Order Effects in Elite Gymnastics

Joustra, Sanne J.; Koning, Ruud H.; Krumer, Alex

Published in:
Economist-Netherlands

DOI:
10.1007/s10645-020-09371-0

Publication date:
2021

De Economist (2021) 169:21–35
https://doi.org/10.1007/s10645-020-09371-0

Order Effects in Elite Gymnastics

Sanne J. Joustra[1] · **Ruud H. Koning**[2] · **Alex Krumer**[3]

Published online: 11 August 2020

Abstract
We study the effect of order of performances on the final execution score in artistic gymnastics. We use data from all the finals for both genders at the main gymnastics tournaments in the period between 2009 and 2017 where the order of performances was fully randomised. We find that female gymnasts who perform later receive on average a significantly higher execution score for their routine, however, we find no effect of the order of performances in men's competitions. The existence of subjective evaluation only in female competitions considering artistry, could be one of the possible explanations for gender differences. We also find no relationship between the score of an athlete and his or her immediate predecessor.

Keywords Performance · Order effects · Gymnastics · Subjective performance evaluation

JEL Classification D70 · D00 · L10 · D20 · Z20

1 Introduction

Evaluation of sequential performance is prevalent in many areas of life that include job promotions (Rosen 1986); political campaigns (Klumpp and Polborn 2006); sports (Rotthoff 2015); music competitions (Ginsburgh and van Ours 2003); and even parole decisions of judges (Danziger et al. 2011). A fair order of actions would

Sanne. J. Joustra: BSc student Economics and Business Economics, University of Groningen (2014–2018). She participated in a trainings program for the Olympic Games London 2012 and participated in multiple international competitions, all-around elite artistic gymnastics.

✉ Alex Krumer
alex.krumer@himolde.no

[1] University of Groningen, Groningen, The Netherlands

[2] Department of Economics, Econometrics and Finance, Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands

[3] Faculty of Business Administration and Social Sciences, Molde University College, Britvegen 2, 6402 Molde, Norway

result in the absence of the order effect on the final outcome. However, a possible unfair advantage driven by the order of actions may harm efficiency by reducing the probability of picking the optimal decision. This can result in an inefficient economy that operates below the production-possibility frontier.

Previous research have shown that the probability of success depends on the order of actions in many settings that involve subjective evaluation. For example, Wilson (1977) showed a positive relationship between the serial number of performance and the final scores in synchronised swimming competitions. In the same spirit, Flôres Jr and Ginsburgh (1996), Glejser and Heyndels (2001) and Ginsburgh and van Ours (2003) found that piano finalists who performed later in the final week of the prestigious Queen Elisabeth music contests obtained a higher rank. De Bruin (2005) found similar results for the Eurovision song contest and for the World and European figure skating championships. In a follow-up study with a larger dataset, De Bruin (2006) confirmed her initial finding on advantage of performing later in figure skating competitions. Similar positive effect of performing later was found in the popular Idol series (Page and Page 2010) and in the international New Wave song contest for young performers (Antipov and Pokryshevskaya 2017). Finally, in a completely different setting, Danziger et al. (2011) showed that judges' decisions are more favorable after their food break.

In this paper, we build on efforts of Damisch et al. (2006), and Rotthoff (2015), who studied the effect of order in artistic gymnastics (henceforth, gymnastics). The results of these papers are mixed. For example, Damisch et al. (2006) found a positive relationship between the serial number of performance and the final score, claiming for the overall order bias. In addition, the authors showed a positive relationship between the score of an athlete and his or her immediate predecessor suggesting the existence of the sequential order effect. However, Rotthoff (2015) could only reproduce the result on the positive overall order effect (see also Morgan and Rotthoff (2014)), failing to find the sequential order effect. The one possible explanation on deviation between the results is a non-random allocation of athletes in the data of Damisch et al. (2006) as acknowledged by the authors themselves. In addition, in a recent study, Rotthoff (2020) finds no consistent evidence of the overall order effect, contradicting the results of the three previous papers.

