Skating around the Figures: Analyzing Factors that Impact Figure Skating Scores

By: Michael Pilson, James Pascale, Dominic Blum, and Lily Klucinec

Introduction:

Despite requiring immense skill and rigorous training, figure skating is one of the few sports where performance is judged subjectively. Each skater is judged according to the following three aspects: the element score, the component score, and deductions. The element score reflects the technical difficulty and execution of specific moves. The component score captures the performance's artistic quality—skating skills, choreography, and overall presentation. Deductions are subtracted for errors such as falls or time violations. These values combine into a "total segment score," which determines a skater's final placement in a competition. This unique scoring system blends athleticism with artistry, making figure skating as much about expression as it is about physical precision.

However, figure skating is not without controversy. In the 2002 Winter Olympics, a judge was pressured into voting one skating pair over another for gold ("IOC finds fraud" 2021). As such, there can be several factors outside an athlete's control that can affect their scores and placements, such as the order in which a skater performs. There's also potential judging bias towards specific skaters, moves, and nationalities. For example, Zitzewitz established statistically significant judging bias when considering medal contention (Zitzewitz 2006). Therefore, this project was designed to answer the following research questions:

1) Are there external factors that significantly affect a skater's score that can be identified and quantified?
2) Can we find any potential evidence of judge biases based on differences in scoring behavior?

These questions are both interesting and important because they challenge the assumption that figure skating is judged solely on merit. Uncovering biased judging patterns could spark conversations about transparency, and reform the judging system. This analysis would help ensure that athletes are evaluated solely on the quality of their performances and highlight how subjectivity in scoring can affect careers, reputations, and perceptions of competitive integrity. We concluded that there's a correlation between scores judges awarded to specific aspects of a skater's performance and when they performed. We also concluded that judges have different impacts on a skater's score, despite not contributing the most variance in the multilevel models evaluated.

Data and EDA:

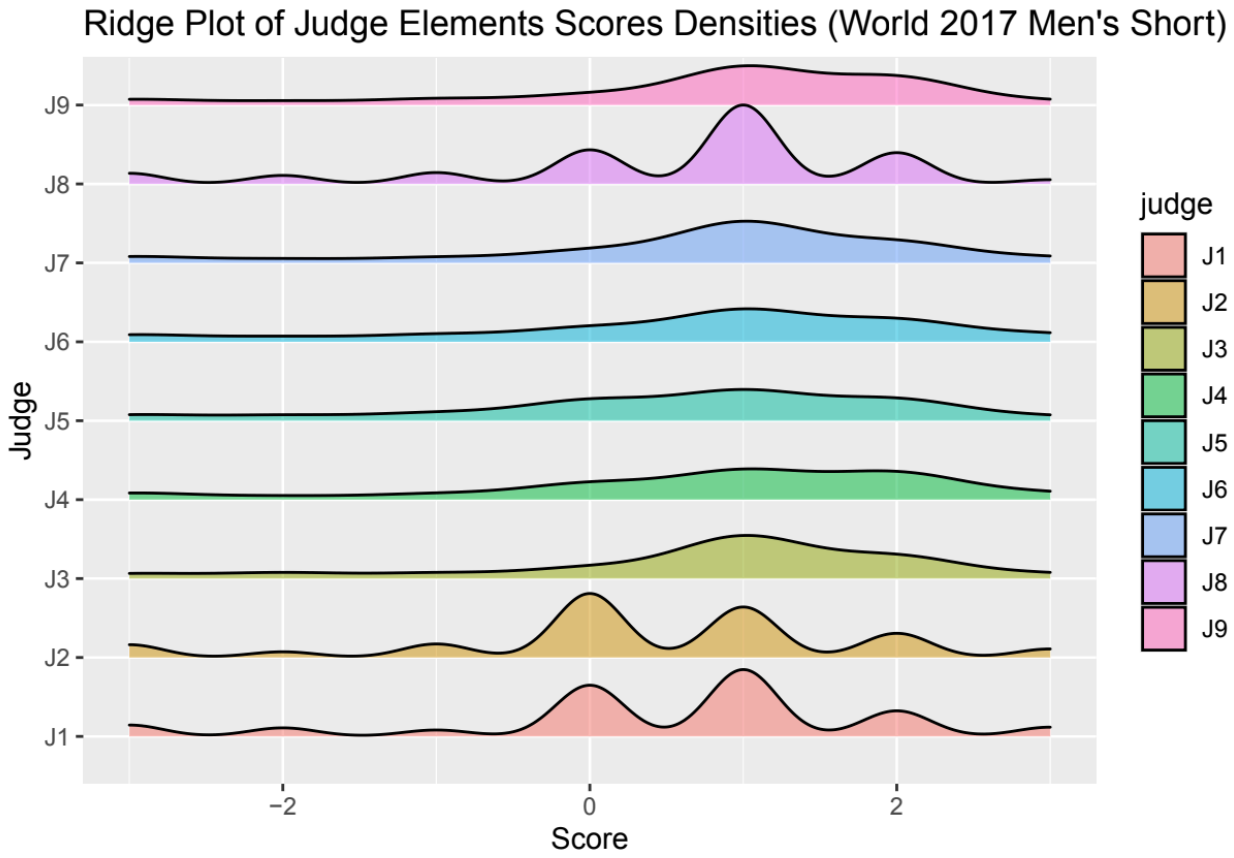
The skating data we used is from BuzzFeedNews/figure-skating-scores GitHub, which is available here. This dataset is derived from official documents that detail the scoring breakdown for each skater's performance published by the International Skating Union (ISU), and spans 17

