ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Economics and Business Economics

Major Behavioural and Health Economics

Imperfections in a sport striving for perfection?

Order bias and its possible mechanisms in women's artistic gymnastics

Name student: Eline Helder

Student ID number: 577099

Supervisor: Elisa de Weerd

Second assessor: David Gonzalez Jimenez

Date final version: 28-06-2023

Abstract

Many human interactions happen sequentially where people, consciously and unconsciously, impact each other's performances. Examples are job interviews, school examinations or sports competitions. This thesis investigates the existence of order bias in international elite women's artistic gymnastics competitions by reviewing the results found in previous research and investigating two new mechanisms to order bias. With the use of randomisation of the start position in apparatus finals, multiple regression analyses are performed to test the existence of overall and sequential order bias. No evidence for overall and sequential order bias on the difference in qualification and final performance was found. However, this does not mean that order bias is not present in the sport, as there could be mechanisms that work in opposite directions and cancel each other out. The quality of the first performance does not significantly affect later performances. The time gap between the warmup and competition, researched through the existence of a one-touch warmup, has also not been found to have a significant impact on order bias. However, when only looking at the subsample where a one-touch warmup was allowed, evidence for overall order bias was found. Gymnasts with start position 2, 3 and 7 perform significantly better than gymnasts with start position 1. More extensive research on overall order bias and its mechanisms, using experiments, is crucial to bring back fairness to gymnastics and other sequential events.

Acknowledgements

This thesis marks the end of an eventful and insightful bachelor's degree at Erasmus School of Economics. I would like to deeply thank my supervisor Elisa de Weerd for her trust in my thesis topic and her helpful comments and discussions along the way. Additionally, I could not have written this thesis without the continuous help and constructive critique from my boyfriend Tom van den Broeke. Next, I would like to thank my mother and sister for their support over the past three years. Lastly, as this thesis is about gymnastics, I want to express my gratitude to my gymnastics coaches, Essie and Ton. You have shown me what beautiful things I am capable of, both within and outside of the gym, and have taught me much about perseverance and helping me grow into the person I am today.

Contents

1. Introduction

Performance evaluations do not merely show the objective quality of performances. Sequential performance evaluations might be influenced by a number of factors other than the objective quality of the performance, resulting in lower economic efficiency. Order bias is described as the order of events impacting human decision making. Decision makers are anything but immune to the phenomenon of order bias, as has been revealed in research. Judicial decisions have been shown to be subject to the order in which evidence is shown (Pennington & Hastie, 1988) and the time of day during which the verdict is made (Danziger et al., 2011). Similarly, job interviews have been shown to be influenced by the ordering of events. Job acceptance decisions are highly influenced by the characteristics of the previous applicant (Rowe, 1967). If the previous applicant was perceived to be unfriendly, the acceptance decision is more favourable for the next applicant, independent of their own characteristics (Holmes & Berkowitz, 1961). Order effects have also been found in teachers grading written exams. Exams graded later on in the sequence and after streaks of extreme events were more prone to order bias (Goldbach et al., 2022).

Not only decision makers, but also competitors are also subject to order bias, often studied in sports. Drivers in NASCAR races show evidence of sequential behaviour by following the decisions to pit from the car ahead of them (Deck et al., 2014). Furthermore, Hill (2014) argues that strategic interactions and peer effects occur during qualification races in track and field tournaments. Qualifications includes multiple sequential heats in which competitors try to qualify for the finals, resulting in the later heats having more information on what times to run. Sequential start positions have previously been shown to affect the outcomes of competitions in many different sports, such as figure skating, synchronised swimming and artistic gymnastics, favouring competitors competing in later start positions (De Bruin, 2005; De Bruin, 2006; Wilson, 1977; Joustra et al., 2021).

Unlike many of the previously discussed settings, such as examinations and interviews, sports events often have well documented data and randomised competition orders. This allows for the separation of order bias from other effects that determine performance. Apparatus finals in women's artistic gymnastics are an example of a sequential setting in which the competition order is randomly determined, leading to the following research question.

What is the influence of order bias on the results of apparatus finals in women's artistic gymnastics?

Women's artistic gymnastics is an Olympic sport where competitors perform routines consisting of multiple elements to earn a score as high as possible. In each of the four apparatus finals, the gymnasts perform in a sequential setting with a randomised start order, where their routines are scored by a panel of judges. Although everyone has an equal chance of being drawn the first spot, the

last spot or anywhere in between, the start position might influence the outcome of the competition. Not only the start position, but also the performance of the previous competitor might be of importance. Due to the randomised setting combined with the sequential nature of the sport, gymnastics is the ultimate setting to compare to other life events, such as job interviews. The interviewer ("judge") sequentially talks to job candidates ("gymnast") and afterwards decides who they think is the best suitable for the job. Other examples are product inventors ("gymnasts") presenting their innovations to investors ("judges") or students ("gymnasts") taking an oral exam graded by their teacher ("judge"). However, the effect of order on performance and judgement in these type of settings is hard to measure due to a lack of data and the existence of other possible confounders. The randomised start order in gymnastics apparatus finals allows for the separation between order bias and other effects. In the world of gymnastics, order bias is possibly influenced by what the previous competitors have done, the atmosphere in the arena or the nerves of the gymnasts. The same holds for the judges, they might be influenced fatigue or reference points and their strictness can differ within parts of a final.

Existing literature often makes use of experimental data or data from a single competition and are therefore to a limited extent externally valid. Joustra et al. (2021) uses a more extensive dataset including around 1400 recent observations. However, no mechanisms explaining why order bias exists have been investigated so far, making it unclear which effects extend to other context outside of artistic gymnastics. This thesis will therefore add to the existing literature by using an even more extensive and recent dataset, while also investigating two possible mechanisms that might influence order effects in women's artistic gymnastics. These mechanisms run through the performance of previous competitors and allowing warmups on the competition apparatus. The quality of the first performance may be used as a reference point for the rest of the competition. This type of reference performance is not only the case in sports, but extends to other contexts as well. The time between the warmup and competition routine captures two things. Firstly, it shows the different lengths of time between a preparation and performance, which might be of influence on the performance. Secondly, the split in the competition induced by the one-touch warmup allows for a break for the judges. This might reduce fatigue and enhance sharpness. Additionally, this thesis sets itself apart from the existing literature by measuring the quality of performance as the difference between the performance in the apparatus finals versus the qualification round. These differences are taken with respect to the execution score (E-score) and total score. Measuring the quality of performance by taking the differences between the finals and qualifications has the advantage that it adjusts for individual differences between gymnasts and what is an attainable score for them. It also adjusts for differences in level and strictness in scoring across competitions and apparatus.

The findings of this thesis will inform both competitors and decision makers on order bias and make them more cautious of this. Fairness is an important aspect in sports as well as other decisions. Each competitor should be evaluated independently of their start position or the performances of previous competitors. Mechanisms are a key starting point to be able to target the existence of order bias and reduce the impact it has on human decision making in every context.

2. Literature review

Because Women's Artistic Gymnastics is a subjective sport, biases or other mistakes are prevalent in the judgement and performance of the routines. This thesis will investigate the presence of order bias in international elite women's artistic gymnastics. Order effects have mainly been researched in different settings such as sports and music contests due to the presence of recorded data. De Bruin (2005) has found evidence that final results are more favourable for contestants performing later on in the Eurovision Song Contest. Similar evidence has been found by Antipov and Pokryshevskaya (2017) for the New Wave Song Contest and by Page and Page (2010) in the Idol series. Page and Page (2010) argued that this effect runs through two mechanisms, the ability to remember all competitors and the direct comparison to the previous competitor. The first mechanism is only applicable in contests when end-of-sequence judgements are given, whereas this should not be an issue in step-by-step judgements like gymnastics scoring. Research within the field of sports shows similar findings. In figure skating, synchronised swimming and artistic gymnastics, existing literature suggest a systematic higher score for athletes later in the competition line up (De Bruin, 2005; De Bruin, 2006; Wilson, 1977; Joustra et al., 2021). Making use of the different heats in track and field qualifications, Hill (2014) argues that order effects could be explained by an asymmetry in information between different start times. Those competing in a later heat know what performances to beat in order to get the desired results, which is not the case for the earlier heats. Outside the world of music and sports contests, order bias has been shown to exist in the grading of examinations (Goldbach et al., 2022), judicial decisions (Danziger et al., 2011) and editorial decisions (Orazbayev, 2017). All three of these studies describe the phenomenon of fatigue when reviewing materials later on in a sequence, resulting in the decision maker being less strict.

The current literature on the existence of overall order bias in women's artistic gymnastics is limited of scope. Experiments where videos of team competitions were edited to manipulate the within-team order have shown significant results in favour of the later gymnasts (Ansorge et al., 1978; Scheer & Ansorge, 1975). Data from the 2009 World Championships also confirm the existence of overall order bias (Rotthoff, 2015). Gymnasts with a later start position score significantly higher on their E-score than gymnasts with an earlier start position. The magnitude of the effect is rather small, less than a hundredth point per position, however, gymnastics competitions have been decided on margins of a few hundred points[1]. Using a dataset including competitions from 2009 to 2017, Joustra et al. (2021) also found that competing later on in a randomised competition order has a positive effect on the gymnasts E-score. The effect size of start order on a standardised E-score differs between 0.04

[1] Judges at the E-panel cannot give deductions of a hundredth point, however, these small differences in scores between gymnasts can arise due to the averaging of the E-scores given by the different judges at the E-panel.

and 0.07 points, depending on the model specification. Contrarily, further research by Rotthoff (2020), using data from the 2013 World Championships, suggests that overall order bias might have reduced over time and is no longer present in the sport of gymnastics. However, both papers by Rotthoff (2015; 2020) utilise data from only one tournament. A combination of more tournaments might reveal an effect. A possible effect could be expected through the focus of individuals being best at the beginning of a sequence (Feenberg et al., 2017). This could lead to the judges being more focused and strict for the first gymnasts. Based on previous research, a small but positive effect of start position on the performance measures is hypothesized.

Hypothesis 1: A later start position leads to better performance measures.

Sequential order bias describes the type of bias induced by the performance of the previous competitor. In the context of gymnastics, this could run through both the gymnast or the judges. The gymnast, or more generally the competitor or contestant, might be influenced by the quality of the previous performance. For example, if the previous competitor falls off the apparatus, this might result in nerves and a less cleanly performed routine for the following competitor. Another option is that the following competitors aim to perform better than the previous competitors, rather than aiming for a certain score. The goal is therefore dependent on the performance of the previous competitor. This idea of reference dependence and loss aversion in sports has been shown in golf tournaments (Elmore & Urbaczewski, 2021). When the reference score was better, competitors performed better as well. The other path could be through the judges, or more generally, the decision makers. Each routine is scored directly after the performance and the following gymnast has to wait until the score is posted. The final E-score of a gymnast is the average score of all E-panel judges. Therefore, a judge can be influenced by the score the other judges gave to earlier routines, for example if the final E-score is substantially lower than what they scored the routine. Research has shown that decision makers are indeed influenced by this. In the field of education, it has been shown that sequential order bias is present after streaks of similar performances or after extreme results when grading of exams (Bhargava, 2008; Goldbach et al., 2022). This effect would be even greater at the end of a grading sequence, possibly due to fatigue and being less precise. Even in the world of speed dating, sequential order bias finds a way in. Bhargava and Fisman (2014) find a negative relationship between the perception of attractiveness of the previous date and the outcome of the following date. A similar patter arises with job interviews. Job acceptance is not solely based on the qualities and qualifications of the candidate, but also highly influenced by these characteristics of the previous candidate (Holmes & Berkowitz, 1961; Rowe, 1967). Contrasts are highlighted to make a candidate following a less likable candidate seem more preferred.