We wish to retest the overall and sequential order effects in gymnastics by contributing to the literature in several ways. First, we use a significantly larger dataset compared to the previous papers. For example, Damisch et al. (2006) used only data from 2004 Athens Olympic Games with non-random allocation of order. The papers of Morgan and Rotthoff (2014) and Rotthoff (2015) also used only one event, namely the 2009 World Championship, whereas Rotthoff (2020) utilized data from the 2013 World Championship only. The authors of the three later studies justify the choice of these tournaments by omission of team competitions which may affect the decision making of coaches in allocating the athletes into different starting positions. In addition, in these tournaments the authors used only data from qualification (preliminary) round, where coaches could still strategically allocate athletes into

different positions[1]. Thus, omission of team events does not exclude the possibility of non-random allocation of athletes. In addition, the usage of only the qualification round may jeopardize the estimation strategy because some athletes do not wish to maximise their efforts for two reasons. The first is that some athletes know that they will qualify with a high probability even without maximising their effort. The second reason is that some athletes give up a certain apparatus since they only aim for achieving a good result in another apparatus.

In our paper, we use data from all the main gymnastics tournaments in the period between 2009 and 2017 that include nine European Championships, seven World Championships and two Olympic Games. Moreover, we only use data from finals where athletes have the incentives to perform their best. More importantly, the allocation of athletes in finals is fully randomised, which substantially simplifies credible causal inference (Manski 1995).

In addition, given contradicting findings on overall order effect in Rotthoff (2015, 2020), and given that Rotthoff (2020) finds some inconsistent evidence on gender differences in the overall order effect, we investigate the order related effects for each gender separately. The additional reason for that is that there is a notable difference in the judging system for male and female gymnasts considering artistry, which may interact with order effect (for example via sharp differences between two consecutive athletes). In addition, previous studies have found gender differences in performance in competitive settings in general (Niederle and Vesterlund 2007) and in sequential tournaments in particular (Cohen-Zada et al. 2017). Thus, it is important to separate between the genders in the analyses on the effect of the order of actions on performance.

Finally, we also contribute to the growing debate about the importance of replication studies. For example, Open Science Collaboration (2015) could replicate the results of only 36 out of 100 experimental and correlational studies that were published in top academic journals in psychology. In addition, Ioannidis and Doucouliagos (2013) discuss the empirical evidence on the lack of a robust reproducibility culture in economics and business research. Therefore, replication of original findings is an important scientific task.

Based on evaluation of 862 male and 572 female routines, we are able to reproduce only some of the previous findings. First, in line with the results of Rotthoff (2015), we find no sequential order effect for both genders, suggesting that there is no relationship between the execution scores of the two consecutive gymnasts. Second, the results of the overall order effects are mixed. We find no order effect on the execution score among men. However, female gymnasts who perform later in the final receive on average a significantly higher execution score for their routine. More specifically, the overall order effect is estimated between 0.01 to 0.03 execution points on average per additional later performance. To put that result into perspectives, the Romanian gymnast Catalina Ponor finished fourth in the final of the balance beam competition at London 2012 Olympic Games having the same number

[1] Note that there is no explicit mention in Rotthoff (2020) about the exact round of competition it uses except for a mention that it follows Rotthoff (2015) that only used data on qualification round.

of points as the bronze medallist Aly Raisman from the USA. The Romanian athlete lost the tie-breaker, which prioritized execution score over difficulty score. Interestingly, Ponor performed second in the final, whereas Raisman's routine was the last in the competition[2].

It is important to note that our data do not allow us to identify the exact mechanism of the overall order effect in general and gender differences in particular. The possible reasons could be that female athletes enhance their performance when they know which result they have to achieve, whereas male athletes are not able to take advantage of the additional information they have and choke under pressure (Cohen-Zada et al. 2017). Thus, it is possible that information advantage and choking under pressure cancel each other. On the other hand, it is possible that referees are those who are responsible for that effect. The reason is that in women's artistic gymnastics the judges have more room for subjective evaluation then in men's artistic gymnastics due to the scoring of artistry. For example, female gymnasts perform their floor routine on music, whereas the male gymnasts do not. As a consequence, a female gymnast can lose points for the lack of synchronization between movement and music.

The remainder of the paper is organised as follows: Sect. 2 describes the institutional settings in gymnastics. The data and descriptive results are presented in Sect. 3. The results are contained in Sect. 4. Finally, in Sect. 5 we offer concluding remarks.