major competitions from October 2016 to December 2018. The original PDFs with this information are public and downloadable.

This data is organized into four main CSV and JSON files. The file “programs.csv” contains the events for any given competition. The file “performances.csv” includes a row for each performance, detailing names, countries, and total score breakdowns. The file “judged-aspects.csv,” lists a breakdown of scores for each element and component. Finally, “judge-scores.csv” logs individual scores from each judge for every element and component, making it possible to analyze patterns in judging behavior.

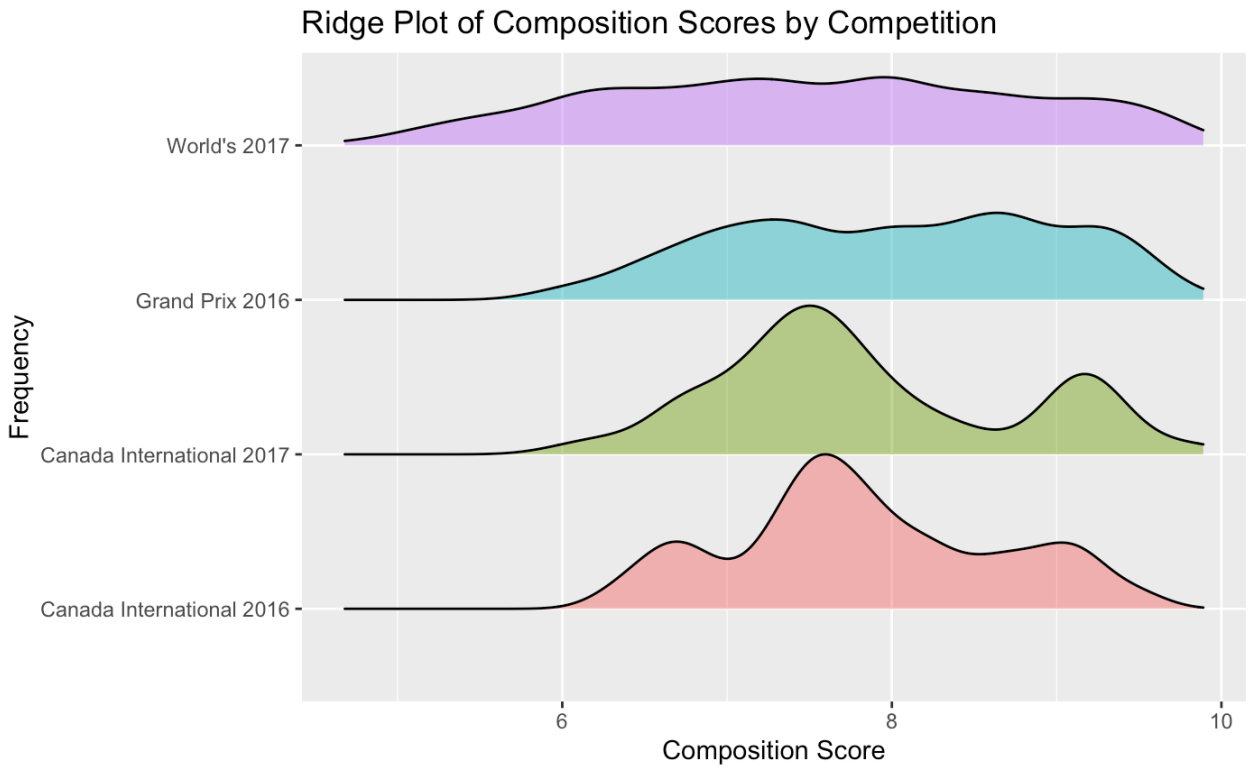
For preprocessing, we first combined the separate CSV files using shared performance IDs and scored aspects (elements and components). Next, we grouped the data by competition and event, which was essential to accurately separate out individual judges since they’re labeled as “J[number]” instead of their names. Thirdly, we separated scoring aspects by element and component scores, as they’re scored differently. Finally, for the analysis, we filtered data for only the 2017 World Championship, as it is the most elite competition with the most scrutiny and consistency, which is favorable for modeling purposes.

