Men's Gymnastics: The Impact of Missed Routines on Athlete Performance

Disha Dasgupta, Kimberly Tran, Joey Ringer, Niki Saelou

Department of Management Science and Engineering
(disha01, kimdroid, josephr5, nsaelou)@stanford.edu
Stanford University - MS&E 125

Abstract

In gymnastics, team scores are based off of cumulative individual performances within six different events. While athletes perform routines as individuals, this research looks at how an individual's performance impacts other athletes' scores. Specifically, it explores the effects of missed routines on the next athlete's score. Looking at scoring data from the Stanford Men's Gymnastics team between 2014–2019, we analyzed the impact of missed routines with a linear regression model and propensity score matching analyses. Based on the overall impact of missed routines, we aim to determine the best lineup orders to minimize score reductions and maximize overall team score.

1 Background

Gymnastics is a sport that began in ancient Greece more than 2000 years ago. It was included in the Greek Olympics, and was incorporated into people's everyday lives because of the muscles it developed for close combat. Gymnastics was introduced to the United States in the late 1800's as the military used it for training, and from there it gained popularity within the public ("Gymnastics History"). Men's gymnastics became recognized as a college sport by the NCAA in 1938. Now, how does college men's gymnastics work? How does it differ from the Olympics? What events are included in college men's gymnastics? How does the scoring system function?

To answer these questions, the basic terminology and fundamentals of men's gymnastics need to be understood. Men's gymnastics is comprised of six apparatuses which are otherwise known as events. These events include floor, pommel horse, rings, vault parallel bars, and high bar. Men compete on these six events by putting together routines composed of a series of ten skills. Every skill has a difficulty value associated with it that ranges from the letters A to H. An A value skill is worth 0.1, a B value skill is worth 0.2, and so on with an H being worth 0.8. The final score a gymnast receives is made up of two different numbers. First is the start value which is the summation of skill values performed in a routine (e.g. 5.5). Therefore, more difficult routines have the potential to score higher. The second number that is used in determining the final score of a routine is the execution score. The execution score is determined by a judge and starts from a 10. A judge takes deductions ranging from -0.1 for a small form break (e.g. knees bent, feet apart, toes not pointed) up to -0.5 for a large form break. The maximum deduction possible is from a fall off the equipment or from not landing on your feet which is a -1.0 point deduction from the final execution score (e.g. 8.5). To get a gymnast's final score, people add the start value and execution score together (e.g. 5.5 + 8.5 = 14.0).

College men's gymnastics consists of a team of student athletes that train in all six events, specialize in few, or train on just one. The goal for a team in college men's gymnastics is to put out their best group of fifteen athletes to achieve the greatest team score. From this group of fifteen, five athletes are assigned each of the six events, with some athletes competing in more than one event per competition. The order in which the five athletes compete for each event is called the lineup. The exact lineup of gymnasts competing on an apparatus is determined by the coach with the intention of trying to achieve the highest event total composed of the five routine scores added together. The final team score consists of the sum of the six event score totals. A lot of strategy goes into selecting the fifteen representatives of the entire men's gymnastics team to fill in the five-athlete lineups for all six events in order to allow for the greatest team score possible in a competition against other colleges.

2 Literature Review

Most previous research has focused on the possible biases of judges, but not specifically in relation to lineup order. For instance, researchers at the University of Ljubljana determined that while judges' ratings were typically

consistent, they were not always valid due to systematic bias (Pajek et al. 50).

Diane M. Ste-Marie, Sheri M. Valiquette, and Gail Taylor, researchers at the University of Ottawa, solidified this point through their research on memory-influenced biases in gymnastics judging. They determined that no matter how long the delay since a certain previous memory (i.e. seeing a gymnast warm up before the meet or remembering a gymnast's performance from a previous meet), their scorings were influenced by these memories as compared to other judges who had no previous memories associated with a certain performance or gymnast (Ste-Marie et al. 420).

The findings from the aforementioned studies show that judges can be biased in multiple ways, indicating the necessity of understanding how these biases specifically come into play when it comes to lineup order. Some research on lineup order has been conducted by Henning Plesser, a researcher at the University of Heidelberg; competitor routines were filmed and shown to judges, where each judge was told that the routine was either the first or last one in the lineup (regardless of whether that was true or not). Plessner found there was in fact a difference in the average scores based on the competitor's stated place in the lineup (Plessner 137).

