Journal of Sports Analytics 6 (2020) 1–11
DOI 10.3233/JSA-200272
IOS Press

Revisiting difficulty bias, and other forms of bias, in elite level gymnastics

Kurt W. Rotthoff*
Department of Economics and Legal Studies, Seton Hall University, Stillman School of Business, South Orange, NJ, USA

Abstract. Difficulty bias was found in Morgan and Rotthoff (2014), claiming that trying something harder leads to a higher overall ranking, *ceteris paribus*. This was measured in elite level gymnastics, which at some meets allows for a unique measurement of this form of bias. Multiple studies also find evidence of an overall order bias. I analyze if these two forms of bias still exist in elite level gymnastics. I continue to find evidence of difficulty bias with this updated dataset and find strong evidence that these biases are different across genders – females have evidence of a large and significant difficulty bias, whereas males do not. However, the evidence of an overall order bias does not exist in this dataset.

Keywords: Difficulty bias, Judging bias, Reference bias

JEL: L10, L83, D81, J70, Z2

1. Introduction

Judging bias can have effects on the outcomes of competitions as well as the strategic behavior of those involved in the competition itself. In Morgan and Rotthoff (2014), they find that there is a difficulty bias; people who attempt more difficult tasks, in their case gymnastics routines, are artificially awarded higher scores. They employ a unique data set from elite level gymnastics that allows for the ability to test for this difficulty bias. At that point in time, there was only one gymnastics meet, the 2009 World Gymnastics Competition (in London, England), that employed both the new scoring system, separating difficulty and execution scores, and did not have a team competition, which has a built-in sequential order bias, simultaneously.

This type of gymnastics meet occurs once every four years. Thus, the 2013 World Artistic Gymnastics Championships, which was held in Antwerp, Belgium, is the second event that includes the unique features that allow for testing these forms of bias. This competition again has the structure of the Morgan and Rotthoff (2014) study: no team competition while utilizing the new scoring system (which allows for a random assignment of athletes to events). With this data, I now have the ability to update the findings in their study to test if difficulty bias still exists, or if their finding was simply an anomaly.

I also test for evidence of an overall order bias (i.e., going earlier or later in an event impacts their overall rank), which is found in Flores and Ginsburgh (1996), Gleiser and Hendels (2001), Bruine de Bruin (2005), Page and Page (2010), Morgan and Rotthoff (2014), and Rotthoff (2015). Many studies have found evidence of an overall order bias – when an athlete competes impacts their score. This finding, however, is not present in the current data.

Biased judging can impact the scores of a gymnast, as well as the strategic behavior of how they approach a meet (it is argued that at least one specific country tries artificially high difficulty levels, sacrificing their execution score, as an optimal behavior; showing that they are changing their strategy in a manner

*Corresponding author: Kurt Rotthoff, Seton Hall University, JH 674, 400 South Orange Ave, South Orange, NJ 07079, USA. Tel.: +1 973 761 9102; E-mail: Kurt.Rotthoff@shu.edu.

ISSN 2215-020X/20/$35.00 © 2020 – IOS Press and the authors. All rights reserved
This article is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License (CC BY-NC 4.0).

consistent with the findings in Morgan and Rotthoff, 2014). These efficiencies and strategic decisions have impacts on gymnastics, diving, and figure skating, as well as non-sports behaviors such as orchestra auditions, job interviews, marketing pitches, and even academic publishing (should an author make a given paper look artificially difficult to impress a referee/editor). I continue to find evidence of a difficulty bias and have data that is detailed enough to use athlete level fixed effects - finding significant evidence of a difficulty bias that is driven by women's gymnastics (and no evidence of a difficulty bias in men's gymnastics). And contrary to many previous studies, I do not find any evidence of an overall order bias.

The next section looks at the existing literature on potential biases in sequential order events. Section three presents the data and methodology. Section four presents the results for difficulty bias and overall order bias. Section five concludes.

2. Forms of bias

In this study, I test for two of the major forms of bias found in the literature: overall order bias and difficulty order bias. Although there is a lot of overlapping research within these measurements of bias, I look at these biases individually below with a brief discussion of some other forms of bias.

2.1. Overall order bias

Overall order bias is typically referred to as primacy and recency in the psychology literature; which is defined by the idea that it is either better to go first (primacy) or last (recency), but not in the middle, when being judged. This leads to a representative U-shaped function (Gershberg and Shimamura 1994, Burgess and Hitch 1999, and Mussweiler 2003). However, it is not clear that the U-shaped function is found in the economics literature. It is argued that there is an efficient suppression of early scores in Rotthoff (2015). Assuming that positions and abilities are randomly distributed, the probability, p, of any individual in the population, n, being the best is presented in the simple equation: $p = 1/n$. Thus the odds that the first person is the best are low, even if they perform very well. Because of this, judges efficiently withhold the highest scores for later in the competition; in the case a better performance comes later.