Within the sport of gymnastics, an experimental design performed by Scheer et al. (1983) used the posting of falsified scores to test if the judges were influenced by previous (falsified) scores, resulting in a confirmation of sequential order bias. Sequential order bias has also been previously researched using data from gymnastics competitions. The existence of this phenomenon cannot be shown by Joustra et al. (2021) and Rotthoff (2015; 2020). On the other hand, Damisch et al. (2006) have found evidence for the existence of an effect of the previous performance on the next competitor. They argued that the direction of this effect depends on the degree of similarity between consecutive athletes. If athletes are perceived as similar by the judges, the scores given to the next gymnasts were more similar to the previous score than when the athletes were not perceived to be similar.

Hypothesis 2: The quality of the performance of the previous gymnast positively influences the performance measures of the next gymnast.

A possible mechanism determining the existence of order bias is the quality of the first performance. The quality of the performance of the first gymnast might have a more extensive effect than only on the directly following competitor. The performance of the first gymnast may be used as a reference point for the rest of the competition. Similar to sequential order bias, this might influence the gymnasts to come. However, if the first gymnast performs an exceptionally good or bad routine, this might not only affect the next gymnast, but also the rest of the field as the reference point is set at an extreme. This might influence the direction of the overall order bias, for example if the effects move in opposite directions. It could be the case that the effect of start position is positive when the reference point has been set at one extreme and negative at the other extreme. The effect of the quality of the first performer on the following performances, however, has not been thoroughly researched. The effect of start position is expected to differ depending on the quality of the performance of the first gymnast. This could be driven by unconscious effects on the judges or the gymnast, or tactical decisions by gymnasts and their coaches. For example, a higher risk could be taken by gymnasts later in line to try to get a score that is higher their opponents score. If an effect is found, it is likely that this extends to other sequential contexts where individuals observe each other's performances.

Hypothesis 3: The quality of the performance of the first gymnasts positively influences the performance measures of the following gymnasts.

Another mechanism driving order bias could be the time since the warmup. For apparatus finals, all gymnasts get an opportunity to warm up before the competition in a separate warmup hall. As of September 2021, gymnasts are also allowed a 'one-touch warmup' in the competition arena shortly before the apparatus finals, which was not allowed before (Fédération Internationale de

Gymnastique, 2021). This policy change was implemented after too many (dangerous) mistakes occurred during the Tokyo 2020 Olympic Games. A one-touch warmup typically gives the gymnast 30 to 60 seconds to warm up on the apparatus before their competition routine. Firstly, the gymnasts who compete in the first half of the final will do their warmup, followed by their competition routines. After the first four competition routines have been performed, the second half of the rotation will follow with their warmup and competition routines. This break in the competition also allows for the judges to take a micro-break from scoring routines. A meta-analysis done by Albulescu et al. (2022) shows that micro-breaks are effective in reducing fatigue in many different contexts. Giving the judges a break in the middle of a competition might enhance their sharpness and precision in scoring the routines to come. On the contrary, Danziger et al. (2011) show that judicial rulings are much more favourable after the judge has had a break. The percentage of favourable rulings lied around 65 percent after a break and drops to close to zero in the following session up until the next break, when it will jump back up to 65 percent.

The one-touch warmup was already allowed for qualification, all around and team finals. Groups of gymnasts warm up at roughly the same time, but compete sequentially. An apparatus final typically lasts half an hour to an hour, depending on the apparatus. This means that there is a substantial difference in time between warmup and competition between gymnasts at different start positions, especially when no one-touch warmup is allowed, which might negatively influence the gymnasts later in line. No research has yet been done on this effect, let alone the influence of the recent policy change.

Hypothesis 4: A larger time gap between warmup and competition leads to worse performance measures.

Order bias is not the only type of bias present in gymnastics. Other previously researched biases include national bias, memory bias and difficulty bias. National bias is described as a judge scoring gymnasts from their own nationality in a more favourable way. This phenomenon has been empirically shown by Leskošek et al. (2012) and Heiniger and Mercier (2021) by using data from 2011 and 2013 to 2016 respectively. The international governing body of the sport has taken measures to minimise such bias. The technical regulations of the Fédération Internationale de Gymnastique (2023) state that during apparatus finals, no judge on the E-panel can be of the same nationality as any of the finalists, nor the first reserve gymnast. After eliminating judges from the same nationalities of the gymnasts, the selection of judges to judge the apparatus finals are determined by a drawing of lots. In the context of gymnastics, memory bias is described as a lower level of accuracy in scoring an element when the judges are less familiar with the element (Ste-Marie, 2003). Difficulty bias shows a positive

relationship between the difficulty of a routine and the E-score (Morgan & Rotthoff, 2014; Rottoff, 2020). Reputation bias could also play a role in gymnastics, although not yet researched in this sport. Findlay and Ste-Marie (2004) show that better known athletes scored higher on the technical execution in figure skating, even with a similarly executed routine. From the athletes side, they might increase their level of difficulty, and therefore the risk they take, when 'superstar' athletes like Simone Biles are present (Meissner et al., 2021). Furthermore, the presence of superstars has been shown to influence performance in tennis, where it would reduce the probability of other top 20 players to advance to following rounds (Deutscher et al., 2022). These biases possibly relate to order bias, as they could be more pronounced at the beginning or end of the competition. More research between the interactions of biases is needed to determine the roleplay between the biases.

3. Research context

Women's artistic gymnastics consists of four apparatus that gymnasts compete on: vault (VT), uneven bars (UB), balance beam (BB) and floor exercise (FX). Since 2004, each routine is scored on its difficulty (D-score) and execution (E-score) by a panel of judges, making the maximum attainable score differ for each competitor (Fédération Internationale de Gymnastique, 2023). The D-score summarises the difficulty of the routine and consists of the composition requirements (CR), difficulty value (DV) of the skills and connection value (CV) of the gymnast linking skills together. The D-panel consists of two drawn judges. The difficulty of the elements gymnasts can perform is revised after each Olympic cycle, after which a new Code of Points (CoP) is released in which all rules and values are stated. The E-score summarises the execution of the routine and starts from 10 points. Each mistake a gymnast makes results in a deduction that lowers the E-score. Worse mistakes lead to bigger deductions. For example, a fall will result in a 1.0 deduction, whereas a slightly bent leg will result in a 0.1 deduction. The E-panel consists of five randomly drawn judges.

International elite gymnastics tournaments start with a qualification round where gymnasts and (country) teams can qualify for all around, team and apparatus finals (Fédération Internationale de Gymnastique, 2023). The top eight qualifiers advance to each final. If a qualifier is unable to compete in the final, for example due to an injury, a reserve gymnast (the ninth gymnast in qualifications) will take their place. Furthermore, only two gymnasts per country are allowed to compete in a final. If three gymnasts of the same country place in the top eight during qualification, only the top two gymnasts of that country will advance to the final. The ninth best gymnast in the qualifications will also advance to the final. If many gymnasts from the same country qualify in the top eight, only the best two can compete in the final and the next best gymnasts from different countries will compete in the final. This explains why sometimes the fourteenth gymnast in the qualification round competes in a final.

An essential fact for this research is that the order of competition in apparatus finals is determined by means of a random draw. This is not the case for the all around and team finals, where competition groups and start positions are determined based on qualification ranks. Therefore, this thesis will focus specifically on apparatus finals. Each gymnast has an equal chance of competing first, last or somewhere in between, independent of their qualifying rank or score. This makes apparatus finals the perfect setting to research the effects of competition order on performance, as on average, the only difference between the gymnasts is their start position.

4. Data

4.1. Data description

The research question will be investigated using empirical data from the most recent international elite women's artistic gymnastics tournaments. A dataset has been created specifically for this research including data on qualifications and apparatus finals. The data have been extracted from the official results books of the tournaments using web scraping (*Fédération Internationale de Gymnastique - Results*, 2023; *GYMmedia.com*, 2023; *Gymnastics Results*, 2023; *The Gymternet*, 2023). When any of the data was not available in document form, it was extracted from the official recordings of the competitions. Strikingly high or low scores have been verified with the recordings as well.

The sample consists of tournaments from 2018 to 2023, including Olympic Games, World Championships, Continental Championships and World (Challenge) Cups and consists of 1684 individual-level observations. Each observation includes the total score, E-score, D-score and penalty for both the apparatus final and qualification of a gymnast on one apparatus. Additionally, the ranks of both the qualification round and final have been collected, as well as the start position for all apparatus finals. Thus, a total of 3368 routines are included in the dataset. To qualify and compete in a vault apparatus final, a gymnast has to compete two different vaults. Therefore, the vault scores documented in the dataset are the final average vault scores. For the uneven bars, balance beam and floor exercise apparatus finals, a gymnast only gets one chance to perform their routine. Lastly, some descriptive variables have been included, such as the apparatus, the year, location and type of tournament, the Code of Points, nationality of the gymnast, and the type of warmup in the finals.

To study the effect of start position on the performance of a gymnast, two continuous measures for performance will be used throughout this thesis. Firstly, the dependent variable is difference between the E-score in the apparatus final and the qualification round. Secondly, the dependent variable is the difference between the total score in the apparatus final and the qualification round. Taking the difference between the final and qualification score will account for individual differences between gymnasts, as well as for differences between tournaments. Judges might be more or less strict in certain tournaments compared to others, which will be accounted for using this method. The D-score will not be used as a dependent variable, as it is often decided on by the gymnast and coach before the competition takes place and is measured objectively, whereas the E-score measures the performance of the skills. If an overall order bias might occur, this is more likely to be in the E-score than in the D-score. The performance of a routine is slightly different each time and valued subjectively by the judges, which makes it more prone to biases.

The independent variable in all of the regressions will be the start position of the gymnasts and will be taken as a categorical variable to not make any assumptions on the mathematical shape of a possible bias. The first start position is taken as a reference category. As mentioned before, the order of performance in apparatus finals is decided by a random draw. This allows for the use of randomisation of treatment, which will be verified in paragraph 5.5.

4.2. Descriptive statistics

A total of 54 tournaments have been included in the dataset, of which 1 Olympic Games, 4 World Championships, 12 Continental Championships and 37 World (Challenge) Cups. The tournaments were hosted in 21 different countries. This selection of tournaments has been made because of the availability of data as well as them being at the highest level. The best judges are selected to judge at these tournaments and the gymnasts typically have been training for a long time and are experienced in competing under pressure. If there would be an order bias in these tournaments, it would be reasonable to assume that order bias is also present in lower level tournaments.

The dataset contains 50 vault finals, 53 uneven bars finals, 54 balance beam finals and 53 floor exercise finals. Finals in which less than 8 gymnasts competed have been removed from the dataset as these only occur when a small number of gymnasts compete during the qualification round, resulting in everyone qualifying to the apparatus final, even if the qualifying routine was a failed routine. 4 of the finals included in the dataset, 2 balance beam and 2 floor exercise, included 9 gymnasts. This happens in the rare case that the 8^{th} and 9^{th} ranked gymnast in qualification score exactly the same and the tie-breaking rules cannot break the tie.[2]

Table 1 shows the descriptive statistics of the most used variables in the analysis. The means of the performance variables in the qualification round are on average higher than those in the finals. This can be explained by the bigger field of competition in the qualification round and only the top gymnasts qualifying to the finals. The competitors that made big mistakes in the qualification round are less likely to qualify to a final. As the reported performance measures for the qualification round only includes gymnasts that did qualify to the final, this mostly includes successful qualification routines, whereas this is not the case in the finals. Every gymnast has to perform their routine again and this routine is scored independently of their qualification routine. The dataset contains all routines

[2] The technical regulations of the International Gymnastics Federation define the tie-breaking rules for qualifying to apparatus finals as follows. In the case of a tie, the gymnast with the highest E-score prevails. If the E-scores are equal, the gymnast with the highest D-score prevails. Penalties are a separate component of the total score and is not taken into account when comparing E-scores and D-scores in case of a tie-break (Fédération Internationale de Gymnastique, 2023).

performed in the apparatus finals, regardless of the quality of the routine. Although the gymnasts in the finals are expected to perform better on average than gymnasts who did not qualify, the finals scores are not by definition higher. The scores are compared to the gymnasts own qualification scores, instead of to all gymnasts who competed in qualifications. Whereas the qualification routines were filtered to mostly contain successful routines, this is not the case for the apparatus finals routines. This line of reasoning can be extended to the higher standard deviations in the finals than in the qualifications and the negative signs of the differences in scores between finals and qualification. Table A1 shows the descriptive statistics of the scores sorted by apparatus. It can be inferred that the average E-scores and total scores in finals are highest for vault in this sample. Scores are generally lowest on balance beam.