3 Motivation

Stanford's varsity NCAA men's gymnastics program has consistently been one of the top colleges qualifying for the NCAA Championships during the past two decades. Furthermore, they have won a total of five NCAA Championship titles since Stanford started the program in the 1950s. Their last championship was won in 2011 and they are going for their sixth national title this season.

One of the members conducting this final research project, Joey Ringer, is a junior on Stanford's men's gymnastics team. From his experiences on the team, we understand some of Stanford's strategy in deciding the lineup order for each event. The general trend that Stanford follows when picking a lineup is to have the final scores of each athlete ascend throughout the lineup. Therefore, the lineup is typically organized from lowest start value to greatest start value because of the potential to score higher with more difficult routines. However, choosing the order isn't necessarily this simple, as there are many other variables for Stanford to consider as well.

Stanford's head coach, Thom Glielmi, considers the first and last positions of a lineup to be the most important. The first person sets the tone for the rest of the lineup, so Coach Glielmi generally puts someone who he knows will hit their routine in this spot. A hit routine on an event is a routine with no major form breaks, while a missed routine is a routine with a major form break or fall. The last athlete in an event has the job of closing out the lineup with a high note and is supposed to get the highest score for that event. According to Coach Glielmi, to be able to capitalize on the last person going in the lineup to get the highest score possible, it's essential that everyone previously in the lineup hits their routine. Coach Glielmi believes this will lead to the greatest event total possible. Thus, we want to know how detrimental a missed routine is to the event total in order to determine a more effective way to organize the lineup on an event to maximize the final team score.

As mentioned before, previous research has only focused on the first and last routines in a lineup, and hasn't consider the relationship between other routines within the lineup. As a result, all of this has led us to our research question: how does a missed routine within a gymnastics lineup affect the performance of later team members in the lineup?

4 Data Collection and Modification

At the end of each meet, the execution scores for all competitors within the 6 events are listed in a final score sheet. In addition to the execution scores, the final score sheet also lists the order of competitors per event, the running total team scores after each event, and the overall team score for every team at the end of the meet. Scoring sheets for all the gymnastics meets are available publicly (sorted by year) — these sheets will be our main data sources.

Scoring Sheet/Dataset Example can be found at this link: **https://gostanford.com/documents/2018/1/25//UIC_Iowa_Stanford_results.pdf**

For our research, we decided to use an 8.25 as the execution score cutoff threshold. Any execution score above an 8.25 is considered a hit, and anything below an 8.25 is considered a miss.

5 Data Exploration

Before conducting any analyses, we explored the data to look for any preliminary trends. In our case, we decided to look at the average execution scores of the athletes who competed before and after the first person in the lineup who missed their routine. When running this preliminary analysis, we expected that the gymnasts competing before the first missed routine would average higher execution scores than the gymnasts competing after the first missed routine.

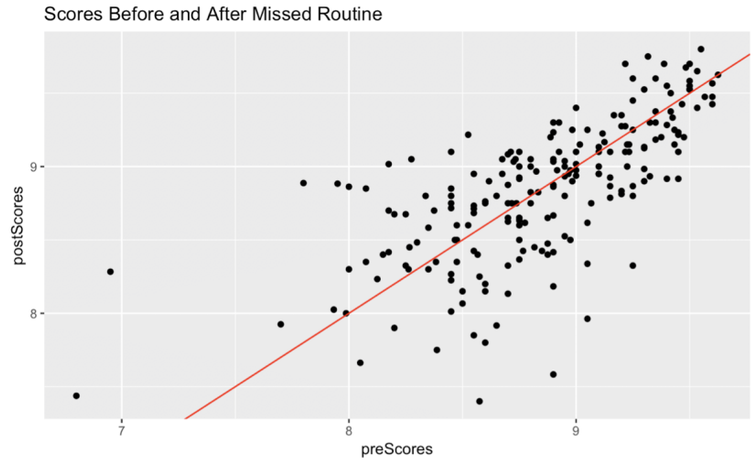

Figure 1: **Average execution scores from before/after first missed routine**

Figure 1 shows a comparison of the average execution scores from before and after the first missed routine. Based on our expectations, we believed

that most of the points would lie under the x = y line (the red line). However, we see that the points are pretty evenly distributed across the x = y line, which motivated our following analyses to delve deeper into the impact of a missed routine on the routine conducted directly after it.

