



Evidence of Similarity Preference Among Diving Judges in the Summer Olympics 2021

Is the Olympic Diving Judge Panels susceptible to Compatriot Preference?

Nokkvi Dan Ellidason, Shiyu Dou, Zehao Dou

19 October 2021

Abstract

A cognitive bias is a subconscious error in thinking that leads you to misinterpret information from the world around you and affects the rationality and accuracy of decisions and judgments. Cognitive biases are unconscious and automatic processes designed to make decision-making quicker and more efficient. One of the most popular sports at the summer Olympics games, diving has long been plagued with controversy due to its unique judging system. Diving presents avenues for biased scoring that don't exist in other summer Olympic sports. In the 2021 Olympics, the panel of judges consists of 25 judges, where 17 of them had a shared nationality of a competitor. The judges who seem to have compatriot preferences are, Canada, Germany, Ukraine, China, Mexico and Korea.

Contents

1	Introduction	2
1.1	Insights	2
1.2	Diving Knowledge Essentials and The Difficulty of Judging	2
2	Data	3
3	Statistical Methods	4
3.1	Average Calculation and Baseline T-Tests	4
3.2	Randomization With Permutation Test	5
3.3	Clustering By Country in Europe	6
4	Results and Discussions	7
5	Appendix	8
	References	10

1 Introduction

Many sports, such as gymnastics, diving, ski jumping, and figure skating, rely on judges' objective judgements to determine the winner of a competition. Judges usually follow a consistent rating scale (e.g., Diving: 0.0 - 10.0). Sport governing bodies have the responsibility of setting and enforcing quality control parameters for judge performance. Given the judging scandals in figure skating at the 1998 and 2002 Olympics, judge performance received greater scrutiny. The purpose of this article is to investigate if nationality can affect judges' grading in either direction in diving at the 2021 Olympic Games. Empirical studies have been conducted on judges' nationality bias. For example, Nationalistic Judging Bias in the 2000 Olympic Diving Competition (Emerson and Meredith 2010) and Racial Bias in National Football League Officiating (Eiserloh Dawson G., Foreman Jeremy J., Heintz Elizabeth C. 2020). This Case Study aims to stand on the shoulders of those giants and hopefully impact the realm of improving the judging system in sports by investigating the cognitive compatriot preference of judges. In this report we denote a judge and diver pair as "compatriot" if they share the same NOC code (Wikipedia, collective, n.d.), and "alien" if they do not share the same NOC code.

1.1 Insights

Although the total evolution of our species has lasted for years on end, humanity, as we know it today, is only about 200,000 years old. During this time, we have learned to take advantage of specialized psychological and sociological aspects that we can even continue through the millennia. These doings are rarely conscious but nevertheless, a vital part of our complex thought process, shaped by the long history of the species. It has long been known, for example, that familiar stimulus is seen as more attractive than exotic stimuli. This effect has been called "the blot-exposure effect" or "the familiarity principle" (The Decision Lab, n.d.). Most often, these factors reach evolving importance. The so-called familiarity principle has helped either an individual to survive or even to reach a certain a social role so that we can create large communities where solidarity and harmony are more important than an individual's well-being. Although these factors certainly have many positive characteristics, they involve a risk of systematic error in decision-making and interpretation. We may value what we know better than it is exotic.

For example, performance assessments need to be examined considering these effects. Performance evaluation is part of almost every workplace. Workers get evaluated by superiors, students by teachers, and team leaders by executives. This chain of evaluations creates an environment that makes most careers depend not only on the performance itself but also on the perception and evaluation of performance by others. Interestingly, there seems to be a sizeable possibility for biases in such judgments (Meyer, Mary and Booker, Jane 2001). An example of this is that in a recent study (Lyngstad, Torkild Hovde and Härkönen, Juho and Rønneberg, Leiv Tore Salte 2020), for example, found strong evidence of bias in sport performance evaluations in ski jumping.