Table 1. Descriptive statistics

Variable	Obs.	Mean	Std. Dev.	Min.	Max.
E-score final	1684	7.817	0.835	2.525	9.566
D-score final	1684	5.085	0.640	0.400	6.700
Total score final	1684	12.869	1.170	2.300	15.399
Rank final	1684	4.509	2.306	1	9
E-score qualification	1684	7.997	0.624	4.500	9.566
D-score qualification	1684	5.124	0.608	2.100	7.200
Total score qualification	1684	13.107	0.868	9.500	15.666
Rank qualification	1684	4.746	2.565	1	14
E-score difference (f - q)	1684	-0.181	0.613	-5.241	2.500
D-score difference (f - q)	1684	-0.040	0.295	-3.800	1.200
Total score difference (f - q)	1684	-0.238	0.826	-10.434	2.466
Rank difference (f - q)	1684	-0.238	2.375	-9	7
Type of tournament					
Olympic Games	1684	0.019	-	0	1
World Championships	1684	0.077	-	0	1
Continental Championships	1684	0.223	-	0	1
World (Challenge) Cup	1684	0.681	-	0	1
One-touch warmup allowed	1552	0.428	-	0	1

Note: The E-score starts from 10 points and points are deducted for each mistake. The D-score starts from 0 points and points are added for all elements performed and requirements fulfilled The total score is the sum of the D-score and E-score, minus a possible penalty. A penalty is given for possible mistakes that are not included in the E-score, such as exceeding the time limit or stepping out of the permitted areas during a routine. The score differences are calculated by taking the difference between respectively the score or rank in the final (f) and qualification (q) round. The type of tournament and whether a one-touch warmup was allowed are shown in proportions.

5. Methodology

Ordinary Least Squares (OLS) regressions will be used in this analysis. Before continuing with the description of the analysis, the assumptions of OLS will be discussed. The Zero Conditional Mean (ZCM) assumption states that the expected value of the error term conditional on the explanatory variable is equal to zero. The assignment of start position to gymnasts is determined by means of a random draw. Whether the treatment assignment was fully randomised is tested in paragraph 5.5. If there is no reason to suspect that the randomisation was not successful, it can be assumed that the ZCM assumption hold. There would be no other determinants of the explanatory variable. The second assumption of OLS states that the observations should be independent and identically distributed (i.i.d.). This assumption is not fully satisfied, as observations within tournaments could be related. For example, if one person performs in multiple apparatus finals included in the dataset, they might be more tired in the last final. Additionally, if a gymnast is set to go after an exceptionally good or bad routine, this might influence the performance. The observations have been drawn from the same population distribution. A violation of this assumption leads to incorrect standard errors, which can be fixed by using robust standard errors. The third assumption of OLS states that large outliers in the dependent and independent variables are unlikely. Some outliers are present in the dataset with regards to the score differences, however, they are not likely to occur. It is possible that a gymnast performs a really disappointing routine in the final or fails to complete the routine, resulting in a bad score, but this only happens occasionally. Outliers in the independent variable, start position, are impossible by definition, as they range from 1 to 8 with 4 observations having start position 9. Each start position (apart from start position 9) has the same number of observations. There are no issues with the assumptions, meaning that OLS can be used for the analysis.

5.1. Overall order bias

The first hypothesis states that a later start position leads to better performance measures and will be established using single linear regression. The direct relationship between start position and performance will be investigated using two different measures for performance, as is shown in equation 5.1. The dependent variables used in this analysis are the differences between the final and qualification score, looking separately at the E-score and the total score. Measuring the quality of performance by taking the differences between the finals and qualifications has the advantage that it adjusts for individual differences between gymnasts and what is an attainable score for them. It also adjusts for differences in level and strictness in scoring across competitions and apparatus.

The explanatory variable is the start position in the apparatus final. Each start position will be included in the model as a dummy variable to avoid making any assumptions on the mathematical

form of the effect. Adding the start positions as separate dummy variables allows for non-linearity and non-monotonicity of a possible order bias. The reference category for start position is taken as the first start position, as this has the most intuitive interpretation. Lastly, the analysis will be performed for each apparatus separately as well, as the order effect may differ per apparatus. This would yield the same results as an interaction term would, but also gives the correct p-values and significance levels for each apparatus separately.

$$Y_i = \alpha + \boldsymbol{\beta} * Start\ position_i + \varepsilon_i \quad (5.1)$$

$$Y_i \in \{Difference\ E - score_i, Difference\ total\ score_i\}$$
$$\boldsymbol{\beta} = \langle \beta_2, \beta_3, \dots, \beta_9 \rangle$$

This analysis will be informative about a possible overall order bias on performance, but does not yet say anything about the way this bias is built up. It is still unclear what driving forces are behind the possible bias.

5.2. Sequential order bias

The second hypothesis states that the quality of the previous performance influences the performance of the next gymnast. The degree to which sequential order bias is explanatory of the performance of a gymnast will be tested by means of a multiple linear regression, as shown in equation 5.2. The dependent variable remains the same as in equation 5.1. The dependent variable is again the difference between the final and qualification score, taken for both the E-score and total score. The explanatory variable, the performance of the previous competitor, will be the same as the dependent variable, but with a lag of 1 routine. The start position dummy variables will be added as control variables, as this possibly affects both the explanatory and dependent variable (as described by hypothesis 1), but is not influenced by the explanatory variable. All other possible control variables that influence both the dependent and independent variable have been controlled for by taking the differences in final and qualification score for both these variables. Similarly to hypothesis 1, the analysis will be performed for the complete dataset as well as the separate apparatus. The sample size will be reduced by approximately 12% as the gymnast with the first start position cannot be used anymore, due to them not having a gymnast competing before them.

$$Y_i = \alpha + \beta * Performance\ previous\ competitor_i + \boldsymbol{\gamma} * Start\ position_i + \varepsilon_i \quad (5.2)$$

$$Y_i \in \{Difference\ E - score_i, Difference\ total\ score_i\}$$

$$\boldsymbol{\gamma} = \langle \gamma_3, \gamma_4, \dots, \gamma_9 \rangle$$

5.3. Performance of the first competitor

To uncover one of the driving forces behind the possible overall order bias, the third hypothesis states that the quality of first performances influences the performance measures of the following gymnasts. For example, if the first gymnast performs a bad routine, this might negatively influence the rest of the competition field or the judges (Elmore & Urbaczewski, 2021).

The quality of the first performance will be divided into three groups, a 'below average' performance, an 'average' performance or an 'above average' performance. These labels will be assigned based on the differences between final and qualification scores of the individual. If this difference is worse than the 25th quartile of all differences on that apparatus, the performance in the final is labelled 'below average'. If a difference is better than the 75th quartile of all differences, the performance in the final is labelled 'above average'. The remaining routines are labelled 'average'. Using the differences instead of just the finals scores adjusts for individual differences between gymnasts. For example, an average score for Simone Biles (unquestionably the best gymnast of the past decade) could easily be unachievable for many other gymnasts.

The analysis will be done by performing the same regressions as in equation 5.1, with the addition of an interaction term for the performance of the first competitor and the start position for the rest of the competitors and is shown in equation 5.3. The interaction term will separate the effect of start position on the performance measures by the quality of the first performance. If there is a different effect of start position on the performance measures for the different types of first performances, this will be reflected in the interaction terms. As was also the case in for the previous hypothesis, the sample size will be reduced by approximately 12% as the gymnast with the first start position cannot be used anymore.

$$Y_i = \alpha + \boldsymbol{\beta} * Start\ position_i + \boldsymbol{\delta} * Quality\ first\ performance_i + \boldsymbol{\gamma} * Start\ position_i * Quality\ first\ performance_i + \varepsilon_i \quad (5.3)$$

$$Y_i \in \{Difference\ E - score_i, Difference\ total\ score_i\}$$
$$\boldsymbol{\beta} = \langle \beta_2, \beta_3, \dots, \beta_9 \rangle, \boldsymbol{\gamma} = \langle \gamma_1, \gamma_2, \dots, \gamma_{15} \rangle, \boldsymbol{\delta} = \langle \delta_1, \delta_2 \rangle$$

Based on the found results, the effect of a good or bad performance on the rest of the field can be concluded. This does not say anything on why the first competitor performs the way they do.

5.4. One-touch warmup

Another possible mechanism of the effect of start position on performance will be investigated through relative warmup times. The fourth hypothesis states that a larger time gap between warmup and competition leads to worse performance measures. The introduction of the 'one-touch warmup' after the Tokyo Olympic Games in 2021 allowed gymnasts to warm up on the competition equipment right before the start of the competition. This already was, and remains, the case for qualification rounds, but not for apparatus finals as it would disrupt the flow of the competition for spectators and television streaming. The new policy drastically reduced the time between warmup and competition in apparatus finals, especially for the gymnasts with a later start position. For apparatus finals, the one-touch warmup takes place in two separate groups, split by competition order (Fédération Internationale de Gymnastique, 2023). The effect of the one-touch warmup will be researched through the effect of a split in the competition for the last four gymnasts to warm up.

To investigate if the one-touch warmup affects the relationship between individual start positions and the performance measures, regressions will be estimated that includes only those finals where there was a one-touch warmup separated for the first four and last four competitions. If there is an effect of the time between the warmup and competition on performance, then there should be a similar trend for the gymnasts performing at the same start position before and after the split in the competition. The fifth gymnast in the competition is the first gymnast to compete after the second round of warmups. Two regression models will be estimated according to the form of equation 5.4. The first model only includes the gymnasts with start positions one to four. The second model includes the gymnasts with start positions five to eight. The constant and coefficients of the respective start positions will be plotted and compared to conclude whether having the same start position before or after the split has a similar effect. If these effects are indeed similar and such a similarity is not found for competitions where there was no touch warmup, this suggests that there is an effect of warmups shortly before the competition on performance.

$$Y_{it} = \alpha + \boldsymbol{\beta} * Start\ position_{it} + \varepsilon_i \tag{5.4}$$

$$Y_{it} \in \{ Difference\ E - score_{it}, Difference\ total\ score_{it} \}$$
$$\boldsymbol{\beta} = \langle \beta_2, \beta_3, \beta_4 \rangle$$

5.5. Randomisation

To test whether the randomisation of start position in apparatus finals was successful, three checks were performed. These checks are only done with respect to the qualification E-score, D-score and

total score, as these are determined prior to the finals and are therefore independent of any effect of start position in the final. First, the descriptive statistics of the qualification scores, sorted by apparatus and start position in the final, are shown in Table A2a-d. The statistics in these tables do not immediately show a trend or reason to suspect problems in randomisation. Second, to validate these results, Table 2 shows the correlations between the qualification measures and the start position in the apparatus final, sorted by apparatus. This table shows very weak correlations. There seems to be no relationship between the qualification performance of a gymnast and their start position in the apparatus final. Third, using the same measures for qualification performance, 12 Kruskal-Wallis tests were performed to test whether at least one of the start position groups differs systematically in performance. The null hypothesis of no systematic difference in any of the groups cannot be rejected for any of the tests at a significance level of 5 percent, as shown in Table 3. Based on these findings, it can be concluded that there is no reason to suspect that the randomisation has not been successful. Therefore it can be concluded that there are no systematic differences in qualification performance based on start position in the apparatus finals.