6 Models for Analysis

We initially ran a linear regression analysis to determine if there was a linear relationship between an athlete missing their routine and the next athlete's average performance. As this linear regression did not account for covariates, we further conducted a propensity score matching analysis to determine the effect of potential covariates before rerunning our model.

6.1 Linear Regression

We first ran a linear regression model (using the glm model function in R). We used the event, athlete, and an indicator variable z ("0" if the previous athlete didn't miss their routine and "1" if the previous athlete missed their routine) as predictors, and tried to predict the average execution score for each competitor.

The initial linear regression model did not account for other covariates, so after running a propensity score matching analysis (see section 6.2 below), we ran the linear regression model again on the matched data to see if the linear regression could better predict the average execution scores based on whether the previous competitor missed their routine or not.

6.2 Propensity Score Matching

Although our initial linear regression model took into account the specific event type and athlete associated with each routine, as well as whether the previous competitor in the event lineup hit or missed their routine (denoted by the indicator variable z), the model fails to address several other covariates which may also be impacting the routine scores and contributing to selection bias. For instance, apart from the performance outcome of the previous competitor, the execution score for any given routine may also be influenced by covariates like the routine's difficulty score, the athlete's skill level in the event type, or even the number of other routines the athlete has had to perform prior to the one at hand.

As a result, in order to strengthen our investigation on any potential causal effects of z, we implemented the MatchIt package within R in order to conduct a propensity score matching analysis. In particular, we utilized the "Nearest Neighbor" matching technique, and executed one-on-one matching between our control (z = 0) and treatment (z = 1) cases on the basis of difficulty score and athlete skill level. The "Nearest Neighbor" technique matches the treated and control units which are closest to each other in terms of distance measured (as opposed to minimizing average absolute distances across all matched pairs or relying on genetic search algorithms).

7 Results

Using both the linear regression and propensity score matching analyses, we achieved the following results:

7.1 Linear Regression on Un-Matched Data

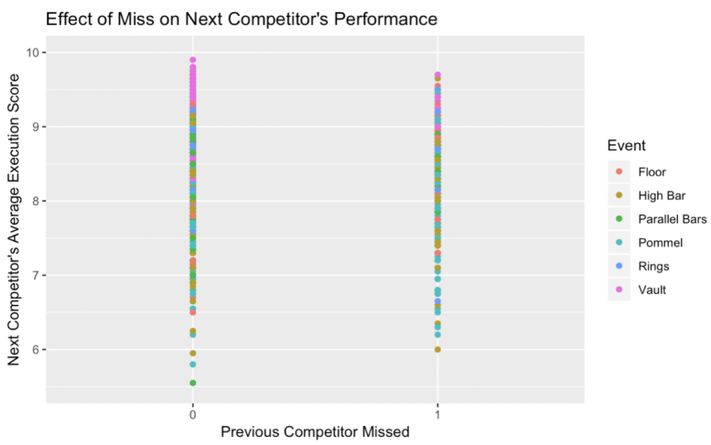

Figure 2: **Effect of previous competitor missing on next competitor's performance (unmatched)**

Figure 2 shows the effect of someone missing their routine on the next competitor's performance. On the x-axis, a score of "0" indicates that the person before a certain athlete did not miss their routine, and a score of "1" indicates that the person before an athlete did miss their routine. The y-axis shows the execution score for the competitor's routine. Figure 3 shows the same effect except on the predicted scores determined from our linear regression model.

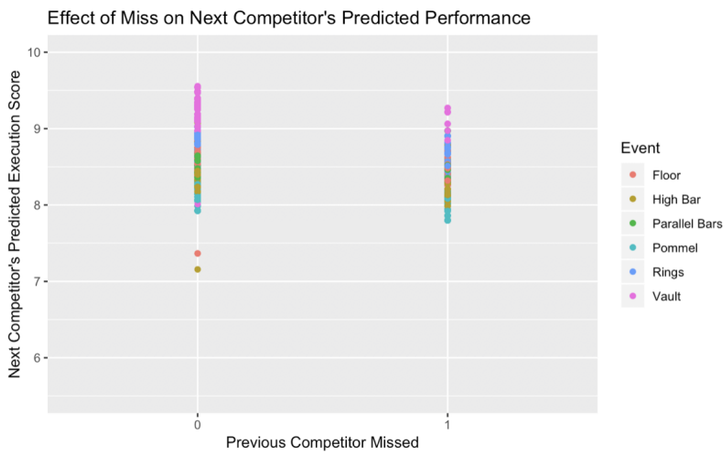

Figure 3: **Effect of previous competitor missing on next competitor's predicted performance (unmatched)**

To clarify the results, we calculated the mean effect of someone missing their routine on the next competitor's performance/predicted performance for each of the 6 events. From figures 4 and 5, we see that for all the events, if the competitor before a certain athlete missed their routine, then the next athlete would be more likely to perform comparatively worse.