2 Institutional Setting

In women's artistic gymnastics, the athletes perform on four apparatus: vault, bars, beam, and floor. On the other hand, men perform on six apparatus: floor, pommel horse, rings, vault, parallel bars, and horizontal bar. The gymnastics competitions we consider are divided into two main rounds. First, there is one qualification round with all gymnasts, and second, there are different final rounds with only the gymnasts who qualified. Usually, the tournament consists of a (country) team competition, an all-around competition, and individual competitions on each apparatus. In this paper, we only consider the individual apparatus finals as a consequence of the randomisation of the order of performance in this final.

In the individual apparatus finals, the best eight gymnasts on that apparatus of the qualification are performing. In the rare event that two gymnasts qualify as eighth, a ninth gymnast is added to the final. The maximum number of gymnasts in the final from the same country is limited to two. So, for example, the third best United States gymnast who qualified as fourth in the qualification women's floor in the Olympic Games 2016 (Laurie Hernandez), was not eligible to perform in the final since two other gymnasts from the United States (Simone Biles and Aly Raisman) had qualified as first and second respectively. In this case, the ninth qualifier also happend to represent the United States, so the tenth qualifier was eligible to enter the final.

[2] For the full competition, see youtube.com/watch?v=VZvoufQy8qc. Last accessed on 12/06/2020.

Performance order in the final on each individual apparatus is randomised, unlike other sports where order of performance is in reverse order of qualification. For example, Regulation 5.1.8.3 reads 'The working order on each apparatus is determined by the drawing of lots' (FIG 2015), section 2, p. 11). Order of performance is *not* randomised in the team and all-around competitions, and for that reason, we do not include those in our study.

The scoring system in artistic gymnastics changed after the 2004 Olympic Games in Athens. Since then, the score consists of two parts: a difficulty score (D-score) and an execution score (E-score). The D-score is determined by the elements a gymnast performs and the E-score is determined on how well a gymnast performs those elements. For both of these scores there is a different jury panel. Usually there are two judges to determine the D-score and five judges to determine the E-score. The D-score is open-ended, because a gymnast receives credits for all elements he or she performs. The D-score is a consequence of the performance chosen by the gymnast: different elements and combinations in the performance are awarded points according to the 'Code of Points'. The D-score can be contested by the gymnast. The judges for the E-score deduct for errors starting at a highest score of 10.00. The highest and the lowest execution scores are dropped and the average of the remaining scores is the final E-score. Moreover, there is a reference panel with two judges that check the E-score given by the judges. The final score of a gymnast is the sum of the D- and E-scores, with a deduction for any penalties. These penalties are given for (multiple) violations of the rules, for example, if the time limit of a performance is exceeded. A new code of points is published after the Olympic Games, so the valuation of different elements in performances changes over time.

Finally, the gymnast with the highest total score wins. If two gymnasts are tied, the E-score is used as a tie breaker.

In women's artistic gymnastics, the judges are almost always females. At the time of writing (July 2020), 705 judges are registered at the Fédération Internationale de Gymnastique (www.gymnastics.sport/site/judges/jud_view.php) licensed to judge women's artistic gymnastics, of which 18 are male. For men's artistic gymnastics, the ratio is similar. There are 852 judges registered of which only 13 are female. So, in artistic gymnastics, the gymnasts are generally judged by the same gender.

3 Data

3.1 Data Collection

We collected data on the apparatus finals of the main elite gymnastics tournaments in the period 2009–2017: nine European Championships (organised every year), seven World Championships (organised every year, except when there are Olympic Games) and two Olympic Games (2012, 2016). For each event, we know the finalists, the order of performance in the final, D- and E-scores in the final, the D- and E-scores in the qualification, as well as the country of origin of the gymnast. Order of performance has been validated by examining television footage, when necessary. We removed evidently failed performances, as these

Table 1 Descriptive statistics E-score

	Period I			Period II			Period III		
	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.
Men									
Horizontal bar	8.23	0.60	64	7.99	0.67	64	7.64	0.81	16
Men's floor	8.75	0.37	64	8.48	0.48	64	8.09	0.36	17
Men's vault	9.13	0.29	63	9.01	0.41	63	9.03	0.37	15
Parallel bars	8.83	0.34	65	8.54	0.55	65	8.57	0.73	15
Pommel horse	8.47	0.60	64	8.43	0.56	64	8.26	0.70	16
Rings	8.61	0.39	64	8.73	0.26	64	8.76	0.28	15
Women									
Beam	8.16	0.71	64	8.15	0.62	64	7.59	0.72	17
Uneven bars	8.34	0.59	64	8.35	0.61	64	8.20	0.59	16
Women's floor	8.57	0.43	63	8.39	0.40	64	8.33	0.24	15
Women's vault	8.70	0.33	62	8.87	0.43	63	8.87	0.26	16

performances are not informative about any order effect. For example, a gymnast may not have been able to complete the performance after a fall and subsequent injury. As a consequence, the number of observations per tournament varies between 78 and 81. As discussed above, the observations we have are subject to three different Code of Points: 2009–2012, 2013–2016, and 2017–2019. We refer to these periods as 'Period I', 'Period II', and 'Period III' respectively.