1.2 Diving Knowledge Essentials and The Difficulty of Judging

In individual diving competitions, divers are evaluated by a panel composed of seven judges with strict rules for determining the diving excellence of each diver. After each dive, all judges evaluate the performance simultaneously. While judges know the other members of a jury, they neither can observe their evaluations nor are allowed to communicate with them. Scores range from 0-10 based on the execution and degree of difficulty. The highest and lowest scores from a divers six dives are excluded, while the rest are weighted for difficulty. Regarding grading these dives there are five elements the judges must evaluate. Each part of the dive is evaluated as part of the overall score a diver receives. They are starting position, approach, take off, flight (including overall height achieved), and entry into the water. The difficulty of evaluating these different scores for an overall score of a dive can be hard. The multivariate aspect of observing the dive and simultaneously evaluating different aspects of it exposes a possibility for an unconscious decision of preference.


Imagine five people in a room watching a movie about fencing and then immediately splitting the people into rooms where they tell you what they thought about the movie, you are most likely going to get five very different answers. Hence it is vital to investigate the cognitive preference of judges and try to choose judges panel based on the fairest result.

2 Data


The modern Olympics comprises all the Games from Athens 1896 to Tokyo 2021. The Olympics is more than just a quadrennial multi-sport world championship. It is a lens through which to understand global history, including shifting geopolitical power dynamics, women's empowerment, and the evolving values of society. The Olympics foundation has gathered data from the games for a long time but only recently started digging into the patterns of the data. The data this report uses comes from 24 (fairly*) identical PDFs made by the Olympics Committee and distributed to students enrolled in Statistical Case Studies at Yale University. From talking to Elliot Schwartz, the performance Data Liaison for the US Olympic Committee we have no reason to not trust that the data is consistent with results, names, and judges. A sample size of the competition data can be seen here below:



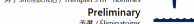
Tokyo Aquatics Centre
東京アクアティクスセンター
Centre aquatique de Tokyo




TOKYO 2020
MON 2 AUG 2021
Start Time 15:00



Diving
飛込 / Plongeon



Men's 3m Springboard
男子3m跳板飛込 / Tremplin 3 m - hommes



Preliminary
予備 / Eliminatoires

Detailed Results

結果詳細 / Résultats détaillés

Rank	Name	NOC Code	Dive No.	DD	J1	J2	J3	J4	J5	J6	J7	Dive Points	Dive Rank	Total Points	Overall Rank	Points Behind
1	WANG Zongyuan	CHN	407C	3.4	8.0	8.0	8.0	8.0	8.0	8.0	8.0	81.60	3	81.60	3	5.10
			5337D	3.5	9.0	8.5	9.0	8.5	9.0	8.0	8.5	91.00	1	172.60	1	
			5156B	3.9	8.0	7.5	7.5	8.0	7.5	8.5	7.5	89.70	1	262.30	1	
			307C	3.5	8.0	8.0	8.0	8.0	7.5	8.5	7.5	84.00	=5	346.30	1	
			207C	3.6	8.0	8.5	9.0	8.5	8.0	8.0	9.0	90.00	3	436.30	1	
			109C	3.8	8.0	7.5	8.5	8.0	8.0	8.5	8.5	95.00	4	531.30	1	
2	XIE Siyi	CHN	5154B	3.4	8.5	8.0	9.0	8.5	8.5	8.5	9.0	96.70	1	96.70	1	
			5353B	3.3	7.5	7.0	8.0	7.5	7.5	8.0	74.25	6	160.95	2	11.65	
			207C	3.6	8.5	7.0	7.0	6.5	7.0	73.80	8	234.75	3	27.55		
			307C	3.5	8.5	8.5	8.5	8.0	8.0	8.5	89.25	=1	324.00	3	22.30	
			407B	3.7	8.5	9.0	9.0	8.5	8.5	8.5	9.0	96.20	1	420.20	2	16.10
			109C	3.8	9.0	8.5	9.0	8.5	8.5	9.0	100.70	1	520.90	2	10.40	
3	PACHECO MARRUFO Rommel	MEX	5154B	3.4	7.5	7.5	6.0	7.5	7.5	8.0	76.50	5	76.50	5	10.20	
			205B	3.0	7.5	8.0	8.0	8.5	8.5	8.5	8.5	76.50	=3	153.00	3	19.60
			407C	3.4	8.0	8.0	8.5	9.0	8.5	7.5	8.0	81.60	4	234.60	4	27.70
			5353B	3.3	7.5	7.5	7.0	7.5	7.5	7.0	7.5	74.25	9	308.85	4	37.45
			307C	3.5	7.0	6.0	6.5	7.0	7.0	7.0	73.50	12	382.35	4	53.95	
			109C	3.8	8.5	8.5	8.0	8.0	8.0	8.5	96.50	=2	478.25	3	52.05	

