Bruin Sports Analytics Dec 17, 2020 8 min read

Execution and Difficulty Trends in Elite Gymnastics

By: Kaylee Diller and Avery Robinson

Source: Emmanuel Dunand/AFP/Getty Images

In elite gymnastics, a gymnast's score is broken up into two components: execution and difficulty. The execution score is out of ten, and can be thought of as a grade of how perfectly the routine was performed. Introduced after the 2004 Athens Olympic Games, the difficulty score is the summation of the difficulty value of each skill in the gymnast's routine, and technically has no limit. The final score a gymnast receives is the sum of the two components.

Given that both components contribute to the final score, a gymnast will want to optimize both scores as they approach the pinnacle of a gymnastics quad -- the four years leading up to an Olympic Games -- in hopes of gaining a coveted spot on an Olympic gymnastics team. Consequently, elite gymnasts will attempt to perform harder, and more perfect, routines.

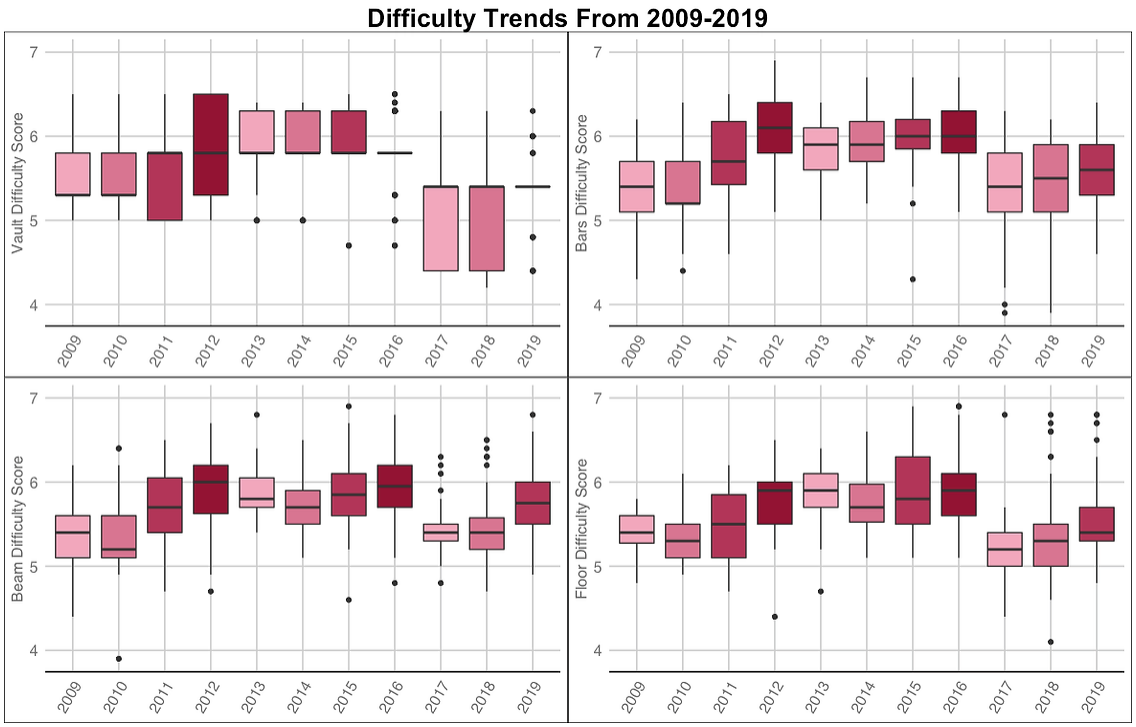
Two common notions within the gymnastics community are that difficulty scores and execution scores

Two common notions within the gymnastics community are that difficulty scores and execution scores increase as a quad progresses, indicating that overall scores increase as well. An alternative belief, however, is that there is a trade-off between these two components: as a gymnast increases their difficulty, risk for error grows, and that risk may be reflected in poor execution.

To investigate the merit of these beliefs, we gathered data from the USA gymnastics official website from three different quads: 2009-2012, 2013-2016, and 2017-present. We focused on American senior elite female gymnasts and took data from major competitions each year. With this data, we aimed to find evidence for or against the beliefs mentioned above, and further, whether the difficulty score or execution score is more important to a gymnast's final ranking on each event.

Difficulty Analysis

To begin, we used box plots to analyze the change in difficulty scores through 2009-2019 for each event.