Table 2. Correlation table of qualifications performance and start position, sorted by apparatus

VARIABLES	Start position (VT)	Start position (UB)	Start position (BB)	Start position (FX)
E-score qualification	-0.043	-0.014	-0.002	0.077
	(0.390)	(0.778)	(0.964)	(0.112)
D-score qualification	-0.002	0.029	0.029	0.070
	(0.963)	(0.557)	(0.551)	(0.147)
Total score qualification	-0.023	0.010	0.016	0.084
	(0.651)	(0.833)	(0.742)	(0.082)

Note: Each cell shows the correlation between a qualification score component and the start position for one of the four apparatus. The standard errors are shown in parentheses.
**** p<0.01, ** p<0.05, * p<0.1*

Table 3. Kruskal-Wallis test of systematic differences in scores, sorted by apparatus

VARIABLES	Vault	Uneven bars	Balance beam	Floor exercise
E-score qualification	$\chi2$(8, N=400) = 8.123, p=0.322	$\chi2$(8, N=434) = 3.968, p=0.832	$\chi2$(8, N=434) = 3.968, p=0.860	$\chi2$(8, N=424) = 9.041, p=0.339
D-score qualification	$\chi2$(8, N=400) = 5.093, p=0.649	$\chi2$(8, N=434) = 4.411, p=0.731	$\chi2$(8, N=434) = 9.818, p=0.278	$\chi2$(8, N=424) = 5.393, p=0.715
Total score qualification	$\chi2$(8, N=400) = 6.701, p=0.461	$\chi2$(8, N=434) = 3.370, p=0.849	$\chi2$(8, N=434) = 5.281, p=0.727	$\chi2$(8, N=424) = 9.587, p=0.295

Note: Each cell shows a separate Kruskal-Wallis test for a qualification score component for one of the four apparatus.

6. Results

6.1. Overall order bias

The existence of order bias in women's artistic gymnastics has been shown in numerous previous studies (Ansorge et al., 1978; Joustra et al., 2021; Rotthoff, 2015; Scheer & Ansorge, 1975). Using a more recent dataset and a new measure for performance, Table 4 reveals the effect of start position on the difference in E-score between the apparatus final and qualification round. Table 5 shows the effect of start position on the difference in total score between the apparatus final and qualification round. Both tables show the results for all apparatus combined in column 1 and the results for each apparatus separately in columns 2 to 5. The null hypothesis of no effect of start position on the difference in E-score and total score in the population that the sample was taken from, cannot be rejected at a significance level of 5 percent when looking at column 1 of Table 4 and 5. The same can be said about vault, uneven bars and balance beam score differences in columns 2, 3 and 4 from Table 4 and 5. When looking at the results for floor exercise score differences in column 5 of Table 4 and 5, it can be concluded that gymnasts starting second on floor exercise, on average, respectively score 0.143 and 0.433 higher than they did in qualification on both E-score and total score compared to the first start position. Nothing can be concluded with respect to the other start positions relative to the first start position. Each model is estimated with respect to these two dependent variables to check the robustness of the results. If one model does show an effect and the other does not, the result is less likely to be robust, as could be the case for start position 3 in column 3 of Table 4.

It cannot be concluded that the first hypothesis, stating that gymnasts with later start positions have better performance measures, holds. There are multiple possible explanations for the lack of any found effect. In terms of fairness, the most hopeful scenario is that there is no order bias anymore, as suggested by Rotthoff (2020). This could be the case with the used dataset, as it contains competitions from 2018 until 2023 at the highest possible level of the sport. The judges for these type of competitions are very carefully selected and have to go through extensive training, including much experience, to be able to judge at this level. The same holds for the gymnasts. They have competed in many competitions before being able to compete at the level of competition included in the dataset. When looking at lower levels of competition, one might be able to find evidence for order bias.

Another option is that there are multiple mechanisms that add to the existence of order bias that possibly work in different directions and cancel each other out. Therefore, it can still be interesting to study two of these possible mechanisms in the rest of this thesis.

Table 4. Regression results overall order bias on E-score difference

VARIABLES	(1) E-score difference	(2) E-score difference	(3) E-score difference	(4) E-score difference	(5) E-score difference
Apparatus	All	VT	UB	BB	FX
Start position					
2	0.0373	-0.0601	0.0135	0.0474	0.143**
	(0.0554)	(0.109)	(0.126)	(0.125)	(0.0667)
3	0.0986*	0.0244	0.248**	0.0802	0.0382
	(0.0532)	(0.0474)	(0.115)	(0.144)	(0.0852)
4	0.0152	-0.0119	0.0517	0.131	-0.114
	(0.0531)	(0.0537)	(0.134)	(0.124)	(0.0858)
5	-0.0124	-0.0329	0.0548	-0.115	0.0437
	(0.0582)	(0.111)	(0.125)	(0.140)	(0.0725)
6	0.0103	0.0657	-0.00540	-0.0575	0.0430
	(0.0555)	(0.0476)	(0.145)	(0.138)	(0.0610)
7	0.0990*	-0.0397	0.214*	0.184	0.0282
	(0.0561)	(0.111)	(0.110)	(0.140)	(0.0777)
8	0.0298	0.0246	-0.0141	0.0219	0.0866
	(0.0591)	(0.0535)	(0.132)	(0.168)	(0.0677)
9	-0.173			0.00991	-0.308
	(0.183)			(0.145)	(0.337)
Constant	-0.215***	-0.0785**	-0.295***	-0.318***	-0.158***
	(0.0354)	(0.0324)	(0.0846)	(0.0909)	(0.0504)
Observations	1,684	400	424	434	426
R-squared	0.004	0.006	0.020	0.014	0.033

Note: The score differences are calculated by taking the difference between the score in the final and qualification round. The standard errors are shown in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table 5. Regression results overall order bias on total score difference

VARIABLES	(1) Total score difference	(2) Total score difference	(3) Total score difference	(4) Total score difference	(5) Total score difference
Apparatus	All	VT	UB	BB	FX
Start position					
2	0.0951	-0.0870	-0.0375	0.0623	0.433**
	(0.0820)	(0.153)	(0.149)	(0.148)	(0.196)
3	0.0648	0.00253	0.231	0.0136	0.00987
	(0.0963)	(0.0619)	(0.189)	(0.162)	(0.280)
4	0.0555	-0.0688	0.00455	0.148	0.130
	(0.0788)	(0.0838)	(0.161)	(0.150)	(0.205)
5	0.0536	-0.0857	0.108	-0.116	0.304
	(0.0827)	(0.156)	(0.142)	(0.155)	(0.198)
6	0.0720	0.101*	0.0267	-0.0483	0.213
	(0.0783)	(0.0593)	(0.160)	(0.162)	(0.195)
7	0.133	-0.0876	0.201	0.143	0.264
	(0.0837)	(0.169)	(0.132)	(0.161)	(0.201)
8	0.0798	0.0707	-0.0537	-0.0466	0.351*
	(0.0804)	(0.0634)	(0.148)	(0.181)	(0.199)
9	-0.206			0.130	-0.357
	(0.309)			(0.133)	(0.579)
Constant	-0.307***	-0.0746*	-0.340***	-0.389***	-0.409**
	(0.0613)	(0.0425)	(0.0974)	(0.108)	(0.188)
Observations	1,684	400	424	434	426
R-squared	0.002	0.010	0.014	0.011	0.033

Note: The score differences are calculated by taking the difference between the score in the final and qualification round. The standard errors are shown in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

6.2. Sequential order bias

The analysis performed in section 6.1 and the analysis that will follow in sections 6.3 and 6.4 investigates the possible existence of overall order bias. Sequential order bias captures the effect of the performance of the previous competitor on the performance of the following competitor, regardless of the overall order. Table 6 and 7 respectively show the regression results of sequential order bias on the difference in E-score and total score. Column 1 includes all observations, whereas columns 2 to 5 only include the observations for specific apparatus. Table 6 and 7 reveal no evidence for a sequential order bias on the difference between qualification and final E-score and total score on all apparatus combined nor any of the separate apparatus at a significance level of 5 percent. It cannot be concluded that there is a significant effect of the performance measures of the previous competitor on the following gymnasts performance measures. This is in line with the findings of Rotthoff (2015; 2020) and Joustra et al. (2021).

Table 6. Regression results sequential order bias on E-score difference

VARIABLES	(1) E-score difference	(2) E-score difference	(3) E-score difference	(4) E-score difference	(5) E-score difference
Apparatus	All	VT	UB	BB	FX
E-score difference previous	0.0103	-0.0118	0.0701	-0.0491	-0.0131
	(0.0308)	(0.0187)	(0.0459)	(0.0690)	(0.0553)
Start position					
3	0.0609	0.0838	0.233*	0.0351	-0.103
	(0.0583)	(0.110)	(0.121)	(0.141)	(0.0825)
4	-0.0231	0.0485	0.0208	0.0874	-0.256***
	(0.0582)	(0.113)	(0.139)	(0.122)	(0.0819)
5	-0.0499	0.0271	0.0377	-0.156	-0.100
	(0.0629)	(0.149)	(0.130)	(0.137)	(0.0685)
6	-0.0268	0.125	-0.0227	-0.111	-0.0990*
	(0.0604)	(0.110)	(0.150)	(0.135)	(0.0558)
7	0.0616	0.0212	0.201*	0.134	-0.114
	(0.0610)	(0.149)	(0.117)	(0.136)	(0.0740)
8	-0.00849	0.0842	-0.0425	-0.0165	-0.0556
	(0.0637)	(0.113)	(0.139)	(0.165)	(0.0633)
9	-0.206			-0.0828	-0.448
	(0.186)			(0.158)	(0.335)
Constant	-0.175***	-0.140	-0.261***	-0.287***	-0.0177
	(0.0434)	(0.105)	(0.0934)	(0.0900)	(0.0452)
Observations	1,474	350	371	380	373
R-squared	0.005	0.006	0.025	0.017	0.036

Note: The score differences are calculated by taking the difference between the score in the final and qualification round. The standard errors are shown in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table 7. Regression results sequential order bias on total score difference

VARIABLES	(1) Total score difference	(2) Total score difference	(3) Total score difference	(4) Total score difference	(5) Total score difference
Apparatus	All	VT	UB	BB	FX
Total score difference previous	0.0274	-0.0108	0.0484	-0.0329	0.0782*
	(0.0266)	(0.0156)	(0.0459)	(0.0658)	(0.0473)
Start position					
3	-0.0329	0.0886	0.270	-0.0466	-0.457**
	(0.0922)	(0.154)	(0.198)	(0.158)	(0.221)
4	-0.0414	0.0183	0.0308	0.0858	-0.304***
	(0.0737)	(0.164)	(0.172)	(0.146)	(0.0935)
5	-0.0431	0.000567	0.145	-0.174	-0.139*
	(0.0778)	(0.210)	(0.153)	(0.152)	(0.0829)
6	-0.0246	0.187	0.0589	-0.114	-0.244***
	(0.0732)	(0.153)	(0.169)	(0.159)	(0.0767)
7	0.0362	0.000539	0.237	0.0796	-0.186**
	(0.0790)	(0.220)	(0.144)	(0.157)	(0.0883)
8	-0.0190	0.157	-0.0259	-0.104	-0.103
	(0.0754)	(0.154)	(0.160)	(0.176)	(0.0840)
9	-0.288			0.0351	-0.812
	(0.313)			(0.144)	(0.565)
Constant	-0.203***	-0.162	-0.361***	-0.339***	0.0560
	(0.0556)	(0.147)	(0.114)	(0.106)	(0.0575)
Observations	1,474	350	371	380	373
R-squared	0.002	0.010	0.017	0.013	0.045

Note: The score differences are calculated by taking the difference between the score in the final and qualification round. The standard errors are shown in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

6.3. Performance of the first competitor

As stated before, finding no evidence of order bias under the first and second hypothesis does not mean that it can be concluded that order bias does not exist. It could be the case that the effect of start position differs between the competitions based on the performance of the first competitor. If the first gymnast performs a below average routine, this might negatively affect the rest of the competition field as this could be a warning sign or be used as a reference point. The second start position is taken as a reference category, as the first start position is omitted from the analysis due to them not having a reference point before their routine. For the quality of the first performance, the group of average routines is taken as a reference category.