The variable of interest in this paper is the E-score, and not the D-score. The reason for this is the following. Gymnasts have to execute a performance in limited time. They practice the elements comprising the performance with great investment of effort and time, so the performance (that is, the D-score) is planned in great detail at the beginning of the final. It is possible that some elements are not carried out completely (for example, a planned triple rotation may become a double rotation because of an imbalance), and as a consequence the D-score of the performance may deviate a little from the planned D-score. Even though the D-score is judged, there is no room for any order bias, as the gymnast can appeal to the D-score as evaluated by the jury. The E-score is more subjective, and measures the quality of the performance, and cannot be appealed to. We take the E-score as the dependent variable in our analyses. A summary of the E-scores in the dataset is provided in Table 1.

We see in Table 1 variation within apparatus over periods with different Code of Points, and between different apparatus within one period with the same Code of Points, both as far as mean score and standard deviation of the score is concerned. In particular, the standard deviation in horizontal bar and pommel horse (men) and beam and uneven bars (women) tend to be larger, since one could lose balance and get a relatively large penalty deduction. Average E-score on vault is highest, both for men and women, while the lowest average scores are for horizontal bar (men) and beam (women). To allow for heterogeneity over time and between apparatus and

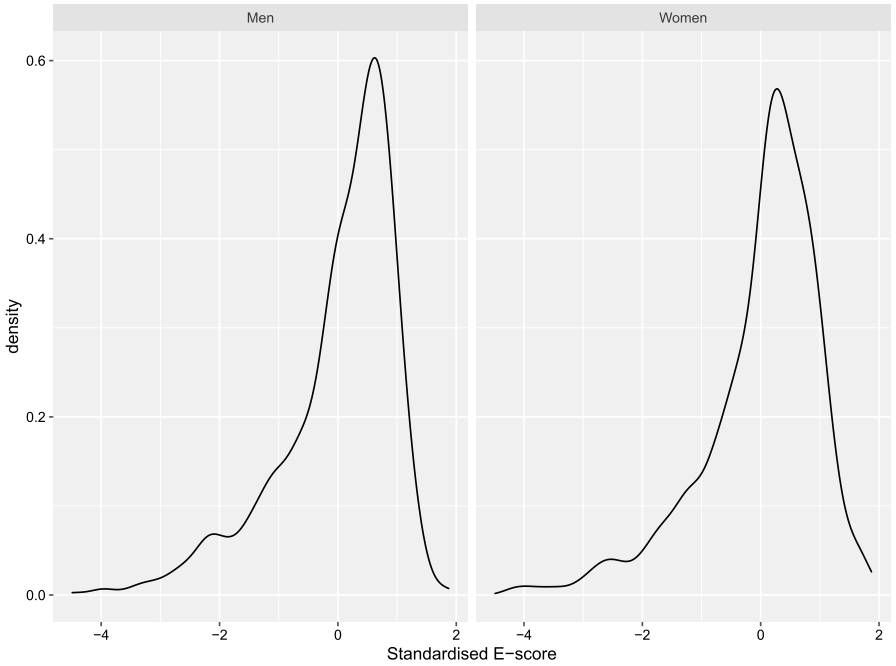

Fig. 1 Kernel density estimate of distribution of standardised E-scores, by gender

gender, we standardise the E-score within each apparatus-gender-period combination. This is similar to Rotthoff (2015, 2020).

In Fig. 1 we provide a kernel estimate of the distribution of standardised E-scores, by gender. The distributions are similar: unimodal with a bit of a tail to the left. This tail reflects performances with serious errors (such as falling, stepping from the beam, etc.). Similar distributions are obtained if made for each gender-apparatus combination.

