Analysis of Girls' Basketball Teams in Southwest Riverside County Using Advanced Statistics

Word Count: 4302

Introduction

Statistics in sports have grown exponentially in the last decade. This has much to do with advanced data collection technology in the modern age. Today, almost all sports in all age groups are utilizing statistics to further understand their game performance. As a result, these statistics have become increasingly complex as statisticians hope to gain an edge over their opponents through a quantitative lens. Advanced statistics or advanced metrics are the terms most commonly used to describe such intricate statistics. As stated by renowned baseball statistician, Nate Silver, advanced statistics can be defined as a combination of qualitative and quantitative data insights used to identify new information as scores on a hard-numerical scale (Silver, 2012). This means that game information can be manipulated through the use of mathematical formulas to illustrate the traits of a team or player. Many of these types of statistics have become commonplace for their use in sports, even at lower levels of play.

Specifically, this research is focused on adjusted advanced statistics. Adjusted advanced statistics are directly influenced by a team's opponents, and indirectly by those opponents' opponents (Zimmerman, 2016). This means that the pace of the game is accounted for within the level that the statistic is being used for. This makes the statistic more accurate and easily transferrable to different levels of play whether that be due to gender, age, or other factors of competitiveness. Although advanced statistics have been gaining traction in the NBA and NCAA, that same progress is lacking in the women's sector. Currently, the NBA has both advanced player and team statistics; however, the WNBA has only started using advanced player statistics in the past couple of years. As a result, this research focuses on team statistics at the high school level for girls due to both the accessibility and the lack of development that is

currently represented in that sector. This research should provide insight about the accuracy of advanced statistics and its potential for growth in areas that it is not currently being prioritized in.

Advanced statistics provide the opportunity for coaches and teams to gain valuable insight about their performance that cannot otherwise be seen in the moment during a game or with common place statistics. Thus, by advancing statistics that are currently being used by high school level basketball teams and analyzing the opinions of coaches at this level, it can be determined whether advanced statistics can be suited for intermediate and women's level basketball. Doing such could largely benefit the game of basketball by diminishing disparity between amateur and professional divisions and those of the women's or men's sectors and therefore, increase competitiveness for the sport in the future.

Literature Review

Data analysis in sports takes root in baseball sabermetrics. This is derived from SABR, Society for American Baseball Research, which was founded in 1971; however, their statistical approaches didn't become largely popularized until 2002 with the Oakland Athletics (Birnbaum, n.d.). The team had unprecedented success due to their reliance on objective baseball statistics. This soon set a precedent for the rest of baseball and later, the majority of sports to follow. In their journal entry on "The Datafication of Everything", Brad and Rob Millington correlate this development to the trend of 'Big Data' (Millington & Millington, 2015). Big Data refers to the constantly growing, complex data sets that can be analyzed by computers to find patterns. In sports, it is being used to uncover information in a multitude of ways. Authors of "State of the Art of Sports Data Visualization" for the Computer Graphics Forum consider three main data types which include: box-score, tracking, and meta data (Perin et al., 2018). Box-score data

regards game details from scorecards, tracking data utilizes machine vision and motion sensing to produce specific real-time player and equipment data, and meta data is contextual gameinfluencing data. Box-score data is the most commonly used in sports today due to its simplicity. These developments, coming first from baseball, spread rapidly among all sports and demonstrate the impact statistics have.

The groundwork for sports statistics in academia is sparse due to the fact that it is centered around a leisure activity. This in itself is a gap in the body of knowledge because the sports industry is worth about $75 billion today in the United States (Gough, 2019). However, sports analysts such as Dean Oliver and John Hollinger have been able to document developments in advanced basketball statistics. Dean Oliver's "Basketball on Paper" published the foundation of basketball statistics while John Hollinger provided annual "Pro Basketball Forecasts" that analyzed every player and team in the NBA. Oliver's work was focused on the four main factors that contributed to basketball statistics: shooting, turnovers, rebounds, and free throws, with the weights of their significance in a game being 40%, 25%, 20%, and 15% respectfully (Oliver, 2004). These factors are the main box-score statistics that have become mainstream throughout leagues from every age, gender, and caliber of play. And they can be combined or manipulated to form the basis of advanced basketball statistics.