We first analyzed the densities of element scores awarded by each judge in different programs. For context, elements, which are the individual moves a skater does, can be scored negatively if they are done poorly. Components are general aspects of the performance, such as storytelling, and are only scored positively. While judges may differ on individual scores, it should be expected that the distributions of scores given are similar between judges. However, we found this to not be the case.



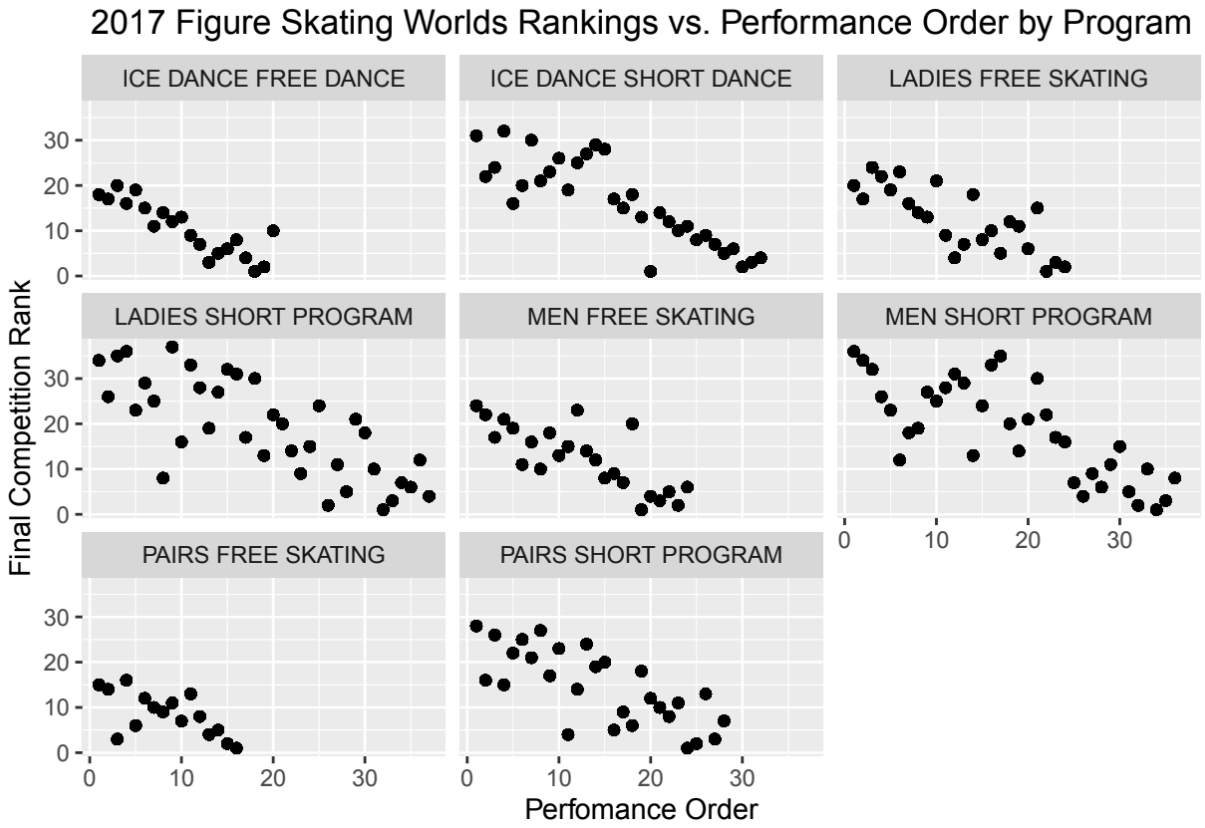
The densities of scores given by the nine judges for elements in the World's 2017 Men's Short Program are shown above. We observe that Judges 1, 2, and 8, give significantly more whole number scores than other judges, who have more flat distributions. These differences in scoring distributions were also found for the component scores, indicating potential inconsistencies based on specific judges.