Figure 1: Sample of Data

Along with this data we also have information about the gender, nationality of each judge in a panel of the round and event a dive is performed in. Combining these two datasets is done for the analysis. For statistical reasons the dives which scored zero have been excluded from the dataset. This is the case as it is clearly defined when a dive has been failed and scores a zero, little to no assessment of the judges goes into that and thus no personal preference to estimate. For example, the rule D 6.28 states that “When a diver refuses to execute a dive, the Referee shall declare a failed dive.” (Fédération Internationale De Natation 2017). In the original data there are four dives which scored zero, two of those dives can be seen here: *Arantxa Chavez failed dive* (s9v, Youtube Channel 2021) and *Pamela Ware failed dive phrase* (Arvi Viva, Youtube Channel 2021).

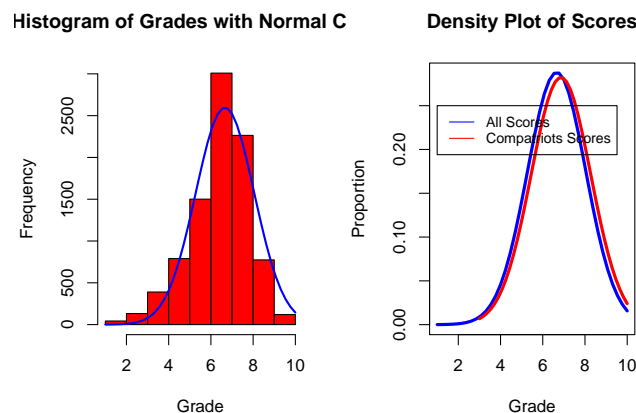


Figure 2: Left: Distribution of grades. Right: Difference in compatriots and alien judges

By Figure 2 we can see the average difference between the compatriot judges and alien judges. It may not seem a lot but when looked closer for each NOC code we can see that it may might be significant. This is what we will discuss in the chapter of Statistical Methods.

3 Statistical Methods

3.1 Average Calculation and Baseline T-Tests

In statistics, a null hypothesis, H_0 is a statement assumed to be true unless it can be shown to be incorrect beyond a reasonable doubt. The idea is that the null hypothesis generally assumes that there is nothing new or surprising in the population. Our test will have the following hypothesis schema:

H_0 : There is not a grade inflation over the difference between grades of compatriots and alien judges

H_1 : There is a grade inflation over the difference between grades of compatriots and alien judges

First we need to calculate the differences between the observations of non-compatriot grading and compatriots. The difference can be seen here below in visual form, and can be found in the table 6 in the appendix:

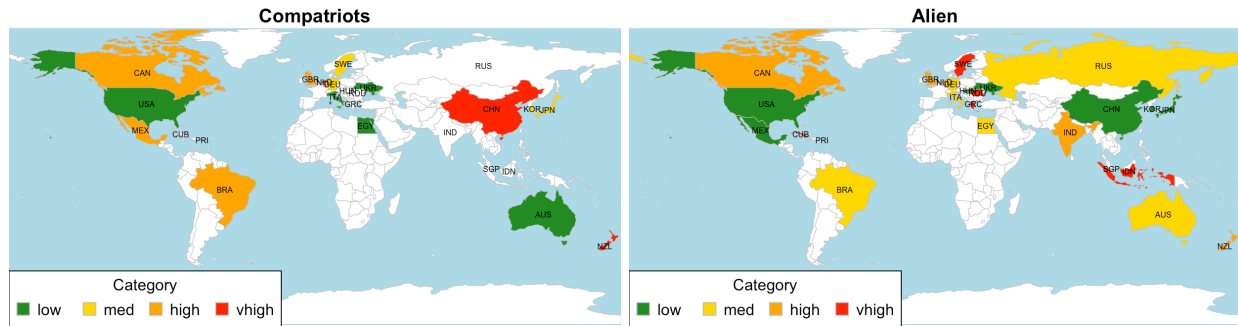


Figure 3: Averages for compatriots, aliens and its differences

The average total difference is 0.0740 where the biggest difference is 0.524 by the Brazilian judge. Regarding these calculations we need to take into consideration that we have way more observations of non-compatriot grading than compatriots.