Overall order bias is found in both end-of-sequence judgments and step-by-step procedures by Bruine de Bruin (2005) in the "Eurovision" song contest and the World and European Figure Skating Contest. End-of-sequence judgments do not require final scores to be given until after every person has competed. However, step-by-step procedures are judged after each individual performs. They find that both contests have an overall order bias, specifically finding that ranking increases towards the end of a group. This finding is separate from another form of bias found in step-by-step judgments, which is termed sequential order bias – the immediately preceding contestant impacts the next person.

Flores and Ginsburgh (1996) and Glejser and Hendels (2001) find that it is optimal for an individual to be towards the end of the order. Page and Page (2010) also find that contestants in the middle, particularly closer to the beginning, are at a disadvantage. Meanwhile, those who are last have an advantage and are judged with higher average scores.

Sequential order contests, such as gymnastics, are suspected to have some serial position effect. Both Morgan and Rotthoff (2014) and Rotthoff (2015), henceforth MRR, found evidence of an overall order bias. They both find, using the same dataset, that going later in the competition is advantageous to the gymnast. Specifically, Rotthoff finds that going in the first session of the day, typically about one-fourth of the sample, is detrimental to the gymnast, whereas anyone going after that point has no differential impact.

The suppressions of the highest scores early in a competition can be magnified when there are fixed ceilings placed on scoring. This occurs in gymnastics because although they have separate scores for execution and difficulty, the execution score is maxed at 10.0, whereas the difficulty score is theoretically infinite. Judges use an efficient suppression of scores, a function of the expected distribution of the quality of future participants and the number of individuals left, when they are limited by a fixed ceiling on scores. The closer judges get to the end of a group, the less impact the score ceiling effect has on participants. The tendency to withhold top rankings from beginning competitors in larger groups explains the overall order bias found and its J-shaped curve.

When examining overall order bias, it is important to distinguish between that and sequential order bias. Overall order bias will result in escalating scores throughout a competition. Sequential order bias is when one contestant's score directly impacts the next contestant's score. There is no evidence of a sequen-

tial order bias in this data (not reported for brevity), so this study focuses on the overall order bias found in the literature.

2.2. Difficulty bias

Following the 2004 Olympic Games, the gymnastics governing body (FIG - Federation Internationale de Gymnastique), after an apparent judging controversy, completely overhauled the scoring system for elite level gymnastics. Gymnastics judges now issue separate execution scores and difficulty scores, determined by completely separate panels of judges. The change occurred after the 2005 World Artistic Gymnastic Championships, which means the first elite level competition with both no team event and the new split scoring system was the 2009 World Artistic Gymnastic Championships. With this data, Morgan and Rotthoff (2014) found evidence of a difficulty bias. Those gymnasts that attempted more difficult routines received an artificial bump in the execution score, even though these scores were determined (and designed to be scored) completely independently.

These findings have major implications for the accuracy of judged competitions that can have different levels of difficulty, as well as the incentive effects of contestants responding to these different forms of bias. This data provides examples in gymnastics, but this is true in other judged sports (diving, figure skating, cheerleading, etc.). However, this also applies the presentation of marketing pitches, the authoring of academic articles, to debates, in interviews, and in musical or acting auditions – in all these cases, if a difficulty bias exists, the person will artificially try/present more difficult skills/statistics/facts purely in an attempt to impress those passing judgment on their performance.

Given that 2009 was the first year that unique data existed to measure this bias, I continue to explore if this bias exists in the second year that this unique data exists, the 2013 World Artistic Gymnastic Championships. This is only the second setup of this data, because in the years following the 2009 World's Championship, each year's World Championships (2010 and 2011), also held a team competition, which biases the structure (order) of the contestants. In 2012, the Olympics were held (instead of a World's Championship, also with a team competition), leaving the 2013 World Championship competition as the second set of data structured to measure for difficulty bias accurately.

2.3. Other forms of judgment bias

There are other forms of bias in the literature that also need to be acknowledged, including racial bias, gender bias, reputation bias, and sequential order bias. These other biases arise in aesthetic sports, such as gymnastics, because panel judging entails human judgment, which inherently creates nonperformance-based bias outside of the evaluation criterion, potentially influencing the scoring and judging process (Landers, 1970; Moormann, 1994).