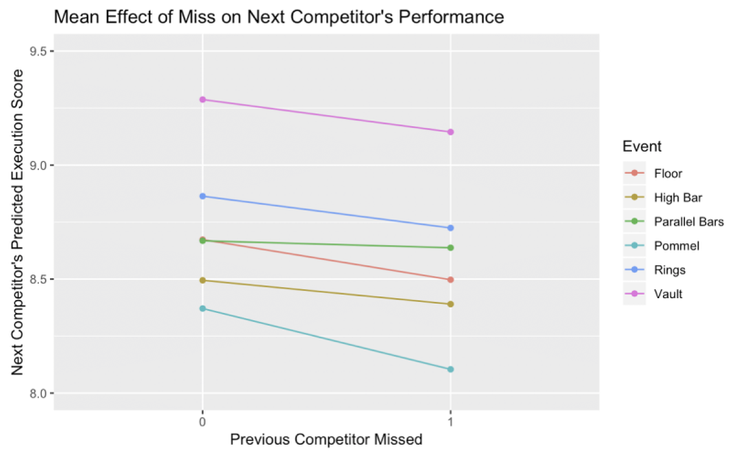

Figure 4: **Mean effect of previous competitor missing on next competitor's performance (unmatched)**

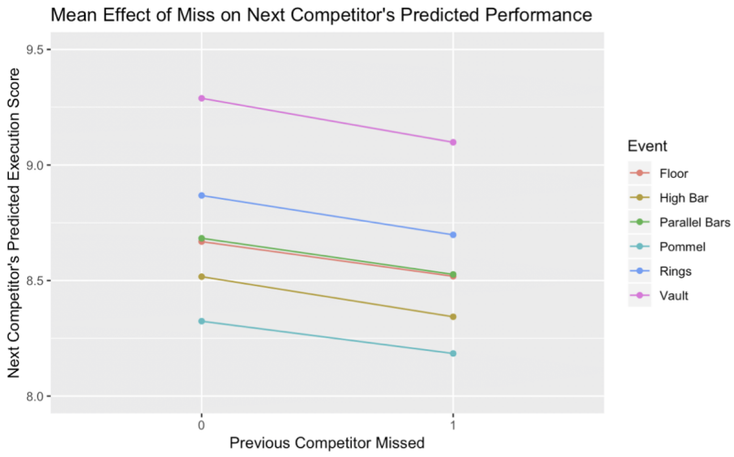

Figure 5: **Mean effect of previous competitor missing on next competitor's predicted performance (unmatched)**

7.2 Propensity Score Matching

After matching our control and treatment cases, we created the following jitter plot to better visualize the results of our propensity score analysis:

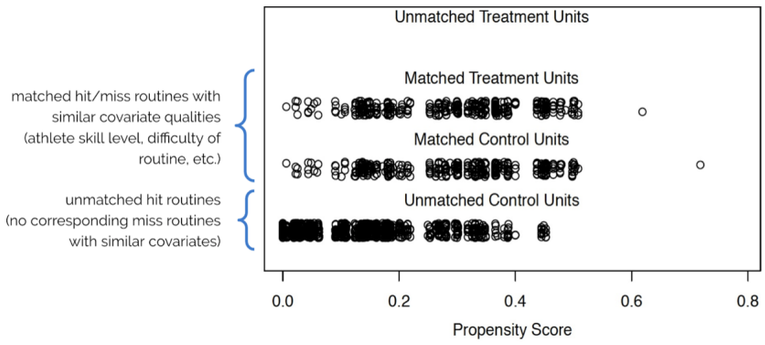

Figure 6: **Propensity scores for each of the matched/unmatched control and treatment groups**

Each individual circle represents the propensity score of a given treatment

or control case, and we can see from the results that the matching process worked well with our particular data set. Although there were a large number of unmatched control cases (z = 0; ie. previous competitor HIT their routine), this was simply a result of the few treatment cases (z = 1; ie. previous competitor MISSED their routine) that we had within our dataset to begin with. Nonetheless, the matching process resulted in a relatively large pool of matched cases, which we then outputted as a data subset for further analyses.