3.2 Randomisation

An important contribution of this paper is to test the existence of an order effect in a truly randomised setting. Even though Rotthoff (2015), when analysing the World Championships 2009, that is also part of our dataset, writes 'The final round is done in traditional gymnastics meet fashion where top talent performs last' (p. 727), this is not correct. In that tournament, the rank correlation between the rank in the qualification and order in the final varies between −0.15 (beam) and 0.36 (uneven bars), as measured by Kendall's τ. If order of appearance were in reverse of rank of qualification, this statistic would be −1. The 0.025 and 0.975 quantiles of the distribution of Kendall's rank correlation τ of two independent permutations of ranks 1 through 8 are −0.57 and 0.57 respectively. In other words, ranking in the qualification and order in the final were not correlated in that tournament. In our full dataset,

Table 2 Regression results, dependent variable: standardised E-score

	All	Men	Women
	(1)	(2)	(3)
Order	0.021*	0.006	0.044**
	(0.011)	(0.015)	(0.018)
Constant	− 0.096*	− 0.027	− 0.199**
	(0.058)	(0.074)	(0.091)
Observations	1434	862	572
R^2	0.002	0.0002	0.011

Standard errors in parentheses

*$p < 0.1$; **$p < 0.05$

Kendall's τ calculated between ranking in the qualification and order in the final varies between −0.87 (women's vault, Olympic Games 2012) and 0.64 (women's uneven bars, World Championship 2010). 95.5% of the 180 estimated correlation coefficients are between the two quantiles listed above. The average correlation over all events is 0.00. We conclude that the order of performance in the final is unrelated to the quality of the gymnasts as measured by the ranking of these gymnasts in the qualification. Order is randomised.

4 Analysis and Results

To test whether or not an order effect exists in the judgment of the jury, we proceed as follows. As noted in Table 1, there is noticeable heterogeneity in the E-scores both between apparatus and periods with different Code of Points. We follow Rotthoff (2015, 2020) and standardise the observations within each apparatus, gender, and period combination. Note that Rotthoff (2015, 2020) does not have to standardise with respect to periods with fixed Code of Points, as he uses data of the qualification of a single tournament only. If we were to standardise within tournaments, we would lose two degrees of freedom on eight observations in each apparatus final. In order to maintain the power of our test, we chose to not standardise within each tournament separately, but within periods with fixed Code of Points. We ignore any dependence between observations as a consequence of this standardisation. As a consequence of the standardisation, effect sizes of the regressions below are at the scale of the standard deviation of the E-score within each apparatus, gender, and period combination.

To test for the existence of an order effect we regress the standardised E-score on the order of performance. The results are given in Table 2, where we present three sets of regression results, one for all data, one for men, and one for women. As argued above, the order in the final is truly randomised and hence strictly exogenous, in particular with respect to ability. Under the null hypothesis of no order effect, the slope of the variable order should be 0.

Table 3 Regression results, dependent variable: standardised D-score

	All (1)	Men (2)	Women (3)
Order	0.008	0.011	0.004
	(0.011)	(0.015)	(0.018)
Constant	− 0.037	− 0.049	− 0.019
	(0.058)	(0.074)	(0.091)
Observations	1434	862	572
R^2	0.0004	0.001	0.0001

Standard errors in parentheses

*$p < 0.1$; **$p < 0.05$

Table 4 Regression results, dependent variable: standardised E-score

	All (1)	Men (2)	Women (3)
Order	0.040**	0.026	0.063**
	(0.018)	(0.023)	(0.028)
HA	− 0.061	− 0.039	− 0.101
	(0.098)	(0.127)	(0.154)
First	0.124	0.160	0.071
	(0.101)	(0.130)	(0.159)
Last	− 0.104	− 0.079	− 0.144
	(0.101)	(0.130)	(0.159)
Constant	− 0.180**	− 0.123	− 0.264*
	(0.085)	(0.110)	(0.135)
Observations	1434	862	572
R^2	0.004	0.002	0.013

Standard errors in parentheses

*$p < 0.1$; **$p < 0.05$

In Table 2 we see that in the population of all gymnasts, there is a small positive order effect, that is not statistically significant at the usual 5% level. However, once we stratify with respect to gender, we find that there is no evidence of an order effect in men competitions ($p = 0.688$), but there is a clear order effect in women competitions ($p = 0.014$). Female gymnasts who perform later in the final get, on average, a higher E-score. The effect size if also meaningful. When multiplied by the standard deviation within each rules period, it varies between 0.01 points per one later performance (floor, period III) and 0.03 points per later performance (beam, period III). In period III, the average difference in total score (D-score and E-score added together) on floor between the gold medallist and bronze medallist, is 0.35 points. So approximately10% of this difference is made up by performing three or four slots later in the final.