Advanced statistics are regularly used in professional and pre-professional basketball to gain valuable insight about the game. Unfortunately, it is rarely used in women's basketball, even at the professional level. Only three years ago did the WNBA introduce advanced statistics on their official website (Martin, 2016). Thus, the disparities between leagues due to gender is highlighted. Further, the complexity of advanced statistics makes it unlikely to be seen in lower levels, such as high school. Nevertheless, diminishing this disparity can be an enormously useful

tool in increasing the competitiveness at this level. Fortunately, pace-drive statistics can be tailored to the speed at which the game is being played. Albrecth Zimmerman's "Basketball Predictions in the NCAAB and NBA: Similarities and Differences" supports pace adjustments and finds them to be the most accurate at representing game data (Zimmerman, 2016). The success of statistics like offensive, defensive, and net ratings in NCAA (the National Collegiate Athletic Association) and NBA games alike appears promising for High School and women's level basketball. These statistics are featured on the NBA's website. They were created by Dean Oliver and they allow a broader analysis of a team by illustrating their play divided into the two components of the game—offense and defense (Kubatko et al., 2007). These statistics can allow for a coach to understand his weaknesses in either sector and improve upon them or analyze another team in their league to find deficiencies.

In my research, I hypothesized that the success of adjusted statistics such as offensive, defensive, and net rating at both the NCAA and NBA levels, can be transferrable to the women's and high school sector. This prompted my research question of: can girls' basketball teams in Southwest Riverside County be accurately depicted by the results of advanced statistics? I used publicly available box-score statistics on five girls high school teams located in my community to perform this research. By comparing the teams' offensive, defensive, and net ratings and their standings, I should identify a positive relationship that proves interchangeability of the statistics between different levels of competitivity. This will then be examined by the coaches of the teams to prove the accuracy and potential implementation.

Methodology

Because statistics are limited for high school basketball, teams are unable to uncover their full potential. However, by implementing advanced statistics into amateur level athletics, teams can make the most out of the information they obtain from game play. Currently, the NBA uses an assortment of advanced statistics to represent both teams and players; however, such is limited for men's college level basketball and even more so for women's and high school level programs. Thus, utilizing such statistics in these sectors can benefit their game insight.

Although there are many insightful statistics, offensive and defensive rating were chosen because they measure a team's efficiency on both ends of the court. Further, they are both pacedriven statistics when adjusted, meaning that they account for the game speed according to the team they are playing. This makes it easy to convert into high school basketball because speed is large discrepancy between the different levels of sports. Moreover, this team statistic was chosen over individual statistics because unlike professional basketball, high school teams are constantly changing in terms of players. Thus, a team statistic is the best representation of performance at this level.

Extracting Data

Team statistics from the high school statistics recording website, MaxPreps, will be used to gain information on girls' basketball teams in Southwest Riverside County. MaxPreps is one of the only high school websites, thus, it is used by the majority of high school athletics to publicly report individual and team statistics. To extract data from this website, Microsoft Web Query will be used. This application was chosen because it is a feature to Microsoft excel which is manageable for a high school student and it also allows for the statistics formulas to be later

applied. To extract the data, MaxPreps was accessed from the query. The appropriate team was then chosen and the data tables including their statistics were imported onto the document. The initial data included information not relevant to the advanced statistics formulas being used in this research, so the data had to be cleaned and organized until only points, field goal attempts (how many times the ball was shot in attempt to score), free throw attempts, offensive rebounds, turn overs, and points against were included. This was repeated for all five teams in the league and then loaded into the excel document.

Transforming Data

The teams' statistics were then ready to be utilized in the advanced formulas. To do this, they first had to be organized by team, then game. These sections were then broken down into the individual team statistics of points and possessions to be used in the formulas. The formulas for defensive and offensive ratings were constructed by Dean Oliver as a means to measure efficiency on both ends of the court with respect to pace of the game (Kubatko et al., 2007). They are as follows:

$$\text{AdjOE} = \text{OE} * \text{avgall teams (OE)} / \text{AdjDEopponent}$$

$$\text{AdjDE} = \text{DE} * \text{avgall teams (DE)} / \text{AdjOEopponent}$$

Where OE is offensive efficiency calculated by $\text{OE} = \text{PPG} * 100 / \text{Possessions}$ and DE is defensive efficiency calculated by $\text{DE} = \text{PAG} * 100 / \text{Possessions}$. Possessions are essential to the game of basketball because they are representative of how many times a team has the ball, and thus, how many chances to score they have. Dean Oliver and his colleagues who wrote, "A Starting Point for Analyzing Basketball Statistics", state that the possessions for two teams should be approximately the same due to the alternating nature of the game (Kubatko et al.,