For this analysis, the outcome of interest lies in the interaction terms between start position and the quality of the first performance. This shows if there is a differential effect of start position depending on the quality of the first performance. The interaction terms in Table 8 and 9 do not show any evidence for such a difference. Although some of the coefficients are statistically significant at a significance level of 5 percent with respect to the reference category of average routines, the sign of the coefficients for below and above average are the same, with overlapping confidence intervals. There does not seem to be a significant difference in coefficients for the interaction terms below and above average first performance and start position. It is therefore unlikely that the quality of the first performance is a systematic factor that determines order bias.

Table 8. Regression results quality of first performance on E-score difference

VARIABLES	(1) E-score difference	(2) E-score difference	(3) E-score difference	(4) E-score difference	(5) E-score difference
Apparatus	All	VT	UB	BB	FX
Start position * quality of first performance					
3 * below average	-0.107	-0.285	0.0937	-0.120	-0.105
	(0.135)	(0.217)	(0.331)	(0.293)	(0.183)
3 * above average	-0.165	-0.330	-0.0564	-0.194	-0.0210
	(0.151)	(0.216)	(0.256)	(0.461)	(0.197)
4 * below average	-0.0643	-0.134	0.186	-0.253	-0.000105
	(0.136)	(0.209)	(0.340)	(0.274)	(0.189)
4 * above average	0.0636	-0.0165	0.282	0.120	-0.0156
	(0.144)	(0.228)	(0.275)	(0.369)	(0.231)
5 * below average	0.0201	-0.0392	0.509	-0.0813	-0.288*
	(0.142)	(0.264)	(0.315)	(0.294)	(0.163)
5 * above average	0.0113	-0.0312	0.170	0.104	-0.177
	(0.157)	(0.273)	(0.304)	(0.443)	(0.169)
6 * below average	0.0226	-0.297	0.295	-0.0348	0.133
	(0.135)	(0.206)	(0.335)	(0.308)	(0.118)
6 * above average	-0.121	-0.216	-0.698	0.273	0.114
	(0.169)	(0.231)	(0.488)	(0.399)	(0.156)
7 * below average	0.0132	-0.0907	0.137	0.0220	-0.0336
	(0.138)	(0.270)	(0.295)	(0.316)	(0.151)
7 * above average	-0.0122	0.0769	-0.239	0.306	-0.133
	(0.155)	(0.272)	(0.264)	(0.424)	(0.235)
8 * below average	-0.177	-0.186	0.105	-0.414	-0.140
	(0.141)	(0.199)	(0.350)	(0.307)	(0.147)
8 * above average	-0.164	-0.559**	0.205	-0.322	0.0375
	(0.189)	(0.231)	(0.271)	(0.652)	(0.159)
9 * below average	0.718***				0.947***
	(0.150)				(0.0903)
Constant	-0.204***	-0.236	-0.267**	-0.263**	-0.0228
	(0.0621)	(0.172)	(0.119)	(0.104)	(0.0655)
Start position	✓	✓	✓	✓	✓
Quality of first performance	✓	✓	✓	✓	✓
Observations	1,474	350	371	380	373
R-squared	0.012	0.042	0.069	0.033	0.075

Note: The score differences are calculated by taking the difference between the score in the final and qualification round. The standard errors are shown in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table 9. Regression results quality of first performance on total score difference

VARIABLES	(1) Total score difference	(2) Total score difference	(3) Total score difference	(4) Total score difference	(5) Total score difference
Apparatus	All	VT	UB	BB	FX
Start position * quality of first performance					
3 * below average	0.103	-0.350	0.603	0.0177	-0.0207
	(0.201)	(0.256)	(0.470)	(0.385)	(0.381)
3 * above average	0.145	-0.291	0.582	-0.170	0.411
	(0.249)	(0.288)	(0.482)	(0.551)	(0.371)
4 * below average	0.00387	-0.0969	0.316	-0.332	0.0894
	(0.179)	(0.250)	(0.423)	(0.354)	(0.244)
4 * above average	0.319	-0.315	0.805*	0.406	0.113
	(0.219)	(0.304)	(0.423)	(0.482)	(0.299)
5 * below average	0.105	0.0240	0.804**	0.0316	-0.591**
	(0.174)	(0.314)	(0.345)	(0.356)	(0.243)
5 * above average	0.132	-0.525	0.717	0.151	-0.111
	(0.234)	(0.409)	(0.442)	(0.534)	(0.218)
6 * below average	0.193	-0.441	0.793**	0.246	0.0699
	(0.168)	(0.269)	(0.382)	(0.377)	(0.167)
6 * above average	0.144	-0.355	0.281	0.442	0.174
	(0.213)	(0.302)	(0.495)	(0.462)	(0.196)
7 * below average	0.120	0.0225	0.311	0.105	-0.0603
	(0.176)	(0.339)	(0.374)	(0.367)	(0.188)
7 * above average	0.136	-0.166	0.173	0.677	-0.323
	(0.228)	(0.364)	(0.410)	(0.508)	(0.274)
8 * below average	-0.155	-0.226	0.106	-0.227	-0.264
	(0.177)	(0.236)	(0.401)	(0.368)	(0.227)
8 * above average	0.0670	-1.091**	0.626	-0.0260	0.238
	(0.276)	(0.437)	(0.427)	(0.738)	(0.195)
Constant	-0.184***	-0.234	-0.262**	-0.244**	0.00112
	(0.0683)	(0.199)	(0.125)	(0.113)	(0.0639)
Start position	✓	✓	✓	✓	✓
Quality of first performance	✓	✓	✓	✓	✓
Observations	1,474	350	371	380	373
R-squared	0.006	0.029	0.057	0.041	0.064

Note: The score differences are calculated by taking the difference between the score in the final and qualification round. The standard errors are shown in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

6.4. One-touch warmup

The second mechanism that is investigated is whether the time gap between warmup and competition influences order bias. Gymnasts with a later start position have to wait longer to compete after the general warmup, which takes place in a separate warmup hall. A one-touch warmup allows for gymnasts to warmup on the competition apparatus just before their competition routine. The competition is split into two parts, as the last four gymnasts warmup in the middle of the competition. Therefore, when looking at the time between warmup and competition, the first four gymnasts would be comparable to the last four gymnasts in terms of time between warmup and competition. For this part of the analysis, the gymnasts with start position 9 have been disregarded, as there is no corresponding start position before the split in the competition to match to the 9^{th} start position. This takes away 4 observations in the complete dataset and 2 observations in the subsample of competitions with a touch warmup.

Figure 1 shows the coefficients and 95 percent confidence intervals of the start positions of gymnasts competing in a final where the one-touch warmup was allowed and split the competition into two parts. The blue coefficient and intervals correspond to the gymnasts with start position 1 to 4 (Table A3 column 1), whereas the red coefficients and intervals correspond to the gymnasts with start position 5 to 8 (Table A3 column 2). The coefficients and intervals in the upper panel of Figure 1 look quite similar before and after the split, indicating that there might be a similarity between the two groups. A similar picture is shown in the lower panel of Figure 1 (Table A4). However, the coefficients of start positions 2 to 4 are not significantly different from each other and from start position 1, making it difficult to compare the coefficients before and after the split in the competition.

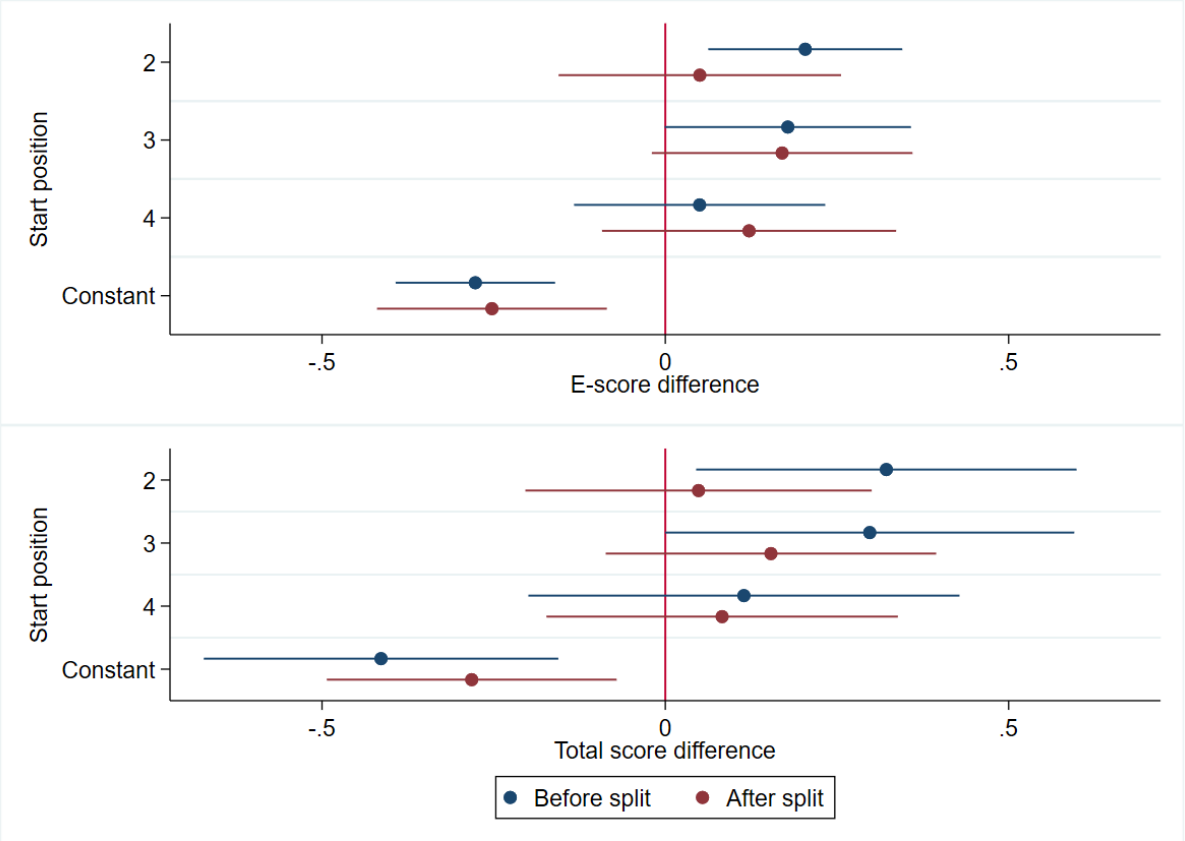

Figure 1. Coefficients of start position before and after split in competition on E-score difference (upper panel) and total score difference (lower panel)

To verify that the similarity is not due to the existence of the touch warmup in a competition, the same analysis is done with respect to the competitions where the touch warmup was not allowed. The competition is again split into two parts, four gymnasts before the hypothetical split and four gymnasts after the hypothetical split. The corresponding figure is shown in Figure 2 (Table A5 and Table A6) and shows a similar trend to Figure 1. The coefficients and intervals are similar for the start positions before and after the hypothetical split. Therefore, the similarity between the gymnasts before and after the split in the competition in Figure 1 cannot be explained by the split in the competition, but can possibly be explained by the overall similarity between all start positions.