Following this, we focused on the densities of scores given to specific aspects of a performance by competition. All baseline values for specific aspects are shared across competitions, so a routine with a storytelling score of 6 at the Grand Prix should also be given a 6 at World's. Therefore, we should expect the densities of scores given to the same aspect to be about the same across competition, but we again found this to not be the case.



The densities of composition scores for 4 different competitions are shown above. We observed that the scores are competition-dependent, as both Skate Canada International competitions had similar distributions, while the Grand Prix and World's had unique distributions. Interestingly, World's has the most uniform distribution and gave lower scores on average than other competitions, which were bimodal and generally had higher peaks, indicating potential scoring discrepancies. Based on this plot, we recognized that focusing on one competition would give the most consistent estimations.

Finally, we explored the relationship between the order a skater performed and their final ranking. At the 2017 World Championships, the bottom 50% would perform first, and the top 50% would perform last. However, the order within these groups was determined randomly. Therefore, we should expect skaters who went later to perform better, with some variation based on the random ordering.



The above plots show each skater's order and final rank for each program at the 2017 World's. Across all programs, there is a negative trend between performance order and final rank. The men's short program has the most variation from this trend, with a few skaters who competed between the 15th and 20th positions being some of the lowest-ranked skaters. This plot suggests that those who compete later in the competition tend to rank higher. Overall, these graphs indicate that a skater's starting order could be a significant factor in their final rank that should be included in our models.

Methods:

To predict figure skating outcomes and their variation based on judges, we decided to model the observation level score assigned by a judge for a move. Given the interest in observing the effect of a judge's bias in the scoring of an event, among other factors, two separate multilevel Gaussian models were fit, predicting judge scores for each component and element. For consistency, the women's free skating event at the 2017 World Championship was selected for this analysis. Based on the merged dataset, where each row is a specific element or component for one judge and one skater, both models include three levels. The first level is the observation level, which is a specific score given by a judge. The second level includes the judge-level random effects and the random effects for a move-skater combination. The third level captures the skater level effect. The exact model structure is specified below, where

n={1….N}, j={1….J}, and a={1….A}, and N is the number of skaters, J is the number of judges, and A is the number of moves performed.

Level One:

$$\text{score}_{jnai} = a_{jna} + b_j \times \text{startingnumber}_i + \beta_2 \,\text{nation}_i$$

Level Two:

$$a_{jna} = \beta_0 + u_{0,j} + v_{0,na} + a_n$$
$$b_j = \beta_1 + u_{1,j}$$

Random Effects:

$$\begin{pmatrix} u_{0,j} \\ u_{1,j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{u_{0j}} & \sigma_{u_{0j}u_{1j}} \\ \sigma_{u_{0j}u_{1j}} & \sigma^2_{u_{1j}} \end{bmatrix} \right)$$

$$v_{0,na} \sim \mathcal{N}(0, \sigma_v^2)$$

Level Three:

$$a_n = \alpha_0 + x_n$$

Random Effects:

$$x_n \sim \mathcal{N}(0, \sigma_n^2)$$

Composite Model:

$$\text{score}_{jnai} = \beta_0 + \alpha_0 + \beta_1 \times \text{startingnumber}_i + \beta_2 \text{nation}_i + u_{0,j} + v_{0,na} + x_n + (u_{1,j}) \times \text{startingnumber}_i$$