Racial preference has been shown by referees in basketball (Price and Wolfers 2010) and baseball (Parsons, Sulaeman, Yates, and Hamermesh 2011). Glejser and Heyndels (2001) find that women (and contestants not from the Soviet Union before 1990) received lower scores in piano in "The Queen Elisabeth Musical Competition." Other studies have found evidence of a nationalistic bias in figure skating: Seltzer and Glass (1991), Campbell and Galbraith (1996), Sala, Scott, and Spriggs (2007), and Zitzewitz (2006). Gift and Rodenberg (2014) have even expanded the referee bias into a height bias, finding a Napoleon Complex in basketball.

There is also empirical evidence that judges are influenced by a specific aspect of a routine (Auweele, Boen, Geest, and Feys, 2004). This can be thought of as a reputation bias, or expected performance level, of a given athlete (Auweele, Boen, Geest, and Feys, 2004, and Ste-Marie, 2004). Morgan and Rotthoff (2014) control for reputation to, at least in part, capture this effect.

Rotthoff (2015) expands the work done by Damisch, Mussweiler, and Plessner (2006) who look at sequential order bias in elite level gymnastics. Damisch, Mussweiler, and Plessner find that there is a positive correlation between a given gymnasts performance and the subsequent performance. However, as argued in Rotthoff, they use the final rounds of a gymnastics meet, which traditionally places the lowest scoring performances from the morning first and the highest scoring performances last. Rotthoff analyzed this sequential order bias in the original gymnastics dataset used by Morgan and Rotthoff (2014), finding no evidence of a sequential order bias (arguing that the unique structures of the World's data allowed for a more efficient estimation of this bias). I do not report the results in this paper, but following Rotthoff with this updated dataset there continues to be no evidence of a sequential order bias.[1]

[1] For these results, please contact the author.

Table 1
Women's events

Summary statistics (women)				
Variable	Vault	Uneven Bars	Balance Beam	Floor
Participants	105	100	109	103
Mean Difficulty Score	5.06	5.06	5.20	5.20
Standard Deviation of Difficulty Score	0.80	0.89	0.65	0.52
Mean Execution Score	8.59	7.22	6.90	7.27
Standard Deviation of Execution Score	0.94	1.01	0.98	0.71

Table 2
Men's Events

Summary statistics (men)						
Variable	Parallel bars	High bar	Rings	Floor	Vault	Pommel horse
Participants	141	134	134	135	120	148
Mean Difficulty Score	5.56	5.45	5.51	5.75	5.20	5.36
Standard Deviation of Difficulty Score	0.86	1.07	0.86	0.67	0.64	0.90
Mean Execution Score	8.02	7.69	7.94	7.89	8.87	7.40
Standard Deviation of Execution Score	0.81	0.69	0.68	0.81	0.44	1.04

Table 3
Normalized data for all events

Variable	Obs	Mean	Std. Dev.	Min	Max
Overall Order	1,239	63.58	37.66	1	149
Order-squared	1,239	5,459	5,200	1	22,201
Normalized Difficulty Score	1,239	0.00	1.00	–6.36	2.55
Normalized Execution Score	1,239	0.00	1.00	–9.14	2.02
Superstar	1,239	0.08	0.27	0	1
Male	1,239	0.66	0.47	0	1

3. Data and model

Following MRR, I use data from elite level gymnastics. The unique aspect of the world championships is that the year immediately following the Olympics the event is held with no team competition. This aspect occurs because the Olympics is the highest level team competition in elite level gymnastics, so the year following the Olympics, they have no team competition and focus on the coming out of the next round of gymnasts that will be targeting the upcoming Olympic Games.

In Tables 1 and 2, I present the summary statistics for the women and men's events. There are over 100 participants in each of the four women's events and more than 120 in each of the six men's events. However, and again following MRR, I normalize the data because there are different means and standard deviations across the various events (normalized to a mean zero, standard deviation one). The normalization occurs separately within women and men, thus there are two sets of data that have been normalized. This also allows for more tests; now I can combine the men's and women's data to increase the observations and increase the accuracy of our estimates.[2] The normalized results are in Table 3.

This data has no team competition, argued to be one of the reasons Damisch, Mussweiler, and Plessner (2006) get a positive result in sequential order bias (which is refuted in Rotthoff, 2015). Also, this competition is unique in that the assignment of each athlete's starting position is randomly assigned throughout (in their session of the day, rotation within each session, and the position in a given rotation). These structures allow for an unbiased estimate of the multiple forms of bias found in the literature.