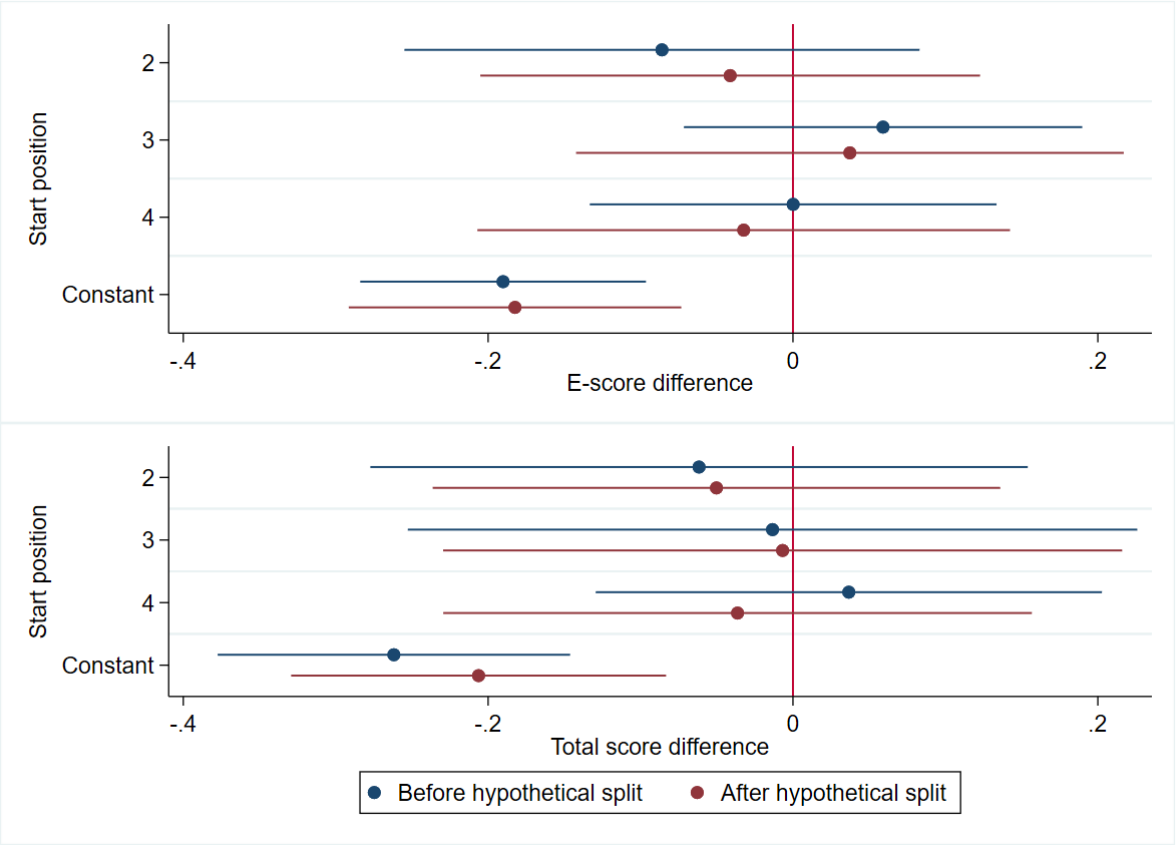

Figure 2. Coefficients of start position before and after a hypothetical split in competition on E-score difference (upper panel) and total score difference (lower panel)

When comparing the coefficients in Figure 1 and 2 (Table A3 to A6), it is remarkable that the magnitude of the estimated coefficients are higher for the subsample of competitions with a touch warmup compared to those without. To further investigate this, a regression in the form of equation 5.1 is estimated for the subsamples with and without a touch warmup. These models are shown in Table A7, and in Figure 3 to compare the coefficients. Strikingly, there does appear to be evidence of overall order bias in the subsample of competitions that did allow for a one-touch warmup. Gymnasts with start positions 2, 3 and 7 score significantly higher compared to the first start position, both when looking at the difference in E-score and total score between the qualification and apparatus final. No such effect is detected when looking at the competitions where no touch warmup was allowed. Randomisation has been checked in Tables A8 and A9 for the subsample of competitions where a one-touch warmup was allowed, and no irregularities have been detected. This means there is no reason to suspect randomisation has failed in this subsample. However, it cannot immediately be concluded that the difference is due to the existence of the one-touch warmup. There could also be potential other differences between the two subsamples, although many of those differences have been adjusted for by using the difference between the qualification score and apparatus final score. This

makes sure that the performances across different competitions in different years are comparable, as the rules are set within each tournament, so the difference between qualification score and apparatus finals score merely shows the difference in performance. A Wald Chi-Squared Test has been performed on each start position coefficient to test whether the coefficients of the two models differ. Only the coefficients for the second start position were found to significantly differ from each other, both for the regression with difference in E-score ($\chi 2$(1, N=206)=6.75, p=0.009<0.05) and difference in total score ($\chi 2$(1, N=206)=4.66, p=0.031<0.05) as the dependent variable.

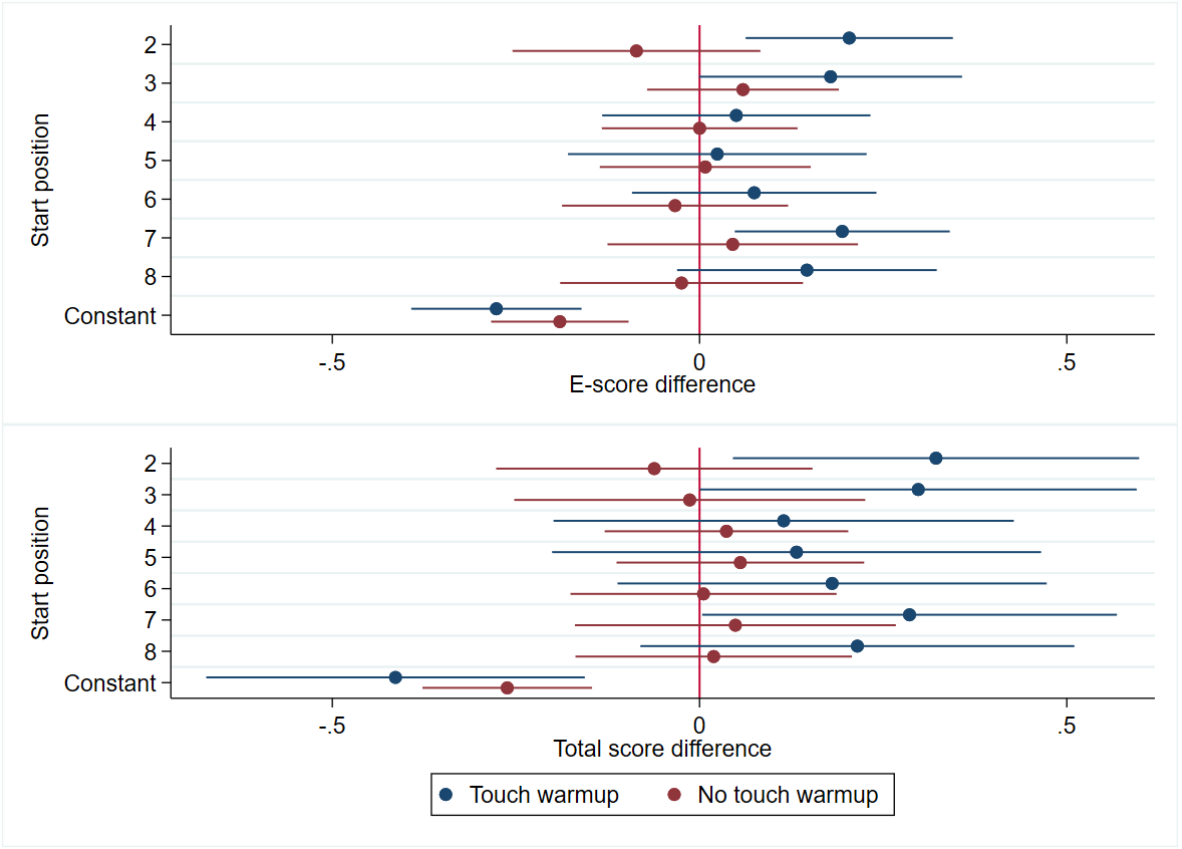

Figure 3. Coefficients of start position for competitions with and without a touch warmup on E-score difference (upper panel) and total score difference (lower panel)

7. Conclusion

This thesis studies the existence of order bias in women's artistic gymnastics apparatus finals by looking at data from international elite tournaments from 2018 to 2023. The sequential nature of the sport and randomised competition order allows to exclude order effects from other determinants of performance. Although often researched in the context of sports, order effects are relevant in many other life events, such as job interviews, product pitches, reviewers grading (oral) examinations, and even the dating scene.

No evidence was found for the existence of overall and sequential order bias when looking at the complete sample, as described by the first two hypotheses. This is in line with the findings of Rotthoff (2020), where no effect for overall and sequential order bias was found. Joustra et al. (2021) and Rotthoff (2015) were able to find evidence for overall order bias, but were unable to find evidence for sequential order bias.

On floor exercise, however, it appears to be the case that gymnasts with start position 2 perform significantly better compared to the reference category of start position 1. No evidence was found for the existence of a driving mechanism of order bias through the quality of the performance of the first competitor. The same holds true for the implementation of the one-touch warmup. It cannot be concluded that the time between warmup and competition has a significant effect on order bias. However, when only looking at the sample of competitions where the one-touch warmup was allowed, there does seem to be evidence for overall order bias. Gymnasts starting at position 2, 3 and 7 perform significantly better than the gymnast with start position 1. It cannot be concluded that this difference is driven by the one-touch warmup.

The analyses performed on the complete sample were unable to confirm the existence of a possible order bias in women's artistic gymnastics. As suggested by Rotthoff (2020), order bias might have disappeared from elite gymnastics competitions over time. However, when only taking the subsample of competitions into account where a one-touch warmup was allowed, some evidence for overall order bias is found. Thus, chances are high that order bias has not completely disappeared from the sport.

This thesis adds to the existing literature by using a more recent and bigger dataset than in previous studies, as well as using a different measure for performance than just the score. By making use of the difference in E-score and total score between the apparatus final and qualification round, individual differences, as well as differences in tournament and competition date, have been controlled for. Additionally, an attempt has been made to uncover mechanisms that drive order bias.

8. Discussion

This thesis attempted to find a relationship between the start position of gymnasts in apparatus finals and their performance compared to the qualification round. By looking at the difference between finals and qualification score, the differences between competitions and gymnasts have been controlled for. The analysis also made use of randomisation in start order, which helped make sure the zero conditional mean assumption holds. Each analysis is done with respect to the difference in E-score and total score between the apparatus final and the qualification round.

The reference category chosen for this research is the first start position, as this has the most intuitive interpretation. The reference category does influence the results obtained from a regression, as in this case we compare if the coefficients of each start position are significantly different from the first start position. The significance of the coefficients would change according to the reference category, however, this has a minimal effect on the conclusions. For example, if the fourth start position had been taken as a reference category, the group with the lowest average performance measures for all apparatus combined, more coefficients would have shown significance, but this does not mean that it can be concluded that order bias exists for many start positions. Rather, it could suggest that there might be a negative effect of starting in fourth place on one's performance.

Although the dataset is larger than in previous research, this thesis looked at international elite gymnastics competitions, as has been done before. One reason for the lack of found evidence of order bias could be the level of competition in the used dataset. The dataset contains only competitions at the highest possible level of the sport. As touched upon earlier, the judges at this level of competition are very carefully selected and have had extensive training to be allowed to judge at these events. Therefore, they might be less prone to bias than for example judges at a district level competition for lower level gymnasts. The same holds true for the gymnasts. Gymnasts with less experience might be more influenced by their start position or the performance of the previous gymnast.