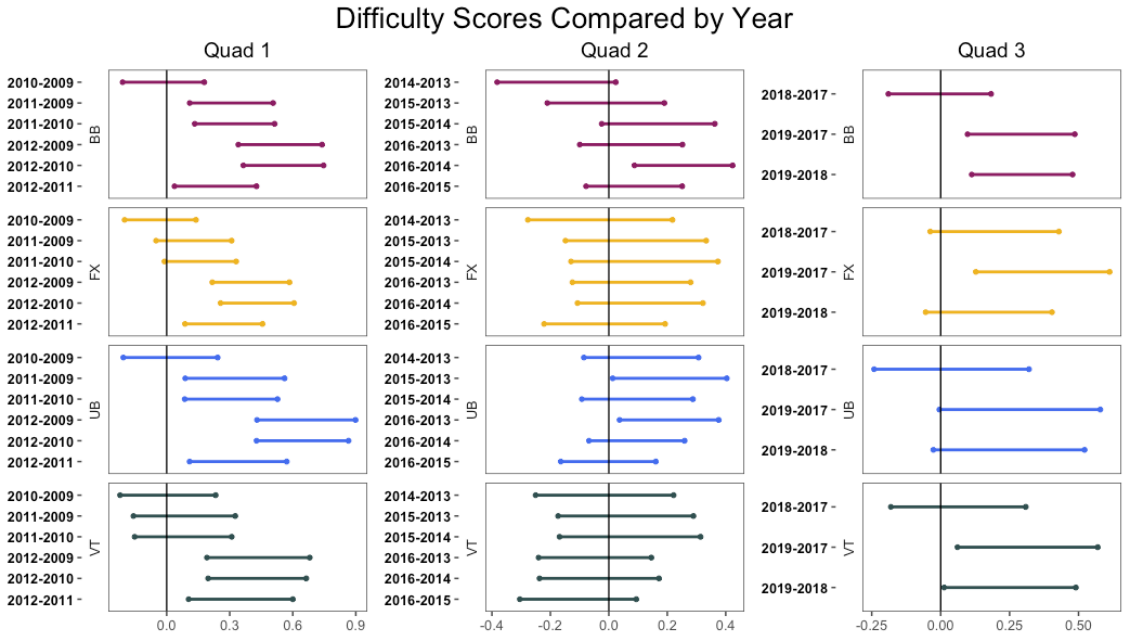


The shade of pink of the box plot corresponds to that particular year's placement in that particular quad, with darker pink meaning further along in the quad. At first glance, it seems that bars, beam, and floor all increase with the progression of each quad. The trend for vault difficulty scores is less clear than the others.

We calculated several F tests within each event and within each quad. In this scenario, we could not combine the quads together. In gymnastics, the code of points is amended after each quad, meaning that the difficulty value of a particular skill is subject to change; therefore, we cannot compare raw difficulty scores between quads, only within. The results are as follows:

Each p-value that is below our predetermined alpha level of 0.05 is significant. For example, we see that the p-value from the F-test for beam difficulty scores from the first quad is 2.99e-14. We interpret this by saying that, within quad 1, at least two of the means of difficulty scores for beam from the four different years are significantly different. The F-tests results are promising evidence in favor of the belief that difficulty scores increase with the progression of the quad. What the F-test results do not tell us, however, is which years within each quad were statistically significantly different. To gather this information, we performed Tukey HSD post hoc analysis.

The graph below shows the results of our post hoc analysis, split up by quad. From top to bottom, the graphs correspond with the post hoc analysis for balance beam (purple), floor (yellow), bars (blue), and vault (dark green). The Y-axis text tells us which two years are being compared, and the bars within each graph display a confidence interval for the difference of the means for the years being compared. The vertical black line in each graph intersects zero on the x-axis.



Consider the top-most bar in the upper left graph, which is comparing balance beam mean difficulty scores in Quad 1 between 2010 and 2009. We can see that the horizontal magenta bar is intersected by the vertical line, and we can interpret this to mean that, since the confidence interval for the difference in mean beam difficulty scores from 2010 and 2009 contains zero, we *cannot* conclude that those two years have significantly different beam difficulty scores. Only the horizontal bars that do not pass through zero can be deemed statistically significantly different.

The later years in the quad are near the end of each box. If we were only looking at Quad 1, we could say that, without a doubt, difficulty scores go up as the quad progresses, especially at the final year of the quad. Quad 3 has the same trend. We see that, as the quad progresses, the differences are either significant or extremely close to being significant. Quad 2, however, tells a slightly different story. For vaulting difficulty scores, only 2016 and 2013 are concluded to be different. For bars, as well, we can conclude that 2015 and 2013, and 2016 and 2013, are significantly different. We cannot, however, make any conclusions about difficulty score differences for vault or floor.

for vault or floor.