Table 5 Regression results, dependent variable: standardised E-score

	All (1)	Men (2)	Women (3)
Order	0.042**	0.026	0.066**
	(0.018)	(0.023)	(0.028)
Previous E-score	− 0.013	− 0.011	− 0.021
	(0.028)	(0.036)	(0.044)
HA	− 0.030	0.013	− 0.091
	(0.106)	(0.140)	(0.162)
Last	− 0.108	− 0.081	− 0.150
	(0.101)	(0.131)	(0.158)
Constant	− 0.189**	− 0.128	− 0.283**
	(0.085)	(0.111)	(0.134)
Observations	1254	754	500
R^2	0.005	0.002	0.013

Standard errors in parentheses

*$p < 0.1$; **$p < 0.05$

The calculation of the standard errors in Tables 2, 3, 4 and 5 is based on the assumption of independent performances. We also calculated standard errors clustered at the level of the individual gymnast (see Tables 6, 7, 8, 9 in Appendix). These are only marginally different from the ones reported in Tables 2, 3, 4 and 5, and do not result in any other conclusions relating to the significance of the results.

One possible concern though, is that gymnasts update their difficulty score as a function of their order, which may affect their execution score (Morgan and Rotthoff 2014). To obviate this concern, we also regressed the standardised D-score on order in the final. As expected, from the results given in Table 3 it is evident that there is no order effect in the D-score, neither for all gymnasts, nor when we separate by gender.

There are different explanations for the difference of the order effect between genders. One explanation is that males and females are judged according to different criteria. Both men and women get points deducted from their E-score if they perform an element incorrectly, for example, if they pause 2 seconds or longer before an element or acrobatic series in a floor exercise. Possible faults and consecutive deductions are listed explicitly. Besides these, artistic criteria are stressed in the female Code of Points. As a result of a lack of artistry, a gymnast can lose up to one full point on beam and up to 2 full points on floor (out of 10 for the E-score). For example, on beam a gymnast can lose point for 'lack of variation in rhythm and tempo in movements' or 'personal style' and confidence. On floor, a female gymnast performs the routine on music, whereas the male gymnasts do not. As a consequence, a female gymnast can lose points for 'music and musicality', 'lack of synchronisation between movement and music' and for 'editing of music'. On both beam and floor, a gymnast can also lose points for composition of the routine. For example, she can lose points for 'insufficient

complexity or creativity of movements'. Therefore, the evaluation criteria for women are more subjective than for men. In this case, the order effect in Table 2 reflects bias on the part of the jury.

To assess the robustness of our result about the overall order effect on the execution score, we include dummies for the first and last performance in the final, and a dummy for home advantage (HA). Home advantage is well known to be prevalent in both team sports and individual sports, in particular when performance is judged by a subjective standard (Balmer et al. 2003). The results of this specification are given in Table 4. The first and last gymnasts are not evaluated differently. Moreover, there is no evidence of home advantage. The effect sizes of the order effect have increased a bit, but again, there is no statistically significant order effect for men, and there is a significant order effect for women.

In a final specification, we also test for the sequential order effect: is the judgment of a gymnast related to the performance of the gymnast immediately preceding. This sequential order bias is tested for by including the previous gymnast's standardised E-score in the regression. As a consequence, we lose the first observation in each final, and we can no longer include the dummy variable indicating the first finalist. The results of this specification are given in Table 5. There is no evidence of a sequential order bias. The conclusions of the lack of order bias for men, and the existence of a significant positive order bias for women remain unchanged, also in this specification. The finding of an order bias in women gymnastics is robust.

5 Conclusions

In this paper we have tested for the existence of an overall order effect as well as for a sequential order effect for both men and women in the most prestigious tournaments in gymnastics where the order of performances was fully randomised. By using data of a higher quality in terms of randomisation and coverage compared to the previous papers, we can replicate only partly the results in Damisch et al. (2006), and Rotthoff (2015). More specifically, we find a significant overall order effect, but only among women, according to which performing later results in a significantly higher execution score. However, we fail to establish the overall order effect among men. In addition, we find no relationship between the score of an athlete and his or her immediate predecessor rejecting the existence of the sequential order effect.