Table 1: Observation Overview

	Proportion	Count
Compatriots	0.0317	286
Alien	0.9683	8737

We have noted the average difference of each compatriot and alien. However, need to get the difference from each judge on a panel compared to the dive so we can account for the dive strength done by diver. We do that in three ways.

Average of all 7 judges for a dive [Noted as 7]

Average of 6 judges for a dive (excluding the judge we are looking at) [Noted as 6]

Median of all 7 judges for a dive [Noted as Median]

Looking at the averages of these differences which we call blunders, along with the count of each group can be seen below:

Table 2: Compatriot Blunders Overview

NOC	Compatriot	Alien	Average Blunder 7	Average Blunder 6	Average Blunder Median
SWE	2	321	-0.086	-0.100	-0.255
EGY	8	373	0.036	0.041	-0.017
CAN	29	679	0.171	0.200	0.187
GER	22	416	0.217	0.253	0.249
NZL	3	418	-0.009	-0.011	-0.145
USA	37	449	0.031	0.036	0.020
JPN	15	273	-0.172	-0.200	-0.157
AUS	21	467	-0.004	-0.005	-0.063
UKR	13	365	0.212	0.248	0.287
CHN	46	494	0.266	0.310	0.241
MEX	31	369	0.255	0.297	0.242
ITA	7	339	0.079	0.093	0.054
GBR	21	316	0.030	0.035	-0.035
KOR	19	301	0.260	0.304	0.295
BRA	3	228	0.362	0.423	0.524
PUR	3	165	0.319	0.372	0.467
NED	6	156	0.127	0.148	0.112

Now we can use our hypothesis schema to answer our questions if those difference are statistically significant. The results from these t-tests can be seen here below:

Table 3: P-Values and Significance for Blunders - Baseline

NOC	Compatriot	Alien	p -value 7	p -val 6	p -val Median
SWE	2	321	0.621	0.621	0.753
EGY	8	373	0.359	0.359	0.557
CAN	29	679	0.000	0.000	0.001
GER	22	416	0.000	0.000	0.000
NZL	3	418	0.524	0.524	0.762
USA	37	449	0.276	0.276	0.345
JPN	15	273	0.905	0.905	0.885
AUS	21	467	0.523	0.523	0.800
UKR	13	365	0.021	0.021	0.011
CHN	46	494	0.000	0.000	0.000
MEX	31	369	0.000	0.000	0.002
ITA	7	339	0.148	0.148	0.247
GBR	21	316	0.375	0.375	0.636
KOR	19	301	0.009	0.009	0.004
BRA	3	228	0.062	0.062	0.105
PUR	3	165	0.116	0.116	0.123
NED	6	156	0.225	0.225	0.302

We can however not conclude that our work is done due to the fact that we do not have the same amount of data for compatriots as alien judges.

3.2 Randomization With Permutation Test

To counteract against the lack of data for compatriots we run a statistical model called permutation test. A permutation test gives a simple way to compute the sampling distribution for any test statistic, under the

strong null hypothesis. Thus we exclude the ones with non-significant p-values in all three differences (7, 6 and median). These are the countries we exclude:

Table 4: Countries That Have No Compatriots

NOC	SGP	CUB	RUS	HUN	IND	INA	GRE	ROU
Compatriot	0	0	0	0	0	0	0	0
Alien	537	465	316	234	423	197	251	185

Now we run a permutation test to determine the statistical significance of the model using our calculated p-values as baselines and then compute many random permutations of that data. If the model is significant, the original test statistic value should lie at one of the tails of the null hypothesis distribution. The results after the permutation test can be seen in visual form here below and in a table 7 in the appendix. In the plot, yes indicates that a judge from the country has compatriot preferences.