7.3 Linear Regression on Matched Data

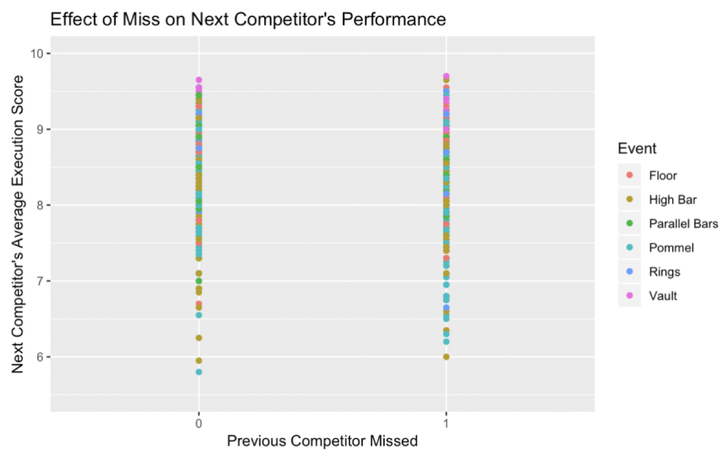

Figure 7: **Effect of previous competitor missing on next competitor's performance (matched)**

After exporting the matched cases from our propensity score matching analysis, we reran our linear regression model on the new, matched dataset instead. As with Figure 3 from section (7.1), Figure 7 displays the effect of a missed routine on the following competitor's performance within an event lineup.

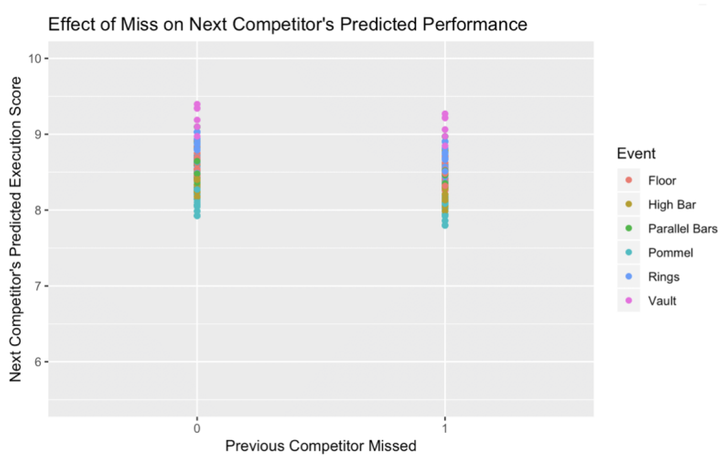

Figure 8: **Effect of previous competitor missing on next competitor's predicted performance (matched)**

Again, the x-axis represents the values of our indicator variable z, with a "0" indicating that the previous competitor hit their routine and a "1" indicating that the previous competitor missed their routine. The y-axis, on

the other hand, indicates the execution score of each routine (where a routine is represented by an individual dot in the plot) out of a perfect score of 10. Figure 8 is characterized by the same axes, but demonstrates the predicted effect of a missed routine determined by our linear regression model after rerunning it on the matched data subset.

Once again, we calculated the mean scores presented in Figures 7 and 8 (by event) in order to present our findings in a more intuitive manner. As shown in Figure 9, a missed routine negatively impacted the scores of following competitors in the lineup for every event except parallel bars, although the predictions of our linear regression model in Figure 10 still indicate that a missed routine would impact following routine scores for all events including parallel bars.

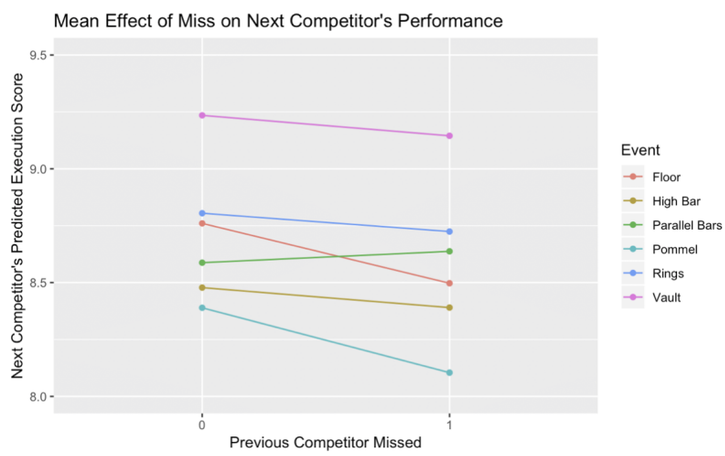

Figure 9: **Mean effect of previous competitor missing on next competitor's performance (matched)**