It is also possible that there are superstar countries that continually place gymnasts in meets that perform higher difficulty routines and higher levels of execution, on average, than those from other countries. Both MRR papers use a proxy to capture this reputation effect from athletes in these types of superstar

[2] I also run the results of men and women separately in case the different events, or differences in genders, could drive different results.

countries (Tables A1 and A2, in the appendix, provide the countries that are used in this reputation measure for this study).

Although MRR also controls for the country of origin from the judges on the execution panels, for this year I do not have that information (and those studies find no relationship between the judges' country and athletes' country). The focus of this study is the measure of overall order bias and difficulty bias; thus the lack of judging countries will not bias our result assuming that they still have no impact on these forms of bias.

As a proxy for overall order bias, in equation 1, I include the overall performance *order* as a measure of a given athlete's relative place in the competition (*order* is a list of 1, 2, 3, ... , *n*; where n is the final gymnast). I also include an o*rder squared* term to allow for a non-linear relationship. It is also possible that there are a few very talented individuals that are influencing the results, so I control for those athletes that come from superstar countries (as a form of reference bias). The *E* vector controls for event specific fixed effects and also include country-level fixed effects, *C*. I estimate the following for each athlete, *i*, aggregating all events, for both men and women, together and separately (and eventually add athlete level fixed effects):

$$ExecutionScore_i = \beta_0 + \beta_1 Order_i + \beta_2 Order_i^2 + \beta_3 Superstar_i + \delta E + \phi C + \varepsilon \quad (1)$$

To measure the existence of difficulty bias in the judge's decision, I add *difficulty* and *difficulty squared* terms in equation 2. This will be the measure to see if the difficulty bias, first found in Morgan and Rotthoff (2014), still exists in the more recent data.

$$ExecutionScore = \beta_0 + \beta_1 Order_i + \beta_2 Order_i^2 + \beta_3 Superstar_i + \beta_4 Difficulty_i + \beta_5 Difficulty_i^2 + \delta E + \phi C + \varepsilon \quad (2)$$

Recall that the execution and difficulty scores are, by rule, determined by judges on separate panels. The execution score has a maximum score of 10.0 and is designed to measure only the execution of the routine. The difficulty section scores the person for the quality of the routine and is theoretically infinite. If difficulty and execution scores are positively related (with the stated controls), this reveals that difficulty bias still exists in the judging process.

The difficulty and execution judges are on different panels and are different people. Each event has their own panel of judges, and these judges remain with the same event throughout the competition. The execution judges are strictly judging how a contestant performs. The difficulty panel is focused on how difficult the routine executed is (i.e., did they connect certain elements on the beam, make a full rotation on the vault or floor, or hit the handstand on the bars). Although the athlete can change the difficulty of their routine, they often do not (at least not intentionally). Given that the judges have judged many competitions, and there are only so many elite level gymnasts, most judges have seen most of these athlete's routines before (or, at a minimum, know of their routine before they perform because they have warmed up the routine in front of the judges). Thus, it is easily assumed within the elite gymnastics world that these judges know what is coming regarding difficulty.

Although overall order bias can be measured in equations 1 and 2, Rotthoff (2015) uses another proxy to measure the existence of overall order bias. I present the use of sessions, as in Rotthoff, as another measure of overall order bias (where *session* is the block in which the gymnast competes in the day, not testing the order within each session). He finds that it is statistically more valuable to not be in the first session of the day, the excluded group in equation 3. I continue to control for the forms of difficulty bias and include both event and athlete-level fixed effects.

$$ExecutionScore = \beta_0 + \beta_1 SecondSexxion_i + \beta_2 ThirdSession_i + \beta_3 FourthSession_i + \beta_4 Difficulty_i + \beta_5 Difficulty_i^2 + \delta E + \varphi C + \varepsilon \quad (3)$$

4. Difficulty bias and overall order bias

The difficulty score evaluates the content of a routine. The athlete determines this score, as they decide what level or complexity of a routine to perform. Theoretically, the score is infinite and exogenous to the judges because they do not determine what difficulty the gymnast will perform[3]. The execution score is determined exclusively by judges on the execution panel. It evaluates how well an athlete performs based on a max score of 10.0 with points deducted for errors

[3]Given the difficulty attempted is a choice variable by the athlete, this can lead to estimation problems. I do my best to control for factors that can influence their decisions and also include athlete level fixed effects to see if there is evidence of this bias across a given athlete's performance.

in technique, form, execution, artistry, and overall routine composition. The two sets of scores are given by separate judging panels. Although the difficulty being attempted in a vault routine is posted prior to the performance, there is still a difficulty panel that confirms, or changes, the difficulty score based on what is actually performed.