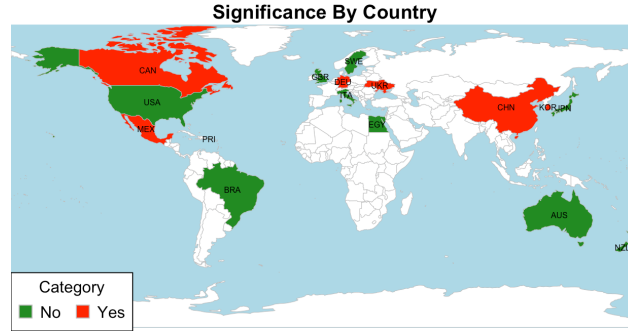


Figure 4: Significance of for judges that have compatriot preferences

From the plot above we can see that there are 6 judges which seem to showcase a similarity preference towards their compatriots. These judges are from the Canada, Germany, Ukraine, China, Mexico and Korea.

3.3 Clustering By Country in Europe

Now we have an idea of the similarity preference that the judges showcase and wonder if there is more under the surface. According to the article, Genes mirror geography within Europe (Novembre, J., Johnson, T., Bryc, K. et al. 2008), our biological structure is built the same way as our compatriots. Thus it might be the case that we also have similarity preferences to others within similar social structure. This part is more suitable for social studies and it takes derby rivalries, history of wars and other similar disputes into consideration. We however decide to do this by looking at the continents and look for similarity preference with in that. We split our data into Europe and others and the split is as follows:

Table 5: Europe Splitted Countries

Europe	EGY	GBR	GER	GRE	HUN	ITA	NED	ROU	RUS	SWE	UKR	NA	NA	NA
Others	AUS	BRA	CAN	CHN	CUB	INA	IND	JPN	KOR	MEX	NZL	PUR	SGP	USA

We now use the same methods as described above for each compatriots but only for Europe against non-Europeans. With the t-test baseline and the randomization using permutation tests we gather these significance relationships between the null hypothesis that grades are inflated by same continent (Europe) judges and divers.

The result here is surprising but nevertheless not unexpected due to the reasons mentioned above and the

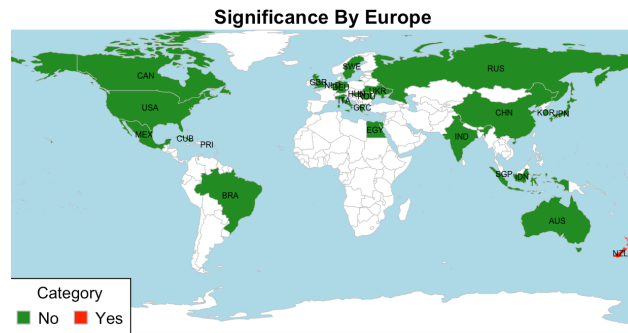


Figure 5: Significance of for judges that have Europe/Non-Europe preferences

fact that the judges are on average very competent. By the table above we see that only the judge from New Zealand seems to have similarity preferences towards the non-European divers. This chapter of clustering by countries should be looked at as an example of further possible analysis rather than exact results as it disregards many historical facts and countries evolution. within the countries

4 Results and Discussions

These kinds of tests give an idea of what can be concluded but disregard many other similarity preferences and thus can not be thought of as definite result. We are however sure of our statistical analysis. The judges that seem to have similarity preferences towards their compatriots are: Canada, Germany, Ukraine, China, Mexico and Korea.

As noted in the introduction the so-called familiarity principle has helped either an individual to survive or even to reach a certain social role so that we can create large communities where solidarity and harmony are more important than an individual's well-being. Thus, it might be interesting to look even further for similarity preferences by divers and judges characteristics such as height, ethnicity, gender and so on.

5 Appendix

Table 6: Averages for compatriots, aliens and its differences

NOC	Compatriots	Alien	Difference
AUS	6.3333	6.6724	-0.3390
BRA	6.8333	6.6469	0.1864
CAN	6.8103	6.6892	0.1211
CHN	8.4022	6.4170	1.9852
CUB	NA	6.9043	NA
EGY	6.2500	6.5898	-0.3398
GBR	6.6429	6.7073	-0.0644
GER	6.4773	6.5865	-0.1093
GRE	NA	6.9143	NA
HUN	NA	6.4017	NA
INA	NA	6.8883	NA
IND	NA	6.8014	NA
ITA	6.3571	6.5324	-0.1753
JPN	6.6000	6.4194	0.1806
KOR	6.5789	6.4535	0.1255
MEX	6.6774	6.4864	0.1910
NED	6.9167	6.7500	0.1667
NZL	7.3333	6.7955	0.5379
PUR	7.1667	6.7333	0.4333
ROU	NA	6.8514	NA
RUS	NA	6.5570	NA
SGP	NA	6.9134	NA
SWE	6.5000	6.9564	-0.4564
UKR	6.1923	6.4534	-0.2611
USA	6.4595	6.4566	0.0029
Total:	6.7371	6.6631	0.0740