Another possible reason for the lack of found evidence for order bias is that the magnitude of the bias would be very small. As discussed before, even a very small effect could be of importance and distort the competition outcomes. If the effect would be of a small magnitude, it could be the case that a bigger and more powerful dataset would be needed to detect an effect. The data also did not allow for the separation of order bias induced by the judges or the gymnasts. This could potentially be done by conducting large scale experiments. Previous experiments have been conducted to test for order bias induced by the judges, however the sample sizes were relatively small and these experiments usually used video recordings of routines. Further research could make use of new experiments to validate past findings and investigate new mechanisms. A new experiment would ideally make use of

a real competition setting where the surroundings are as close to a real competition as possible. Experiments regarding order bias induced by the gymnasts might include several 'actor' gymnasts, that purposely perform a good or bad routine to set a reference point. Experiments regarding order bias induced by the judges might make use of a gymnastics competition where the gymnasts do not observe the routines and scores of the other gymnasts and are unaware of their own start position.

This thesis tried to add to the existing literature on order bias in women's artistic gymnastics by researching two of the possible mechanisms that influence order bias. No evidence was found to conclude that these mechanisms actually influence order bias. For the effect of waiting time on performance, a more extensive dataset that includes the precise times between warmup and competition could be used. Experiments could also be used to draw conclusions on the effect of a one-touch warmup on performance.

Although the literature describes some effects of order bias and its mechanisms in other contexts, it is unclear to what extent the effects are in line with each other. Although the results of this thesis state that order bias is not as pronounced anymore in the sport of gymnastics, it is hard to conclude anything with respect to order bias in other contexts. Gymnastics is comparable to many other contexts on certain aspects, but the competition aspect is very visible, which possibly makes the judges more aware and cautious of a possible order bias. This is not the case in other contexts, like job interviews and grading exams, where individuals are less aware of their competition. If an individual is less aware of a possible order bias, they might be more likely to be influenced by this. Furthermore, by looking more closely at order bias in other contexts, possible new mechanisms could be discovered and researched. This could improve the external validity of this research.

Order bias in international elite women's artistic gymnastics has reduced over time, which can be concluded by relating this thesis to previous research. To even further reduce the possibility of order bias, also on lower levels of competition and non-sports contexts, it is crucial that competitors and decision makers are made aware of the existence of this phenomenon. More research is needed to uncover mechanisms of this bias, which in turn could be used to target the existence of order bias. The goal is to limit order bias until it is no longer something to worry about, returning fairness to competitions, grades and interviews and returning to the goal of perfection in gymnastics for gymnasts at all start positions.

9. References

Albulescu, P., Macsinga, I., Rusu, A., Sulea, C., Bodnaru, A., & Tulbure, B. T. (2022). " Give me a break!" A systematic review and meta-analysis on the efficacy of micro-breaks for increasing well-being and performance. *Plos one*, *17*(8), e0272460. https://doi.org/10.1371/journal.pone.0272460

Ansorge, C. J., Scheer, J. K., Laub, J., & Howard, J. (1978). Bias in judging women's gymnastics induced by expectations of within-team order. *Research Quarterly. American Alliance for Health, Physical Education and Recreation*, *49*(4), 399-405. https://doi.org/10.1080/10671315.1978.10615552

Antipov, E. A., & Pokryshevskaya, E. B. (2017). Order effects in the results of song contests: Evidence from the Eurovision and the New Wave. *Judgment and Decision Making*, *12*(4), 415-419. https://doi.org/10.1017/S1930297500006288

Bhargava, S. (2008). Perception is relative: Sequential contrasts in the field. *[Working paper].*

Bhargava, S., & Fisman, R. (2014). Contrast effects in sequential decisions: Evidence from speed dating. *Review of Economics and Statistics*, *96*(3), 444-457. https://doi.org/10.1162/REST_a_00416

De Bruin, W. B. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta psychologica*, *118*(3), 245-260. https://doi.org/10.1016/j.actpsy.2004.08.005

De Bruin, W. B. (2006). Save the last dance II: Unwanted serial position effects in figure skating judgments. *Acta Psychologica*, *123*(3), 299-311. https://doi.org/10.1016/j.actpsy.2006.01.009

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. https://doi.org/10.1037/1076-898X.12.3.166

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, *108*(17), 6889-6892. https://doi.org/10.1073/pnas.1018033108

Deck, A., Deck, C., & Zhu, Z. (2014). Decision making in a sequential game: The case of pitting in NASCAR. *Journal of Sports Economics*, *15*(2), 132-149. https://doi.org/10.1177/1527002512443828

Deutscher, C., Neuberg, L. & Thiem, S. (2022). Who's Afraid of the GOATs? - Shadow Effects of Tennis Superstars. *[Working paper]*. http://dx.doi.org/10.2139/ssrn.4207867

Elmore, R., & Urbaczewski, A. (2021). Loss aversion in professional golf. *Journal of Sports Economics*, *22*(2), 202-217. https://doi.org/10.1177/1527002520967403

Fédération Internationale de Gymnastique. (2021, September*). FIG Council approves a touch warm-up for Apparatus Finals. https://www.gymnastics.sport/site/news/displaynews.php?urlNews=3312532*

Fédération Internationale de Gymnastique. (2023). Technical regulations. Lausanne: Fédération Internationale de Gymnastique.

Fédération Internationale de Gymnastique - *Results*. (2023). https://www.gymnastics.sport/site/events/searchresults.php#filter

Feenberg, D., Ganguli, I., Gaule, P., & Gruber, J. (2017). It's good to be first: Order bias in reading and citing NBER working papers. Review of Economics and Statistics, 99(1), 32-39. https://doi.org/10.1162/REST_a_00607

Findlay, L. C., & Ste-Marie, D. M. (2004). A reputation bias in figure skating judging. *Journal of Sport and Exercise Psychology*, *26*(1), 154-166.1 https://doi.org/10.1123/jsep.26.1.154

Goldbach, C., Sickmann, J., & Pitz, T. (2022). Sequential decision bias–evidence from grading exams. *Applied Economics*, *54*(32), 3727-3739. https://doi.org/10.1080/00036846.2021.1976390

GYMmedia.com. (2023). https://www.gymmedia.com/

Gymnastics Results. (2023). https://gymnasticsresults.com/

Heiniger, S., & Mercier, H. (2021). Judging the judges: evaluating the accuracy and national bias of international gymnastics judges. *Journal of Quantitative Analysis in Sports*, *17*(4), 289-305. https://doi.org/10.1515/jqas-2019-0113

Hill, B. (2014). The heat is on: Tournament structure, peer effects, and performance. Journal of Sports Economics, 15(4), 315-337. https://doi.org/10.1177/1527002512461156

Holmes, D. S., & Berkowitz, L. (1961). Some contrast effects in social perception. *The Journal of Abnormal and Social Psychology*, *62*(1), 150. https://doi.org/10.1037/h0042168

Joustra, S. J., Koning, R. H., & Krumer, A. (2021). Order effects in elite gymnastics. De Economist, 169(1), 21-35. https://doi.org/10.1007/s10645-020-09371-0

Leskošek, B., Čuk, I., Pajek, J., Forbes, W., & Bučar-Pajek, M. (2012). Bias of judging in men's artistic gymnastics at the European Championship 2011. *Biology of Sport*, *29*(2), 107-113. http://dx.doi.org/10.5604/20831862.988884

Meissner, L., Rai, A., & Rotthoff, K. W. (2021). The superstar effect in gymnastics. *Applied Economics*, *53*(24), 2791-2798. https://doi.org/10.1080/00036846.2020.1869170

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. https://doi.org/10.1111/ecin.12074

Orazbayev, S. (2017). Sequential order as an extraneous factor in editorial decision. Scientometrics, 113(3), 1573-1592. https://doi.org/10.1007/s11192-017-2531-7

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*(2), 186-198. https://doi.org/10.1016/j.jebo.2009.08.012

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 521. https://doi.org/10.1037/0278-7393.14.3.521

Rotthoff, K. W. (2015). (Not finding a) sequential order bias in elite level gymnastics. *Southern Economic Journal*, *81*(3), 724-741. https://doi.org/10.4284/0038-4038-2013.052

Rotthoff, K. W. (2020). Revisiting difficulty bias, and other forms of bias, in elite level gymnastics. *Journal of Sports Analytics*, *6*(1), 1-11. https://doi.org/10.3233/JSA-200272

Rowe, P. M. (1967). Order effects in assessment decisions. *Journal of Applied Psychology*, *51*(2), 170. https://doi.org/10.1037/h0024346

Scheer, J. K., & Ansorge, C. J. (1975). Effects of naturally induced judges' expectations on the ratings of physical performances. *Research Quarterly. American Alliance for Health, Physical Education and Recreation*, *46*(4), 463-470. https://doi.org/10.1080/10671315.1975.10616704

Scheer, J. K., Ansorge, C. J., & Howard, J. (1983). Judging bias induced by viewing contrived videotapes: A function of selected psychological variables. *Journal of Sport and Exercise Psychology*, *5*(4), 427-437. https://doi.org/10.1123/jsp.5.4.427

Ste-Marie, D. (2003). Memory biases in gymnastic judging: Differential effects of surface feature changes. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *17*(6), 733-751. https://doi.org/10.1002/acp.897

The Gymternet. (2023). The Gymternet. https://thegymter.net/

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. https://doi.org/10.2466/pms.1977.44.1.295

10. Appendix

Table A1. Descriptive statistics of apparatus finals scores, sorted by apparatus

	Obs.	Mean	Std. Dev.	Min.	Max.
Vault					
E-score	400	8.670	0.518	3.783	9.499
D-score	400	4.744	0.609	2.500	5.900
Total score	400	13.382	0.939	6.367	15.399
Uneven bars					
E-score	424	7.608	0.817	4.050	8.900
D-score	424	5.413	0.718	1.500	6.700
Total score	424	13.009	1.356	5.133	15.300
Balance Beam					
E-score	434	7.308	0.773	2.525	8.666
D-score	434	5.160	0.503	2.900	6.600
Total score	434	12.461	1.073	7.325	15.266
Floor exercise					
E-score	426	7.743	0.487	5.366	9.566
D-score	426	5.000	0.526	0.400	6.700
Total	426	12.662	1.053	2.300	15.133

Note: The E-score starts from 10 points from which points are deducted for each mistake. The D-score starts from 0 points and points are added for each composition requirement fulfilled, each element performed and each element connected. The total score is the sum of the D-score and E-score, minus a possible penalty. A penalty is given for possible mistakes that are not included in the E-score, such as exceeding the time limit or stepping out of the permitted areas during a routine. The score differences are calculated by taking the difference between respectively the score or rank in the final (f) and qualification (q) round.