Once a gymnast completes their routine, the difficulty score and execution score are added together, with penalties taken out, to give the final mark. For each contestant, their scores are posted immediately following their routine, before the next contestant. The execution score can capture any bias in judging where it exists because the score is completely decided by the judges on the execution panel.

With two different judging panels, I can measure the impact of a difficulty bias. By controlling for the other known biases in the data, I can test if this form of bias still exists in the data (originally found in Morgan and Rotthoff, 2014). Column 1 includes overall order, overall order squared, and superstar controls and includes event level fixed effects. I find that athletes from superstar countries score significantly higher execution scores in their routines and that going later in the competition is better for the athlete. However, the overall order bias disappears when including measures for the difficulty bias in column 2. When including the normalized difficulty score, I find that overall order bias is no longer significant, but the evidence of difficulty bias still exists. A difficulty bias is when an athlete's execution score increases with an increase in the difficulty of their routine; attempting a more difficult routine artificially inflates the execution score. This is also true when including the normalized difficulty bias squared term in column 3. Although it is increasing at a decreasing rate. In columns 4 and 5, I add country-level fixed effects. Continuing to find no evidence of an overall order bias, but finding significant evidence of a difficulty bias. As long as a difficulty bias is controlled for, the evidence of an overall order bias is non-existent.

In events where the difficulty is being determined by a judge, the difficulty bias could have an impact on the results. This can also impact the decisions made by the gymnast being evaluated. For example, in this meet, Larisa Iordache received fourth place in the Women's All-Around. If she would have attempted a one standard deviation increase in her difficulty on the bars, *ceteris paribus*, she would not only have received a 0.89 more points on her difficulty score, (which would have still placed her fourth) but she would have received a bump in her execution score. Due to the difficulty bias, that would have given her enough points for third place in the competition – with the difficulty bias moving her onto the podium.

4.1. Athlete level fixed effects with a difficulty and superstar interaction

To add more confidence in the baseline regression presented here, I include two measures to Table 4: athlete level fixed effects and an interaction of difficulty bias and the super star measure in Table 5. The former allows a measure of the variation within a given athlete's performances across events – testing if there are differences across the athletes' that could

Table 4
Estimating Execution Score

	Execution score				
	(1)	(2)	(3)	(4)	(5)
Order	0.0052*	0.0027	0.0030	0.0036	0.0043
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Order2	–0.0000*	–0.0000	–0.0000	–0.0000	–0.0000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Superstar	0.7545***	0.4273***	0.4721***	0.1414	0.1856
	(0.103)	(0.104)	(0.106)	(0.119)	(0.119)
Normalized		0.2922***	0.2693***	0.1772***	0.1241***
Difficulty Score		(0.028)	(0.030)	(0.036)	(0.038)
Normalized			–0.0357**		–0.0607***
Difficulty Score2			(0.017)		(0.016)
Constant	–0.1763	–0.0677	–0.0459	–0.4913***	–0.4643***
	(0.114)	(0.110)	(0.110)	(0.132)	(0.131)
Event FE	Yes	Yes	Yes	Yes	Yes
Country FE	No	No	No	Yes	Yes
Observations	1,239	1,239	1,239	1,239	1,239
R-squared	0.045	0.122	0.125	0.275	0.283

Standard errors in parentheses. $^{***}p<0.01$, $^{**}p<0.05$, $^{*}p<0.1$.

Table 5
Including athlete level fixed effects, an interaction of normalized difficulty score, and event level fixed effects

Execution score: Testing reputation		
	Interaction	Interaction
	(1)	(2)
Order	–0.0076	–0.0078
	(0.005)	(0.005)
Order2	0.0000	0.0000
	(0.000)	(0.000)
Superstar	0.0086	–0.1058
	(0.206)	(0.203)
Normalized	0.1457***	0.0117
Difficulty Score	(0.047)	(0.052)
Normalized		–0.1101***
Difficulty Score2		(0.020)
Normalized Difficulty	–0.0016	0.2445
Score x Superstar	(0.174)	(0.177)
Constant	0.2487	0.3811*
	(0.197)	(0.195)
Event FE	Yes	Yes
Athlete FE	Yes	Yes
Observations	1,239	1,239
R-squared	0.017	0.053
Number of IDs	392	392

Standard errors in parentheses. $^{***}p<0.01$, $^{**}p<0.05$, $^{*}p<0.1$.

be driving these results (i.e. are those of higher level, who can complete more difficult routines at a lower cost, also the ones getting higher execution scores). This means that the reputation variable cannot be included because there is no variation of the country of representation within athlete. I continue to find evidence of a difficulty bias and the normalized difficulty score and the normalized difficulty score squared are jointly significant in the second column, with results that are similar to the results before including athlete level fixed effects. The later (interaction of difficulty bias and the super star) will capture if the difficulty bias measured in the previous regressions is a difficulty effect driven by the athletes from superstar countries.