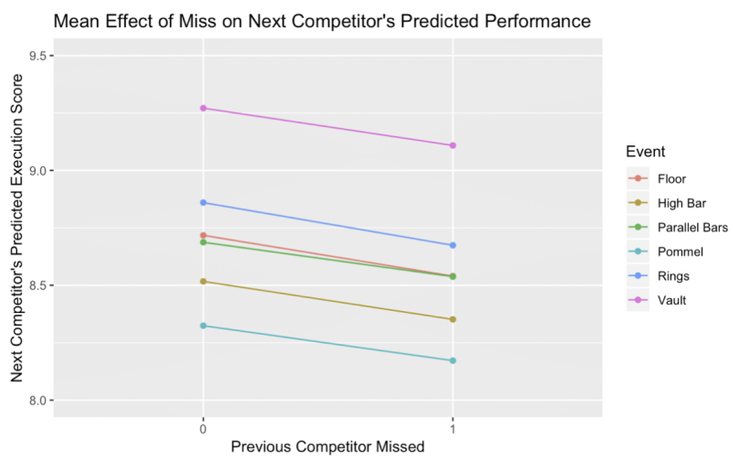

Figure 10: **Mean effect of previous competitor missing on next competitor's predicted performance (matched)**

In other words, even after taking potential covariates into consideration via our propensity score matching analysis, our findings still indicate a correlation between a missed routine and the scores of the following routine.

8 Conclusion

Performing an analysis on the Stanford Men's Gymnastics data over the past five years revealed that missed routines result in minor impacts on the execution scores of routines directly following the miss. This trend of lower scores implicates that missed routines create a trend of score reductions either as an athlete's confidence is affected by watching teammates underperform or as judges create a negative bias towards a team after witnessing a missed routine from one of their athletes. While this may be the case, the differences in performance scores can also be attributed to other factors, such as routine difficulty and athlete skill level. Overall, the observation of small score reductions can implicate large cumulative effects on the overall team score. For example, if the trend of minor score reductions were to remain consistent for all of the athletes left in a lineup following a missed routine, then the overall team score reductions could be significant enough to lower the team's rank in an event or competition.

While this research looks only at the routine immediately following a miss, future research could look at all of the athletes in a lineup following a missed routine. The extent of impact of a missed routine could change the methodology used for determining a lineup for an event. With enough data, one could also look at the impact of a missed routine based on different locations in the lineup. If trends reveal that a specific spot in the lineup tends to be assessed less harshly, then one might opt to place a less-skilled athlete in that spot to minimize score reductions. Another consideration for furthering this research is to incorporate data from other schools to compare the impact of a missed routines between different teams.

Overall, this research can be applied to college gymnastics teams to help determine the best lineup orders as teams gain a better understanding of scoring patterns between judges and the impact of athlete performance on other subsequent athletes in lineups.

Sources

(1) Diane M. Ste-Marie, Sheri M. Valiquette Gail Taylor (2001) Memory-Influenced Biases in Gymnastic Judging Occur across Different Prior Processing Conditions, Research Quarterly for Exercise and Sport, 72:4, 420-426, DOI: 10.1080/02701367.2001.10608979
(2) "Gymnastics History." Where Gymnastics Started, Athnet, www.athleticscholarships.net/history-gymnastics.htm.
(3) Pajek, Maja Bucar, et al. "Reliability Improvement of Real-Time Embedded System Using Checkpointing." 2008 Second International Conference on Secure System Integration and Reliability Improvement, vol. 3, no. 2, ser. 47-54, 2009. 47-54, doi:10.1109/ssiri.2008.15.
(4) Plessner, Henning. "Expectation Biases in Gymnastics Judging." Journal of Sport and Exercise Psychology, vol. 21, no. 2, 10 Dec. 1998, pp. 131–144. Human Kinetics Journals, doi:10.1123/jsep.21.2.131.