This model includes random intercept effects for the judges ($u_{0,J}$) and a nested effect for specific moves within each skater ($v_{0,na}$), along with a random slope effect for the effect of the starting order variable within each judge ($u_{1,J}$). This allows us to see different "biases" between judges. Fixed effects for the skating order (β_1) and a skater's nationality were included. Here, β_2 represents a vector of coefficients that change the intercept based on the difference in effect from the base country. These models assume a linear relationship of covariates, as well as Normal residuals and homoscedasticity. Additionally, the estimates of the random effects are assumed to be from a Normal distribution.

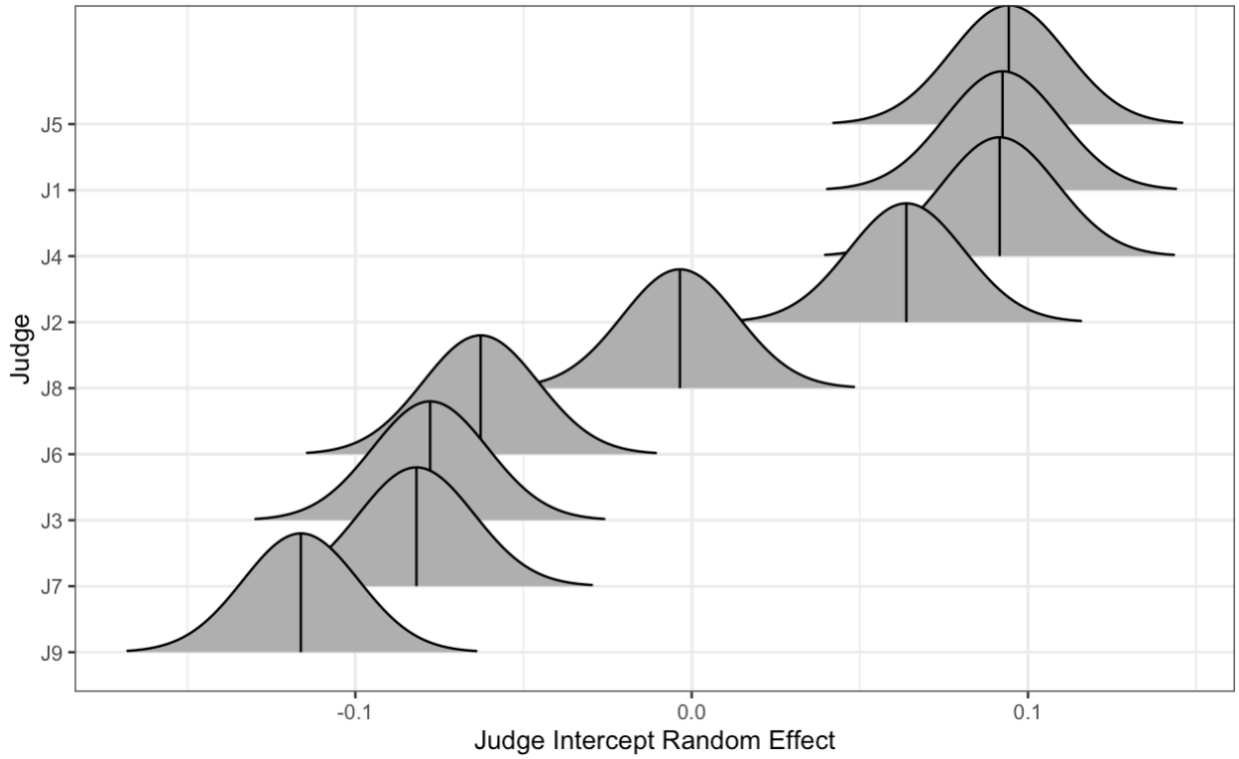
Similar models with rescaled data were fit assuming the Beta distribution and a logit link function to better model the bounded nature of the scoring data. These models have similar assumptions surrounding the random effects, while also assuming a relationship between the mean and variance of the model. Given that component scores fall between 0 and 10, and element scores fall between -3 and 3, this approach is ideal. Rescaling the data between 0 and 1 preserves the distance between data points while allowing the use of the Beta distribution. These models have a similar structure and specification to the previous models, though instead of directly modeling the response variable, we instead model the logit of the expected mean.