4.2. Separating by gender

Although the pooled sample provides evidence of bias, the sample is pooling the male and female gymnasts together. Given that these biases can have a different impact across the genders, I continue by splitting the sample by gender and when controlling for athlete level fixed effects and find that the evidence of difficulty bias in male athletes' becomes insignificant in Table 6. However, the results for the women continue to show strong and consistent evidence of a difficulty bias. As a matter of fact, when taking the male athletes out of the sample, the impact of the difficulty bias on female scoring increases in magnitude as well – it has a much larger impact than any of the previous estimates. This shows that not only does difficulty bias exist, as found in Morgan and Rotthoff (2014), but arguably there are gender differences in

Table 6
Separating out the female and male athletes, including athlete level fixed effects, an interaction of normalized difficulty score, and event level fixed effects

	Execution score			
	(1)	(3)	(4)	(6)
Order	0.0016	0.0008	–0.0098*	–0.0094*
	(0.013)	(0.013)	(0.005)	(0.005)
Order ^{2}	–0.0000	–0.0000	0.0001	0.0001
	(0.000)	(0.000)	(0.000)	(0.000)
Superstar	0.0254	–0.8105**	0.0506	0.2080
	(0.293)	(0.392)	(0.168)	(0.238)
Normalized	0.1951***	0.1179	–0.0093	0.0066
Difficulty Score	(0.070)	(0.073)	(0.068)	(0.070)
Normalized	–0.1470***	–0.1672***	0.0371	0.0456
Difficulty Score^2	(0.021)	(0.021)	(0.036)	(0.037)
Normalized Difficulty		0.8488***		–0.2112
Score x Reputation		(0.269)		(0.225)
Constant	0.1932	0.2552	0.2965	0.2854
	(0.414)	(0.408)	(0.210)	(0.211)
Event FE	Yes	Yes	Yes	Yes
Athlete FE	Yes	Yes	Yes	Yes
Gender	Female	Female	Male	Male
Observations	421	421	818	818
R-squared	0.252	0.278	0.013	0.014
Number of Bib IDs	134	134	258	258

Standard errors in parentheses. $^{***}p<0.01$, $^{**}p<0.05$, $^{*}p<0.1$.

Table 7
Time bias measured in sessions throughout the day

	Execution score			
Second session			–0.0316	–0.0232
			(0.457)	(0.459)
Third Session	0.0453	0.0406	–0.0050	0.0055
	(0.473)	(0.473)	(0.395)	(0.397)
Fourth Session	–0.8211*	–0.8313*	–0.0475	–0.0299
	(0.474)	(0.474)	(0.313)	(0.311)
Normalized Difficulty	0.1494*	0.2026**	0.0018	–0.0089
Score	(0.083)	(0.083)	(0.076)	(0.074)
Normalized	–0.1600***	–0.1466***	0.0382	0.0328
Difficulty Score2	(0.046)	(0.048)	(0.044)	(0.042)
Normalized Difficulty	0.4578**		–0.0848	
Score x Reputation	(0.180)		(0.143)	
Constant	0.2733***	0.3057***	0.0344	0.0225
	(0.076)	(0.080)	(0.234)	(0.235)
Event FE	Yes	Yes	Yes	Yes
Athlete FE	Yes	Yes	Yes	Yes
Gender	Female	Female	Male	Male
Observations	424	424	821	821
R-squared	0.269	0.255	0.006	0.006
Number of Bib IDs	134	134	259	259

Standard errors in parentheses; The excluded group is the first, of four, session of the day.
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

the impact of this form of bias – females face this bias at a significantly different rate than males.

I also find a significant impact on the interaction of difficulty scores on reputation. Again with big gender differences, for males there continues to be no evidence of a difficulty bias and no impact on the interaction term. However, for the female athletes, the interaction of difficulty score and reputation has a large and significant impact on the women's competition. This shows that the women from the best-trained countries also tend to get the biggest increase in their score for trying the most difficult routines.