2007). Because of this, the slight disparity means that whoever had the most possessions in a game should be the winner, thus, making them extremely important to the game and to calculating advanced statistics. Possessions are calculated as: Possessions = FGA-OR+TO+.475*FTA.

Derived from the previous two is net rating (NetRtg). This can be found by subtracting defensive efficiency from offensive efficiency. In this case however, adjusted offensive and defensive efficiencies are substituted in. This formula is commonplace in both NBA and WNBA statistics as a way to measure a team's efficiency as a result of their defensive and offensive performance. The formula is demonstrated as: NetRtg = AdjOE – AdjDE

Using these formulas, basic team statistics will be transformed into advanced statistics. The offensive rating will first be found by inputting a team's points per game and possessions per game into the non-adjusted offensive efficiency formula. This will be done for all five teams for all twelve games of their season. This will be averaged for each team to find their offensive rating. Then, the average of all of the net ratings will be calculated for later input into the paceadjusted offensive efficiency rating. This process will be repeated to find the defensive rating for each team. With this information, the adjusted efficiency would be found by multiplying the offensive efficiency of team 1 by the average offensive efficiency of all teams, then it will be divided by the opposing team's adjusted defensive efficiency. This adjusted defensive efficiency is used from the team's prior game; however, when calculating the results for the team's first game of the season, the adjusted defensive efficiency will be only an average of their defensive efficiency from the season. This process is similar for calculating adjusted defensive efficiency except offensive efficiency is used in place of defensive efficiency and adjusted offensive efficiency is used in place of adjusted defensive efficiency. After this, both the

adjusted offensive and defensive efficiencies will be averaged for each team to give them a rating of their season.

Finally, net rating will be found by subtracting the calculated adjusted defensive efficiency from the adjusted offensive efficiency for each team. The results for these three processes will be recorded into separate charts for each team.

Data Analysis

The recorded adjusted offensive and defensive efficiencies and net ratings with a line of best fit representing the average of each will then be graphed to create a visual representation of the results. Each of the five teams will be represented by a different colored point on the graphs. Teams above the line of best fits are performing better than the average of all of the teams. Using this information, I will compare the results to the league standings at the end of the season to determine whether there is a correlation. The teams' offensive and defensive efficiencies will likely fluctuate due to different strengths and weaknesses. With this information alone, a team can improve upon their deficiencies or determine how they will match up against another team within the league, thus, allowing them to improve their game preparedness. Moreover, the net rating outcomes should correspond to the league rankings to provide backing to the study's results.

Interview

With the resulting statistics, an interview amongst the league's coaches will be conducted. This is done to gain insight on the accuracy and potential application of advanced statistics in girls' high school basketball. The first question examined is "What statistics do you currently use?", this allows me to preface my research scope about the current use of statistics at

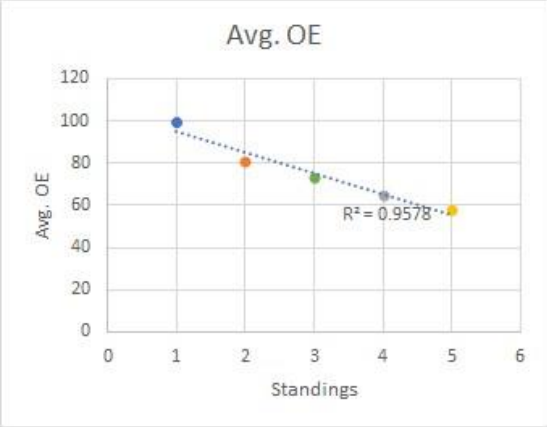
the high school level. Next, asking "Do these statistics benefit your team? How?" provides an understanding of how teams are using statistics. Then, asking "Do you know what advanced statistics are?", will uncover how many coaches are knowledgeable of this type of statistic and it will allow an introduction of advanced statistics for those who do not know what that is. Further, the data analysis results will be shown to the coaches at this time to reveal the results of advanced statistics use. Next, the coaches will be asked "Does the data resulting from the use of advanced statistics accurately represent your team? Why or why not?", to reveal the advanced statistics' accuracy. If the previous question does not apply to their team, the coach will be asked to explain why. If the results do apply to their team, the coach will then answer concluding questions about the possibility of implementing this method into their coaching strategies.