To estimate uncertainty, we used bootstrapping to resample our data within a given skater's observed moves by a judge to estimate distributions for the random effects. Through this method, distributions for the slope and intercept random effects concerning the judges could be examined and compared. This also allows us to quantify the uncertainty of our models.

Furthermore, general model fit for models of both types is measured through metrics such as AIC and RMSE, which allows us to compare these models.

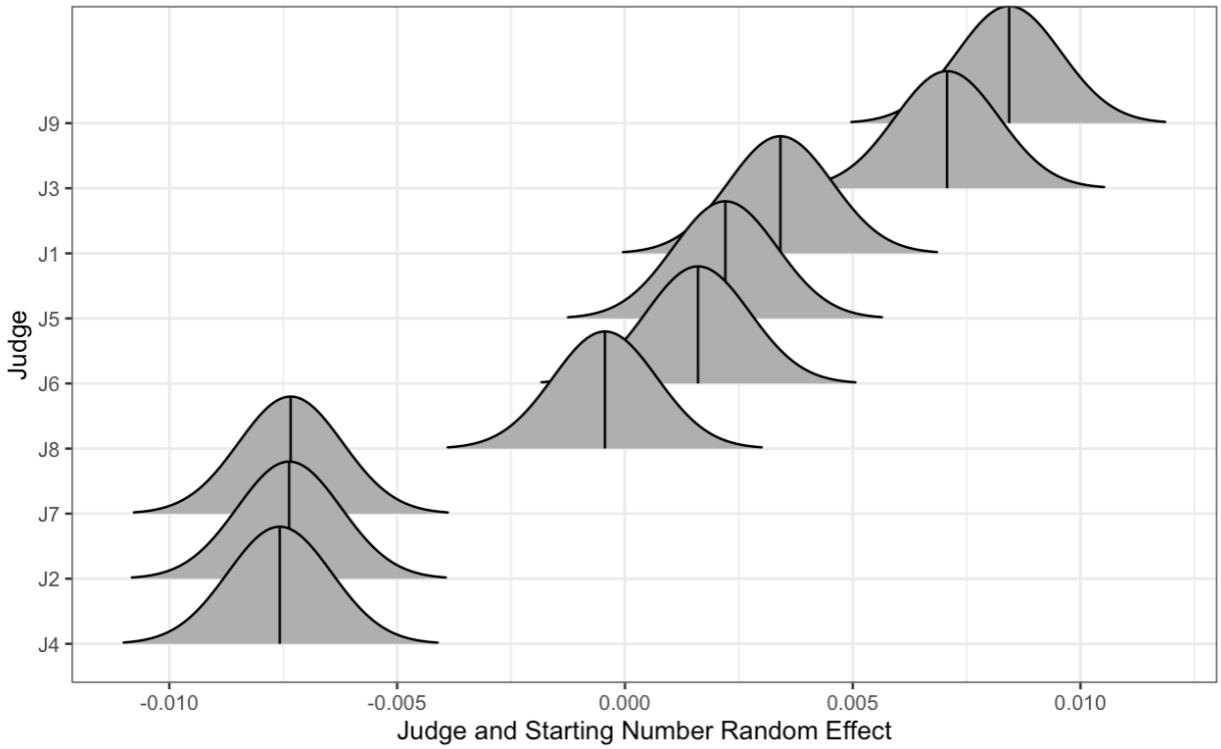
Results:

Starting with the Gaussian models, we found that none of the fixed effects for skater nationality nor starting order were significant. However, for the random effects, we did find a few interesting relationships. For both element and component scoring models, we observed strong correlations between the random intercepts and slopes for a skater's starting position and the judge evaluating them. These correlations are -0.5 for element scores and -0.89 for component scores. Our EDA showed that skaters who performed later tended to do better. However, this suggests that the slope for each judge *also* decreases as skating order increases, so individual judge scores would tend to converge later in the competition. Despite this, the move-skater random intercept had the highest variance at approximately 1.12 for the element scores and that the skater had the highest variance for the component scores at around 0.29. That said, we explored the potential impacts that judges have on scoring to discern any potential scoring biases.