4.3. A Robustness of overall order bias

As a robustness to the search of an overall order bias, another way to test this is by the round the athlete performs in rather than just the rank order they compete (i.e., *order* is a list of 1, 2, 3, . . . , *n*; where n is the final gymnast vs. *session* is the block in which the gymnast competes in the day, not testing the order within each session). Given this different measure of overall order bias, I find weak evidence that the last session of the day for females was actually worse than the other sections – which goes against the previous literature on this topic. Rotthoff (2015) found, measuring overall order in different sessions of the day, that going in any of the later three (not the first session of a four-session day) was best for your overall score. In Table 7, I continue to find strong evidence of a difficulty bias that is present only in the female events, with no evidence of this bias in male events (session two, for females, is dropped do to collinearities with the other controls in the regression).

5. Conclusion

The execution score is designed to be unrelated to the difficulty score to avoid improper scoring, but I continue to find evidence that those who perform a more difficult routine receive a higher execution score and, in reverse, that those who perform less difficult routines are given lower execution scores. These results support the findings in Morgan and Rotthoff (2014), who define this as a difficulty bias. However, I also expand this research and find clear and consistent evidence that males and females are impacted differently by this form of bias. Females have a large and significant impact by difficulty bias, whereas males seem to be unaffected by this form of bias.

It is also important to note that in gymnastics, because the individual gets to choose the difficulty of their routine, it presents situations where competitors can choose to optimize their choice of difficulty. Ideally, it would be best to evaluate this bias by assigning the difficulty level of routines at random to the competitors that do not coincide with their

optimal choice. This is not possible in this scenario but should be applied in other areas where the difficulty level is determined from the outside, not through self-selection.

Competitors that know they have a higher ability will also gain from being able to perform a more difficult routine because they have a lower expected cost, so they will choose a higher difficulty. A gymnast with a lower ability will have a higher expected cost than the more able gymnast. If the execution score were not impacted by the difficulty bias, the spread between gymnasts with differing ability and difficulty but equal execution would be less. With a present and evident difficulty bias, the competitor can control their difficulty score and artificially increase their execution score through self-selection.

Providing an even playing field for those being evaluated is essential to come to a true outcome. During an interview or debate, it is important to ask the same level of questions. If more difficult questions are asked to some candidates, they receive a bump when scoring their answers because they had a more difficult question. Also, when a candidate can pick their level of difficulty, such as in a musical audition or when an author decides the level of their statistical analysis in a paper, this finding will incentivize people to attempt a more difficult piece of music or include more difficult (and possibly unnecessary) statistical analyses. In an attempt to impress the judges and get an artificial bump in score they attempted something with a higher difficulty level.

However, there have been many studies that have found evidence of an overall order bias: Flores and Ginsburgh (1996), Gleiser and Hendels (2001), Bruine de Bruin (2005), Page and Page (2010), and MRR. These studies all suggest that it better to go later in any judged event. However, this study suggests that it is possible that elite level judges may be concerned with this form of bias and have responded to these findings by adjusting their behaviors (or it randomly went away during this competition). If this is true, and true for other judges as well, people no longer need to worry about the optimal ordering of their presentations or job interview time.

Acknowledgment

A special thanks to Stephanie Dunham who did work on the start of this project as an undergraduate student at Seton Hall University. Any mistakes are my own.

References

Auweele, Y., Boen, F., Geest, A. and Feys, J., 2004, Judging bias in synchronized swimming: Open feedback leads to nonperformance-based conformity, *J Sport Exerc Psychol*, *26*, 561-571.

Bruine de Bruin, W., 2005, Save the Last Dance for Me: Unwanted Serial Position Effects in Jury Evaluations, *Acta Psycologica*, *118*, 245-260.

Burgess, N. and Hitch, G., 1999, Memory for serial order: A network model of the phonological loop and its timing, *Psychological Review*, *106*, 551-581.

Campbell, B. and Galbraith, J., 1996, Nonparametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments, *The Statistician*, *45*(4), 521-526.

Damisch, L., Mussweiler, T. and Plessner, H., 2006, Olympic Medals as Fruits of Comparison? Assimilation and Contrast in Sequential Performance Judgments, *Journal of Experimental Psychology Applied*, *12*, 166.

Emerson, J., Seltzer, W. and Lin, D., 2009, Assessing Judging Bias: An Example from the 2000 Olympic Games, *The American Statistician*, *63*, 124-131.

Flôres, R.G. Jr. and Ginsburgh, V.A., 1996, The Queen Elisabeth Musical Competition: How Fair Is the Final Ranking? *The Statistician*, *45*(1), 97-104.

Findlay, L.C. and Ste-Marie, D.M., 2004, A Reputation Bias in Figure Skating Judging, *Journal of Sport and Exercise Psychology*, *26*, 154-166.