Table 7: P-Values and Significance for Blunders - After Randomization

NOC	p -value (7)	p -value (6)	p -value (median)	$p < 0.05$
SWE	0.829	0.829	0.311	No
EGY	0.371	0.371	0.767	No
CAN	0.007	0.007	0.002	Yes
GER	0.001	0.001	0	Yes
NZL	0.965	0.965	0.772	No
USA	0.242	0.242	0.297	No
JPN	0.57	0.57	0.576	No
AUS	0.967	0.967	0.635	No
UKR	0.012	0.012	0.001	Yes
CHN	0	0	0	Yes
MEX	0	0	0	Yes
ITA	0.277	0.277	0.242	No
GBR	0.371	0.371	0.759	No
KOR	0.001	0.001	0.001	Yes
BRA	0.062	0.062	0.014	No
PUR	0.087	0.087	0.019	No



NOC	p -value (7)	p -value (6)	p -value (median)	$p < 0.05$
NED	0.194	0.194	0.188	No

Table 8: P-Values and Significance for Blunders - After Randomization (Europe)

NOC	p -value (7)	p -value (6)	p -value (median)	$p < 0.05$
SWE	0.279	0.279	0.184	No
EGY	0.116	0.116	0.027	No
SGP	0.838	0.838	0.724	No
CAN	0.89	0.89	0.413	No
CUB	0.773	0.773	0.865	No
GER	0.072	0.072	0.152	No
NZL	0.037	0.037	0.036	Yes
USA	0.154	0.154	0.126	No
JPN	0.658	0.658	0.623	No
AUS	0.184	0.184	0.205	No
RUS	0.5	0.5	0.887	No
UKR	0.061	0.061	0.187	No
CHN	0.256	0.256	0.264	No
MEX	0.327	0.327	0.374	No
HUN	0.339	0.339	0.248	No
IND	0.271	0.271	0.144	No
ITA	0.691	0.691	0.89	No
GBR	0.127	0.127	0.08	No
KOR	0.812	0.812	0.746	No
INA	0.033	0.033	0.064	No
BRA	0.112	0.112	0.078	No
PUR	0.098	0.098	0.167	No
GRE	0.017	0.017	0.088	No
ROU	0.365	0.365	0.375	No
NED	0.612	0.612	0.706	No



References

- Arvi Viva, Youtube Channel. 2021. “Pamela Ware Failed Dive (Scores Zero).” Youtube.
- Eiserloh Dawson G., Foreman Jeremy J., Heintz Elizabeth C. 2020. “Racial Bias in National Football League Officiating” 5: 48.
- Emerson, John W., and Silas Meredith. 2010. “Nationalistic Judging Bias in the 2000 Olympic Diving Competition.”
- Fédération Internationale De Natation. 2017. “Fina Diving Rules 2017-2021.”
- Lyngstad, Torkild Hovde and Härkönen, Juho and Rønneberg, Leiv Tore Salte. 2020. “Nationalistic Bias in Sport Performance Evaluations, An Example from the Ski Jumping World Cup” 17 (3): 250–64.
- Meyer, Mary and Booker, Jane. 2001. “Eliciting and Analyzing Expert Judgement, A Practical Guide.”
- Novembre, J., Johnson, T., Bryc, K. et al. 2008. “Genes Mirror Geography Within Europe” 456: 98–101.
- s9v, Youtube Channel. 2021. “Olympic Fail Diving Arantxa Chavez.” Youtube.
- The Decision Lab. n.d. “Why Do We Prefer Things That We Are Familiar With?” The Decision Lab.
- Wikipedia, collective. n.d. “List of IOC Country Codes.” Wikipedia.