Finally, given the high quality data available in gymnastics, we encourage future studies to investigate human's behaviour in sequential contests by using that data. In addition, future research may be in place to investigate a possible mechanism of the gender differences we found. One possible direction is to investigate the additional degree of freedom the referees have when evaluating a very subjective artistry feature that exists only in women's gymnastics competitions.

Acknowledgements We thank a referee for helpful comments.

Appendix

A density plot for each gender-apparatus combination is given in Fig. 2. The shapes are similar in each cell: a bit skewed to the left.

In Tables 6, 7, 8 and 9 we provide the same estimation results as in the main text, but with clustered standard errors at the level of the individual gymnast. Clustering is appropriate if individual athletes have a reputation that exceeds other athletes, and the E-grade also incorporates the subjective appreciation of the jury members

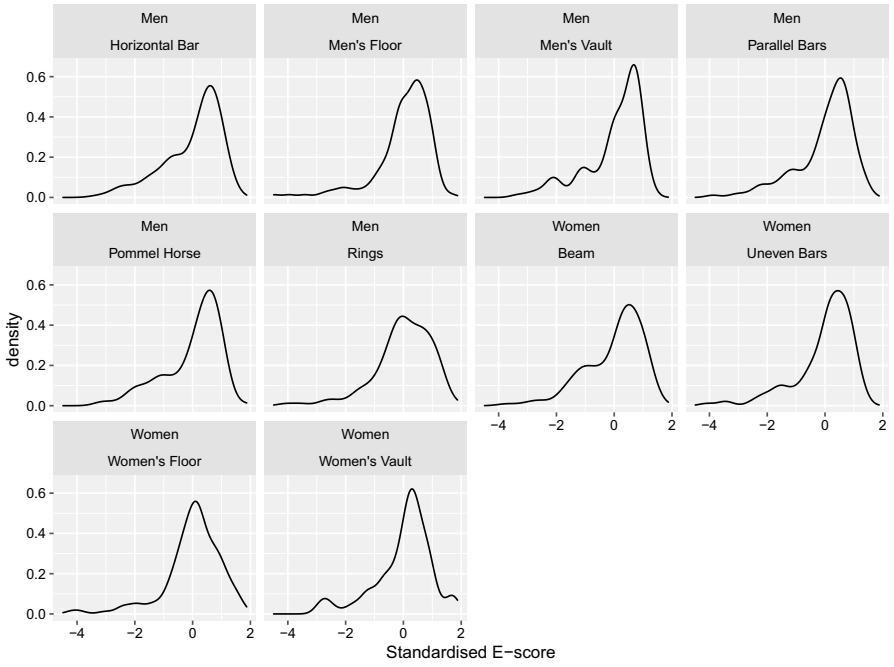

Fig. 2 Density plot of the standardised E-score by gender-apparatus combination

Table 6 Regression results, dependent variable: standardised E-score. Clustered standard errors

	All	Men	Women
	(1)	(2)	(3)
Order	0.021*	0.006	0.044**
	(0.012)	(0.015)	(0.018)
Constant	− 0.096	− 0.027	− 0.199**
	(0.066)	(0.087)	(0.101)
Observations	1434	862	572
R^2	0.002	0.0002	0.011

Clustered standard errors in parentheses

*$p < 0.1$; **$p < 0.05$

Table 7 Regression results, dependent variable: standardised D-score. Clustered standard errors

	All	Men	Women
	(1)	(2)	(3)
Order	0.008	0.011	0.004
	(0.011)	(0.015)	(0.017)
Constant	− 0.037	− 0.049	− 0.019
	(0.066)	(0.086)	(0.103)
Observations	1434	862	572
R^2	0.0004	0.001	0.0001

Clustered standard errors in parentheses

*$p < 0.1$; **$p < 0.05$

Table 8 Regression results, dependent variable: standardised E-score. Clustered standard errors