Table A2a. Descriptive statistics of vault qualification performance, sorted by start position in finals

	Obs.	Mean	Std. Dev.	Min.	Max.
1					
E-score qualification	50	8.741	0.178	8.200	9.133
D-score qualification	50	4.764	0.518	3.800	5.900
Total score qualification	50	13.492	0.590	12.283	14.866
2					
E-score qualification	50	8.784	0.243	8.150	9.450
D-score qualification	50	4.781	0.634	3.700	5.900
Total score qualification	50	13.539	0.680	12.199	15.199
3					
E-score qualification	50	8.755	0.247	8.267	9.566
D-score qualification	50	4.639	0.588	3.600	6.100
Total score qualification	50	13.381	0.773	11.925	15.666
4					
E-score qualification	50	8.779	0.293	7.725	9.466
D-score qualification	50	4.784	0.555	3.600	5.900
Total score qualification	50	13.548	0.707	12.167	15.166
5					
E-score qualification	50	8.738	0.234	7.966	9.416
D-score qualification	50	4.719	0.590	3.600	5.700
Total score qualification	50	13.422	0.709	12.216	14.833
6					
E-score qualification	50	8.775	0.300	7.600	9.350
D-score qualification	50	4.853	0.556	3.400	5.800
Total score qualification	50	13.606	0.745	11.500	15.050
7					
E-score qualification	50	8.702	0.244	7.950	9.249
D-score qualification	50	4.691	0.577	3.800	5.900
Total score qualification	50	13.368	0.627	12.349	14.633
8					
E-score qualification	50	8.741	0.240	8.050	9.300
D-score qualification	50	4.739	0.551	3.500	5.900
Total score qualification	50	13.454	0.694	11.599	15.200

Table A2b. Descriptive statistics of uneven bars qualification performance, sorted by start position in finals

	Obs.	Mean	Std. Dev.	Min.	Max.
1					
E-score qualification	53	7.818	0.491	6.500	8.750
D-score qualification	53	5.472	0.570	4.200	6.600
Total score qualification	53	13.29	0.951	10.700	15.200
2					
E-score qualification	53	7.869	0.571	6.000	8.800
D-score qualification	53	5.321	0.861	2.100	6.500
Total score qualification	53	13.189	1.193	10.033	15.166
3					
E-score qualification	53	7.770	0.658	5.600	8.666
D-score qualification	53	5.409	0.713	3.900	6.700
Total score qualification	53	13.180	1.201	9.500	15.366
4					
E-score qualification	53	7.903	0.614	6.066	8.833
D-score qualification	53	5.575	0.664	4.200	6.700
Total score qualification	53	13.478	1.152	10.766	15.233
5					
E-score qualification	53	7.879	0.475	6.150	8.633
D-score qualification	53	5.466	0.631	4.300	6.500
Total score qualification	53	13.345	0.921	11.150	15.066
6					
E-score qualification	53	7.822	0.547	5.950	8.766
D-score qualification	53	5.445	0.676	3.600	6.400
Total score qualification	53	13.262	1.061	10.950	14.766
7					
E-score qualification	53	7.790	0.656	5.100	8.800
D-score qualification	53	5.577	0.584	4.200	6.400
Total score qualification	53	13.367	1.076	9.500	15.016
8					
E-score qualification	53	7.814	0.580	6.200	8.933
D-score qualification	53	5.391	0.712	3.500	6.500
Total score qualification	53	13.205	1.104	10.800	15.233

Table A2c. Descriptive statistics of balance beam qualification performance, sorted by start position in finals

	Obs.	Mean	Std. Dev.	Min.	Max.
1					
E-score qualification	54	7.598	0.517	6.350	8.700
D-score qualification	54	5.181	0.489	4.100	6.400
Total score qualification	54	12.774	0.828	11.150	15.000
2					
E-score qualification	54	7.519	0.450	6.200	8.466
D-score qualification	54	5.189	0.418	4.300	6.20
Total score qualification	54	12.704	0.686	10.950	14.566
3					
E-score qualification	54	7.567	0.605	4.500	8.533
D-score qualification	54	5.372	0.540	4.000	7.200
Total score qualification	54	12.936	0.743	11.050	14.633
4					
E-score qualification	54	7.660	0.488	6.500	8.566
D-score qualification	54	5.269	0.410	4.300	6.300
Total score qualification	54	12.929	0.750	11.433	14.800
5					
E-score qualification	54	7.623	0.474	5.900	8.366
D-score qualification	54	5.174	0.485	4.300	6.200
Total score qualification	54	12.791	0.709	11.100	14.300
6					
E-score qualification	54	7.658	0.437	6.300	8.466
D-score qualification	54	5.217	0.393	4.500	6.400
Total score qualification	54	12.870	0.681	11.200	14.466
7					
E-score qualification	54	7.540	0.640	5.050	8.633
D-score qualification	54	5.319	0.429	4.500	6.500
Total score qualification	54	12.856	0.818	10.650	14.850
8					
E-score qualification	54	7.534	0.625	4.800	8.650
D-score qualification	54	5.248	0.599	3.800	7.000
Total score qualification	54	12.775	0.951	10.100	14.933
9					
E-score qualification	2	7.800	0.707	7.300	8.300
D-score qualification	2	5.100	0.424	4.800	5.400
Total score qualification	2	12.900	1.131	12.100	13.700

Table A2d. Descriptive statistics of floor exercise qualification performance, sorted by start position in finals

	Obs.	Mean	Std. Dev.	Min.	Max.
1					
E-score qualification	53	7.887	0.298	7.300	8.500
D-score qualification	53	5.002	0.412	4.200	6.100
Total score qualification	53	12.853	0.601	11.650	14.200
2					
E-score qualification	53	7.746	0.371	6.600	8.400
D-score qualification	53	4.977	0.442	4.100	6.200
Total score qualification	53	12.682	0.677	11.400	14.300
3					
E-score qualification	53	7.862	0.325	7.100	8.533
D-score qualification	53	4.989	0.367	4.100	5.800
Total score qualification	53	12.834	0.588	11.600	14.033
4					
E-score qualification	53	7.899	0.362	6.800	8.500
D-score qualification	53	5.017	0.415	4.300	5.800
Total score qualification	53	12.874	0.673	11.100	14.200
5					
E-score qualification	53	7.881	0.365	6.500	8.500
D-score qualification	53	5.074	0.516	3.700	6.300
Total score qualification	53	12.926	0.767	10.200	14.166
6					
E-score qualification	53	7.889	0.313	7.050	8.650
D-score qualification	53	5.021	0.492	4.200	6.700
Total score qualification	53	12.878	0.714	11.500	15.333
7					
E-score qualification	53	7.841	0.331	6.800	8.450
D-score qualification	53	4.979	0.404	4.200	5.900
Total score qualification	53	12.798	0.641	11.200	14.066
8					
E-score qualification	53	7.946	0.322	7.166	8.433
D-score qualification	53	5.136	0.504	4.300	6.600
Total score qualification	53	13.037	0.740	11.500	14.833
9					
E-score qualification	2	7.899	0.094	7.833	7.966
D-score qualification	2	5.200	0.283	5.000	5.400
Total score qualification	2	13.05	0.306	12.833	13.266

Table A3. Regression results split touch warmup on E-score difference

	(1)	(2)
VARIABLES	E-score difference	E-score difference
Split	Before	After
Start position		
2	0.204***	0.0503
	(0.0718)	(0.105)
3	0.178*	0.170*
	(0.0912)	(0.0965)
4	0.0500	0.122
	(0.0930)	(0.109)
Constant	-0.277***	-0.252***
	(0.0590)	(0.0851)
Observations	332	332
R-squared	0.023	0.012

Note: The score differences are calculated by taking the difference between the score in the final and qualification round. The standard errors are shown in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table A4. Regression results split touch warmup on total score difference

	(1)	(2)
VARIABLES	Total score difference	Total score difference
Split	Before	After
Start position		
2	0.322**	0.0484
	(0.141)	(0.128)
3	0.298*	0.154
	(0.151)	(0.122)
4	0.115	0.0828
	(0.160)	(0.130)
Constant	-0.414***	-0.282***
	(0.131)	(0.107)
Observations	332	332
R-squared	0.025	0.006

Note: The score differences are calculated by taking the difference between the score in the final and qualification round. The standard errors are shown in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table A5. Regression results hypothetical split touch warmup on E-score difference

	(1)	(2)
VARIABLES	E-score difference	E-score difference
Hypothetical split	Before	After
Start position		
2	-0.0859	-0.0412
	(0.0860)	(0.0834)
3	0.0591	0.0373
	(0.0665)	(0.0914)
4	0.000131	-0.0323
	(0.0679)	(0.0889)
Constant	-0.190***	-0.182***
	(0.0477)	(0.0555)
Observations	444	444
R-squared	0.008	0.002

Note: The score differences are calculated by taking the difference between the score in the final and qualification round. The standard errors are shown in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table A6. Regression results hypothetical split touch warmup on total score difference

	(1)	(2)
VARIABLES	Total score difference	Total score difference
Hypothetical split	Before	After
Start position		
2	-0.0616	-0.0502
	(0.110)	(0.0947)
3	-0.0134	-0.00678
	(0.122)	(0.113)
4	0.0366	-0.0364
	(0.0845)	(0.0982)
Constant	-0.262***	-0.206***
	(0.0588)	(0.0626)
Observations	444	444
R-squared	0.002	0.001

Note: The score differences are calculated by taking the difference between the score in the final and qualification round. The standard errors are shown in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table A7. Regression results of start position overall order bias on E-score difference and total score difference, competitions with and without a one-touch warmup

VARIABLES	(1) E-score difference	(2) E-score difference	(3) Total score difference	(4) Total score difference
Touch	Yes	No	Yes	No
Start position				
2	0.204***	-0.0859	0.322**	-0.0616
	(0.0718)	(0.0860)	(0.141)	(0.110)
3	0.178*	0.0591	0.298**	-0.0134
	(0.0912)	(0.0665)	(0.151)	(0.122)
4	0.0500	0.000131	0.115	0.0366
	(0.0930)	(0.0679)	(0.160)	(0.0845)
5	0.0241	0.00782	0.132	0.0556
	(0.104)	(0.0732)	(0.170)	(0.0859)
6	0.0744	-0.0334	0.180	0.00538
	(0.0847)	(0.0784)	(0.149)	(0.0923)
7	0.194***	0.0452	0.286**	0.0488
	(0.0746)	(0.0869)	(0.144)	(0.111)
8	0.146	-0.0245	0.215	0.0192
	(0.0900)	(0.0842)	(0.150)	(0.0959)
Constant	-0.277***	-0.190***	-0.414***	-0.262***
	(0.0590)	(0.0477)	(0.131)	(0.0588)
Observations	664	888	664	888
R-squared	0.017	0.005	0.017	0.002

Note: The score differences are calculated by taking the difference between the score in the final and qualification round. The standard errors are shown in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Table A8. Correlation table of qualifications performance and start position in the case of a one-touch warmup, sorted by apparatus

VARIABLES	Start position (VT)	Start position (UB)	Start position (BB)	Start position (FX)
E-score qualification	-0.075	-0.007	-0.001	0.082
	(0.346)	(0.929)	(0.993)	(0.290)
D-score qualification	-0.021	0.116	0.126	0.071
	(0.797)	(0.135)	(0.104)	(0.358)
Total score qualification	-0.052	0.073	0.089	0.093
	(0.513)	(0.348)	(0.251)	(0.233)

Note: Each cell shows the correlation between a qualification score component and the start position for one of the four apparatus. The standard errors are shown in parentheses.
**** p<0.01, ** p<0.05, * p<0.1*

Table A9. Kruskal-Wallis test of systematic differences in scores in the case of a one-touch warmup, sorted by apparatus

VARIABLES	4 - Vault	3 - Uneven bars	1 - Balance beam	2 - Floor exercise
E-score qualification	$\chi2$(7, N=160) = 4.585, p=0.710	$\chi2$(7, N=168) = 1.101, p=0.993	$\chi2$(7, N=168) = 2.165, p=0.950	$\chi2$(7, N=168) = 6.657, p=0.465
D-score qualification	$\chi2$(7, N=160) = 7.001, p=0.429	$\chi2$(7, N=168) = 6.038, p=0.535	$\chi2$(7, N=168) = 6.236, p=0.513	$\chi2$(7, N=168) = 10.638, p=0.155
Total score qualification	$\chi2$(7, N=160) = 6.452, p=0.488	$\chi2$(7, N=168) = 2.844, p=0.899	$\chi2$(7, N=168) = 1.660, p=0.976	$\chi2$(7, N=168) = 10.060, p=0.185

Note: Each cell shows a separate Kruskal-Wallis test for a qualification score component for one of the four apparatus.