main problems in the scoring process: three are directly linked to the algorithm the ISU uses and two indirectly.  
  
1. ISU algorithm flags judges whose scores fall above or below the average of the other scores on their panel. According to the BuffFeed analysis, however, the algorithm was ineffective because ‘the corridor’ is so wide that it caught only the most extreme outliers. The BuzzFeed algorithm analysis was able to find judges that acted suspiciously but were not flagged by ISU. China's judges on average gave Chinese skaters a 4.6-point boost, the largest of any country with at least 50 home-country judgments in the dataset BuzzFeed News Top-Level Figure Skating Judges Consistently Favor Skate analyzed. Italy, Russia, the United States, and Canada, all gave their skaters a boost of more than 3.4 points. There are 27 judges flagged by BuzzFeed’s analysis and their’ home-country preferences are listed at the end of the [notebook](https://github.com/BuzzFeedNews/2018-02-figure-skating-analysis/blob/master/notebooks/home-country-preference-analysis.ipynb)

2. ISU’s system looks at only one performance at a time. Therefore, it wouldn’t detect judges who consistently, in many competitions, favor their home-country skaters. This means the judge could give little higher points to the compatriots in every competition and the algorithm would never detect it if he/she just stayed inside the ‘corridor’.

3. The third problem that remains in the scoring system is the collusion of judges that down-vote certain countries and up-vote the others. This may actually make the judge who does not participate in the complicity an outlier because his/her scores would, therefore, differ from the others.

4. The next problems are related to the judges and their selection process. Each country selects them themselves. Especially in countries that have a lot of international-level judges, this could lead to the selection of the ones that are “home-friendly”. This was also mentioned by long term coach and ice skating enthusiast **Eeva Ketvel** in messenger discussion. According to her, the judges found in the article represent countries that have a lot of competition for judge places. She also pointed out the discussion that some coaches might try to gather a lot of athletes from different countries so that they would have the home advantage as often as possible. Also, the story highlights, many judges are very familiar with the skaters from their own country.

5. The fifth thing is that judges from certain areas might simply like more the style of that area; for example, Asian judges might appreciate more Asian style than the European.

The last two points mean the problem is not only the algorithm but the way the variables (=judges) are fed to it.

**What could be changed to BuzzFeed’s analysis and reporting?**

As mentioned in the story, judges may favor the style of the region they come from. Yet, the concern wasn’t really proven in the story. Perhaps they could havetested if the judges favor the skaters from their own continent, not only their own country, to see if this theory seems to be true.

However, as the writer of the story, John Templon said. the biggest advantage of potentially doing the story now would not be in the ISU's algorithm, but in fact that there would be more judging data available. “We had a little less than a year's worth of data in which to do our analysis because of how figure skating had kept judging secret. This prevented us from doing deeper analyses where we could've looked not just at national bias, but potentially at "trades", where judges would swap votes in competitions to gain preferences for their country”, he said. It would definitely be interesting to test their results with the new dataset.

The story was persuasive and well written. The only lack of reporting we want to raise isthe number of anonymous sources. Templon said it was extremely hard to get comments and pointed out that this differed a lot from his previous experience on covering sports.”It was interesting to be honest because when I reported on tennis, for instance, the TIU authorities who police tennis were very willing to at least talk. I think part of the issue is that ISU members are also judges, so there was a large conflict of interest for them.”, he said. This point could have been better underlined in the article.

**Conclusion**

This story shows the fact we have often discussed: the bias is not (at least only) a technical question. Even the faults in the ISU algorithm could be corrected; it would not solve the issue of how the judges are selected and how they are affected by their personal relationships. Also, if this story was only based on the algorithm and did not include a lot of additional reporting, it would only have been half-right.