It would be interesting to explore further as to why Quad 2 shows different patterns. It could be that, for example, there was a surge in injuries, causing gymnasts to perform lower difficulty skills; there are numerous other factors at play. It would also be interesting to factor in more Quads, rather than the three presented in this article, to understand the trends better. It could be that the pattern of difficulty scores increasing as the quad progresses is least prevalent on vault and floor exercise. More data is needed to analyze such trends.

Still, it seems that there is merit to the idea that difficulty scores increase as the quad progresses, most notably on balance beam and uneven bars.

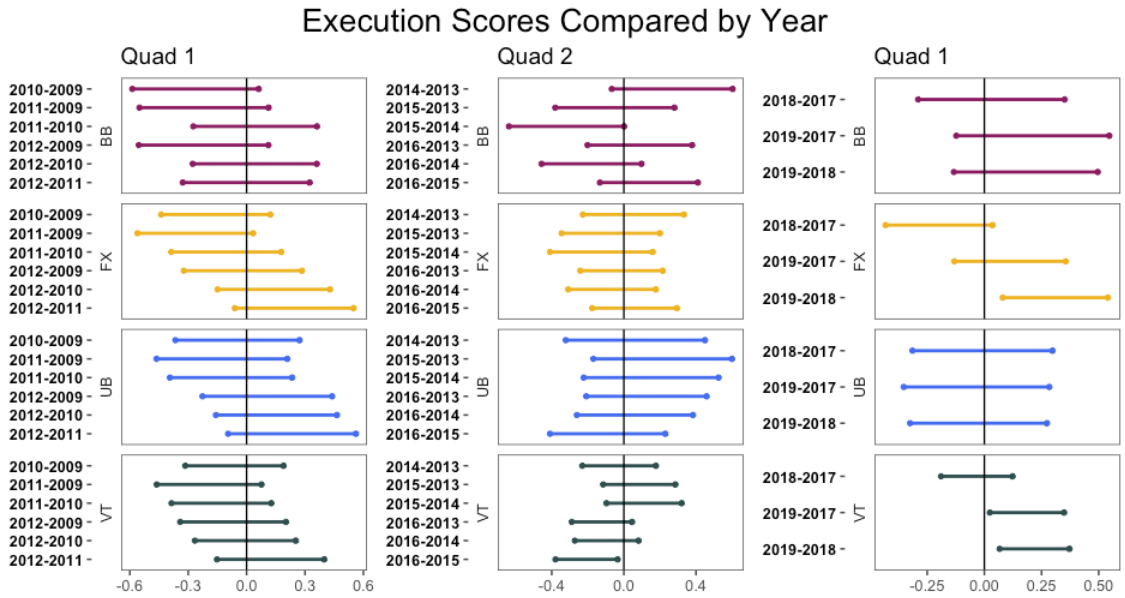
Execution Analysis

Next, we performed the same analysis, but on the trends of execution scores through the quads. We can think of execution as a grade of how well the routine was performed, judged out of 10 points. We are investigating whether execution scores increase as the quad progresses. The idea behind this belief is that, as gymnasts continuously perform their routines, they should progressively get better at performing them. The trends for execution scores can be visualized as follows:

Immediately, it is clear that the trends in execution scores are not nearly as clear as those for difficulty scores. To make sense of the above graph, we started with F-tests:

Using an alpha level of 0.05, we can see that our F-tests comparing means of execution scores result in quite fewer significant results than our F-tests comparing means in difficulty scores. To investigate further, we used Tukey HSD post hoc analysis tests and obtained the following results:

Tukey HSD post hoc analysis tests and obtained the following results:

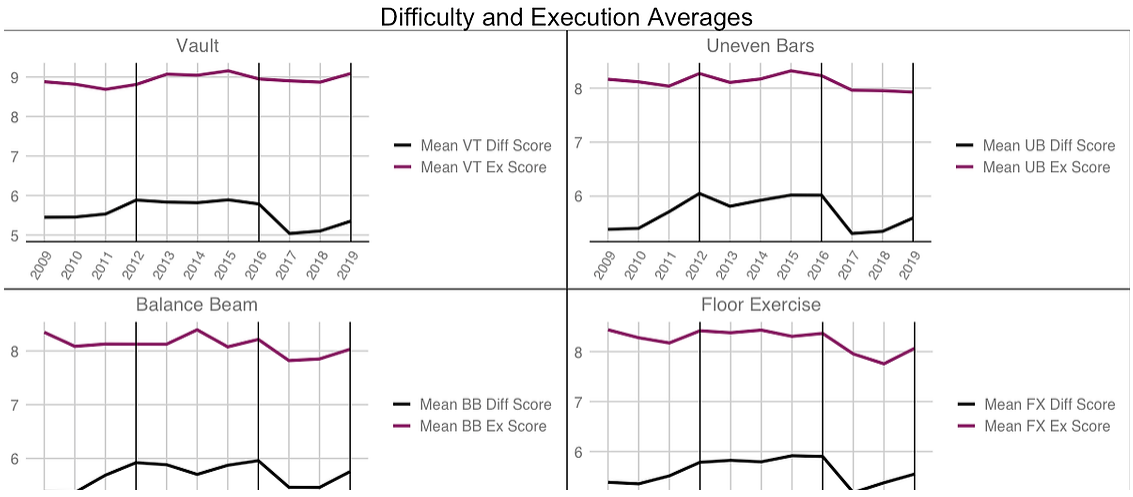


Focusing in on one box, if we see the bars, from top to bottom, increasingly move towards the right, this indicates that the differences are becoming greater in a way that would show us increasing execution scores as the quads progress. As we can see from the boxes above, this trend is observable in quad 1, and slightly in quad 3, but not in Quad 2. Still, even though we see a slight shift to the right in quads 1 and 3, there are still barely any statistically significant differences.

We conclude that there is no clear trend in execution scores as the quad progresses, which goes against the idea that execution scores increase as quads progress. It is worth noting that Quad 2 again has different patterns than the other two quads.

Interaction Analysis Between Execution and Difficulty

By now we have concluded that difficulty scores increase as the quads progress, and also that there are not clear trends on execution scores as the quad progresses. Next, we take a look into difficulty and execution scores together to investigate the idea of a “trade-off” between the two scores.



The graphs above display the mean values of difficulty score and execution score for each event throughout the years. The vertical black lines divide the years by quad. As we know from our previous analysis, difficulty scores trend upward towards the end of a quad, as we can see from the dark green lines in the graphs.

Additionally, we know that there is no clear trend in execution scores as the quads progress. Therefore, it is not surprising that these graphs do not provide evidence for the idea of a trade-off between the two components. To be sure, we built several models to test the interaction between execution score and difficulty score when predicted final score. The t tests for the coefficient of the interaction terms are as follows:

As we can see, two of the coefficients proved to be significant, but after analyzing the total variation explained by each of the significant values, they do not seem to be practically significant. This indicates that, in general, there does not seem to be merit to the trade-off belief.

Finally, to explore which of the two components of the final score had more effect on outcome, we plotted the difficulty scores versus the execution scores for each event, and color-coded the points according to the rank of the final score obtained in the competition. For example, in the upper left-hand scatterplot of the graph below, a dot of dark blue indicates that the score was within the top 25th percentile of vaulting scores for the competition that that particular data point came from. Yellow dots correspond to the next 25% best scores, and so on for magenta dots, and light blue dots, for the third and fourth 25% sections, respectively.

Beam Difficulty (Scaled) | Floor Difficulty (Scaled)

The more separable the colored dots are by a vertical line, the more we can attribute the ranking group to difficulty score. The more separable the colored dots are by a horizontal line, the more we can attribute the ranking group to execution score. It seems that, for vaulting scores, separating the groups by a vertical line is much more feasible than for the other events. To quantify these relationships, we used ANOVA methods, and obtained the following results:

These results tell a similar story to the graphs above. As we can see, the event that has the highest percent of ranking explained by difficulty score is vault. In essence, this means that out of all four events, it makes the most sense to optimize difficulty on vault. If we look back to the graph above, many vaults of high difficulty value but lower execution scores still made it into the top two groups. This differs greatly from bars, for example. In the upper right panel, we see that the colored dots are much more easily split by a horizontal line rather than a vertical line, indicating that execution should be prioritized above difficulty.

Based on this data, it seems that a gymnast wanting to optimize her ranking on each event should focus strongly on the difficulty value of her vault, and once her difficulty value on bars, beam, and floor are a bit above average, to then focus on perfecting her execution on those events rather than adding harder skills.

Conclusion

In conclusion and based on the three quads analyzed above, it appears that there is evidence in favor for the belief that difficulty scores increase with the progression of a quad, and evidence against the two beliefs that execution scores increase with the progression of a quad and that there is a general trade-off between execution and difficulty scores. Additionally, it seems that difficulty score on vault plays a bigger role in ranking than difficulty score on the other events, for which execution scores seem to be more important. This information could help direct a gymnast's attention as to how she might optimize her ranking. Finally, it is worth noting that, while there were numerous clear trends in the data, integrating data from more quads would allow us to make more robust conclusions and perhaps add insight into some of the less clear trends observed.