Analyzing the estimated judge random intercepts for the element score model, we can see that there appears to be one judge whose random intercept estimate appears to be close to 0, and an even split for the remaining judges between negative and positive random intercept estimates. Given that all of these estimates overlap with at least one other distribution, it does not appear that there is one judge with a random intercept value that starkly differs from the others. That

said, we observe that there appear to be four judges whose "baseline" element score is negative, and four who are positive, implying some differences in scoring.



For the random slopes based on a skater's starting position in the element model, we see that the full range of random slopes is small, with the minimum being around -0.015 and the maximum being around 0.015. This aligns with the small variance found for these random slopes, at approximately 4.98×10^{-5} for the element score model. Additionally, we observe that more judges appear to have slopes that are closer to 0, though there are three judges whose random slope estimates are consistently negative.

Shifting to the Beta models, we observed many of the same results, with the highest variance for the element score model and component score model being the same (e.g., move-skater random intercept and skater random intercept). Specifically, the variances for these variables were 1.34 for the move-skater combination for the element model and 0.07 for the skater in the component model. However, unlike the Gaussian models, we did observe that the fixed effect for the nation of Italy was statistically significant. The estimate for this coefficient was 1.15, with a standard error of 0.50. On the other hand, in line with the previously discussed models, we did see strong negative correlations for the judges' random intercepts and slopes based on a skater's starting position. Though these variables did not contribute the most variation, they are still important to consider as a potential external influence on a judge's scores.

Finally, to compare these modeling approaches, we calculated both the RMSE and AIC values for the element and component score models that relied on the Gaussian assumptions and the Beta assumptions. These values are displayed in the following table:

	Beta Component	Beta Element	Gaussian Component	Gaussian Element
RMSE	0.001	0.01	1.11×10^{-5}	3.87×10^{-6}
AIC	-3818.84	-4585.42	1082.10	4623.42

Based on the table above, it is clear that, despite the Gaussian models having lower RMSE values, the Beta models have *much* lower AIC values. Given this information, we can conclude that our specific Beta models are better at modeling the figure skating scores than our chosen Gaussian ones. This is a sensible conclusion, as the scores given are bounded, while the Gaussian distribution is not.

Discussion and Conclusions:

Based on our analysis, we found that the specific moves and skaters showed the most variation in both the Gaussian and Beta models of figure skating scores, and the fixed effects for starting position and skater nationality were mostly insignificant. However, we did determine that there were some observable differences in the random effects and slopes for judges, along with strong negative correlations between these values. Using the comparisons of RMSE and AIC values, we discovered that, although the RMSE for the Gaussian models was lower, the Beta models had much lower AIC values. When comparing these two models, the AIC scores are generally the better route for comparison, which leads us to conclude that using the Beta distribution is best for modeling figure skating scores.

We recognize that there are limitations with this analysis. In particular, we lacked information about the judges themselves, such as their nationalities, which could lead to differences in their scoring behavior. Furthermore, this dataset only had two seasons of data, which prevented us from observing potential trends over time. These factors also made it difficult to evaluate the models out-of-sample, as two judges listed as "J1" might not have been the same across competitions. Finally, our analysis only considered information from one competition which might not show the full picture of factors affecting figure skating scores.

Looking forward, researchers should consider extending this analysis by using data from several competition seasons to model factors that may have changed over time. Additionally, by extending this work to include more information about the judges, other analyses could uncover additional biases, which could be based on judge and skater nationalities combined. Ultimately, we have shown that there is some variation in figure skating scoring based on external factors and determined a model that can be used to analyze these under-researched scores.

Citations:

"IOC Finds Fraud, Awards Second Gold in Winter Olympics Skating Event | February 15, 2002." History.Com, A&E Television Networks, 24 Jan. 2025,

www.history.com/this-day-in-history/February-15/winter-olympics-scandals-figure-skating-2002. Accessed 28 Apr. 2025.

Zitzewitz, Eric. "Does transparency reduce favoritism and corruption? Evidence from the reform of figure skating judging." Journal of Sports Economics 15.1 (2014): 3-30.