Garicano, L., Palacios-Huerta, I. and Prendergast, C., 2005, Favoritism Under Social Pressure, *Review of Economics and Statistics*, *87*, 208-216.

Gershberg, F. and Shimamura, A., 1994, Serial position effects in implicit and explicit tests of memory, *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 1370-1378.

Gift, P. and Rodenberg, R.M., 2014, Napoleon Complex Height Bias Among National Basketball Association Referees, *Journal of Sports Economics*, *15*(5), 541-558.

Glejser, H. and Heyndels, B., 2001, Efficiency and inefficiency in the ranking in competitions: The case of the Queen Elisabeth music contest, *Journal of Cultural Economics*, *25*, 109-129.

Goldin, C. and Rouse, C., 2000, Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians, *The American Economic Review*, *90*(4), 715-741.

Kahneman, D. and Tversky, A., 1996, On the reality of cognitive illusions, *Psychological Review*, *103*, 582-591.

Kingstrom, P.O. and Mainstone, L.E., 1985, An investigation of the rater-ratee acquaintance and rater bias, *Academy of Management Journal*, *28*, 641-653.

Morgan, H. and Rotthoff, K.W., 2014, The Harder the Task, the Higher the Score: Findings of a Difficulty Bias, *Economic Inquiry*, *52*(3), 1014-1026.

Mussweiler, T., 2003, Comparison Processes in Social Judgments: Mechanisms and Consequences, *Psychological Review*, *110*(3), 472-489.

Neilson, W., 1998, Reference Wealth Effects in Sequential Choice, *Journal of Risk and Uncertainty*, *17*, 27-48.

Novemsky, N. and Dhar, R., 2005, Goal Fulfillment and Goal Targets in Sequential Choice, *Journal of Consumer Research*, *32*, 396-404.

Page, L. and Page, K., 2010, Last shall be first: A field study of biases in sequential performance evaluation on the Idol series, *Journal of Economic Behavior & Organization*, *73*, 186-198.

Parsons, C.A., Sulaeman, J., Yates, M.C. and Hamermesh, D.S., 2011, Strike Three: Discrimination, Incentives, and Evaluation, *The American Economic Review*, *101*(4), 1410-1435.

Price, J. and Wolfers, J., 2010, Racial Discrimination Among NBA Referees, *Quarterly Journal of Economics*, *125*(4), 1859-1887.

Rotthoff, K.W. 2015, (Not Finding a) Sequential Order Bias in Elite Level Gymnastics, *Southern Economic Journal 81*(3), 724-741.

Sala, B., Scott, J. and Spriggs, J., 2007, The Cold War on Ice: Constructivism and the Politics of Olympic Skating Judging, *Perspectives on Politics*, *5*(1), 17-29.

Sarafidis, Y., 2007, What Have you Done for me Lately? Release of Information and Strategic Manipulation of Memories, *The Economic Journal*, *117*, 307-326.

Segrest Purkiss, S., Perrewe, P., Gillespie, T., Mayes, B. and Ferris, G., 2006, Implicit Sources of Bias in Employment Interview Judgments and Decisions, *Organizational Behavior and Human Decision Processes*, *101*, 152-167.

Seltzer, R. and Glass, W., 1991, International Politics and Judging in Olympic Skating Events: 1968-1988, *Journal of Sports Behavior*, *14*, 189-200.

Tversky, A. and Kahneman, D., 1974, Judgment and uncertainty: Heuristics and biases, *Science*, *185*, 1124-1131.

Wilson, V., 1977, Objectivity and Effect of Order of Appearance in Judging of Synchronized Swimming Meets, *Perceptual and Motor Skills*, *44*, 295-298.

Zitzewitz, E., 2006, Nationalism in Winter Sports Judging and its Lessons for Organizational Decision Making, *Journal of Economics and Management Strategy*, Spring, 67-99.

Zitzewitz, E., 2010, Does Transparency Really Increase Corruption? Evidence from the 'Reform' of Figure Skating Judging *Working Paper*.

Appendix

Table A1
Superstar countries for women's events

Superstar countries (women)			
Vault	Uneven bars	Balance beam	Floor
USA	USA	USA	USA
China	China	China	Romania
Germany	Russia	Romania	China
Russia	Great Britain		Russia

Table A2
Superstar countries for men's events

Superstar countries (men)					
Parallel bars	High bar	Rings	Floor	Vault	Pommel horse
China	China	China	Brazil	Romania	China
Japan	Germany	Bulgaria	China	Russia	Great Britain
S. Korea	Japan	Italy	Japan	Poland	Hungary
	Netherlands	Netherlands	Canada		Australia
			Romania		