	All	Men	Women
	(1)	(2)	(3)
Order	0.040**	0.026	0.063**
	(0.018)	(0.023)	(0.028)
HA	− 0.061	− 0.039	− 0.101
	(0.096)	(0.097)	(0.192)
First	0.124	0.160	0.071
	(0.102)	(0.132)	(0.161)
Last	− 0.104	− 0.079	− 0.144
	(0.092)	(0.121)	(0.142)
Constant	− 0.180*	− 0.123	− 0.264*
	(0.092)	(0.124)	(0.138)
Observations	1434	862	572
R^2	0.004	0.002	0.013

Clustered standard errors in parentheses

*$p < 0.1$; **$p < 0.05$

Table 9 Regression results, dependent variable: standardised E-score. Clustered standard errors

	All (1)	Men (2)	Women (3)
Order	0.042**	0.026	0.066**
	(0.018)	(0.023)	(0.027)
Previous E-score	− 0.013	− 0.011	− 0.021
	(0.026)	(0.038)	(0.033)
HA	− 0.030	0.013	− 0.091
	(0.108)	(0.116)	(0.205)
Last	− 0.108	− 0.081	− 0.150
	(0.092)	(0.121)	(0.142)
Constant	− 0.189**	− 0.128	− 0.283**
	(0.092)	(0.124)	(0.137)
Observations	1254	754	500
R^2	0.005	0.002	0.013

Clustered standard errors in parentheses

**$p < 0.05$

of such athletes. However, even though the clustered standard errors differ slightly from the ones reported in the main text, the conclusion does not change. The estimation results as presented in the main text are robust in this sense.

References

Antipov, E. A., & Pokryshevskaya, E. B. (2017). Order effects in the results of song contests: Evidence from the eurovision and the new wave. *Judgment & Decision Making*, *12*(4), 415–419.

Balmer, N. J., Nevill, A. M., & Williams, A. M. (2003). Modelling home advantage in the Summer Olympic Games. *Journal of Sports Sciences*, *21*(6), 469–478.

Cohen-Zada, D., Krumer, A., Rosenboim, M., & Shapir, O. M. (2017). Choking under pressure and gender: Evidence from professional tennis. *Journal of Economic Psychology*, *61*, 176–190.

Cohen-Zada, D., Krumer, A., & Shtudiner, Z. (2017). Psychological momentum and gender. *Journal of Economic Behavior & Organization*, *135*, 66–81.

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied*, *12*(3), 166–178.

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, *108*(17), 6889–6892.

De Bruin, W. B. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, *118*(3), 245–260.

De Bruin, W. B. (2006). Save the last dance for me II: Unwanted serial position effects in figure skating judgments. *Acta Psychologica*, *123*(3), 299–311.

FIG. (2015). Technical regulations. In *Technical report*. Lausanne: Fédération Internationale de Gymnastique.

Flôres, R. G, Jr., & Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: How fair is the final ranking? *Journal of the Royal Statistical Society: Series D (The Statistician)*, *45*(1), 97–104.

Ginsburgh, V. A., & van Ours, J. C. (2003). Expert opinion and compensation: Evidence for a musical competition. *American Economic Review*, *93*(1), 289–296.

Glejser, H., & Heyndels, B. (2001). Efficiency and inefficiency in the ranking in competitions: The case of the Queen Elisabeth Music Contest. *Journal of Cultural Economics*, *25*(2), 109–129.
Ioannidis, J., & Doucouliagos, C. (2013). What's to know about the credibility of empirical economics? *Journal of Economic Surveys*, *27*(5), 997–1004.
Klumpp, T., & Polborn, M. K. (2006). Primaries and the New Hampshire effect. *Journal of Public Economics*, *90*(6), 1073–1114.
Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.
Morgan, H. N., & Rotthoff, K. W. (2014). The harder the task, the higher the score: Findings of a difficulty bias. *Economic Inquiry*, *52*(3), 1014–1026.
Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, *122*(3), 1067–1101.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(5251), aac4716.
Page, L., & 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.
Rosen, S. (1986). Prizes and incentives in elimination tournaments. *The American Economic Review*, *76*(4), 701–715.
Rotthoff, K. W. (2015). (Not finding a) sequential order bias in elite level gymnastics. *Southern Economic Journal*, *81*(3), 724–741.
Rotthoff, K. W. (2020). Revisiting difficulty bias, and other forms of bias, in elite level gymnastics. *Journal of Sports Analytics*, *6*(1), 1–11.
Wilson, V. E. (1977). Objectivity and effect of order of appearance in judging of synchronized swimming meets. *Perceptual and Motor Skills*, *44*(1), 295–298.