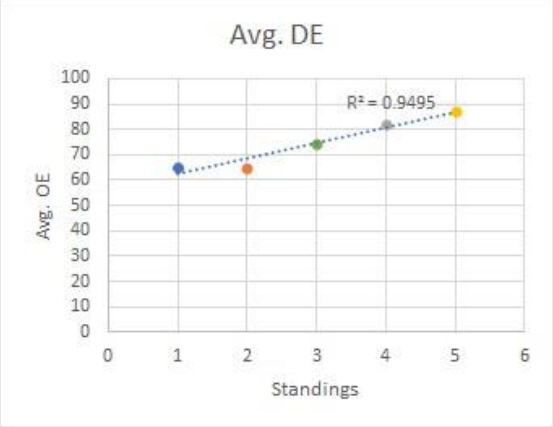
From these interviews, it will be determined if advanced statistics can have an impact on high school girls' basketball. Because high school level teams are comprised of minimal staff, the coach is the team's statistics analyst. Because of this, their opinions are knowledgeable when considering the subject of advanced statistics.

Findings

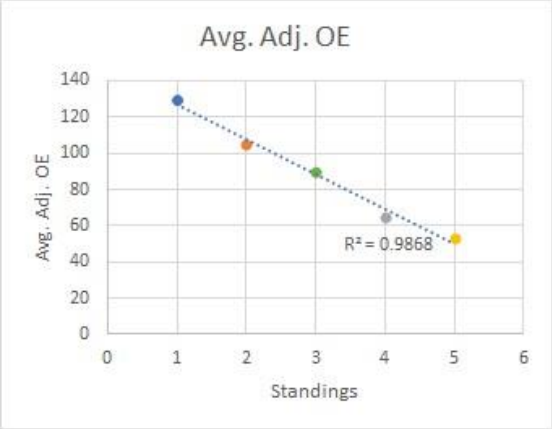
A review of my data from excel reveals that there is a high correlation between defensive and offensive rating and a team's ranking. The graphed adjusted offensive values proved to have a negative relationship with and R squared value of 0.9578. This is statistically significant due to its large, almost 1 value. This shows that as the teams ranked lower, their offensive efficiency decreased. The opposite trend occurred for defensive efficiency which illustrates that the opposing team had less opportunities to score on the higher ranked team, thus, giving them a

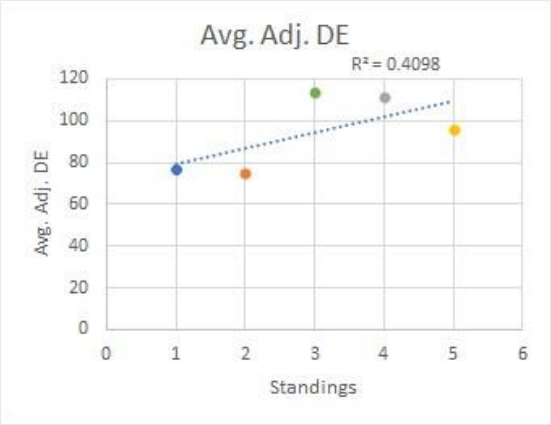
lower and more advantageous defensive efficiency. It had an R squared value of 0.9495 which was significant for my data.





After graphing adjusted defensive and offensive efficiencies along with the rankings, there continued to be the same negative trend for the offense. Its R squared value was 0.9868, which again, was accurate. The adjusted defensive efficiency, however, proved to have some discrepancies in rank vs. Adj. DE. Its R squares value of 0.4098 illustrated that rankings were not significant in the adjusted defensive efficiencies for these teams.

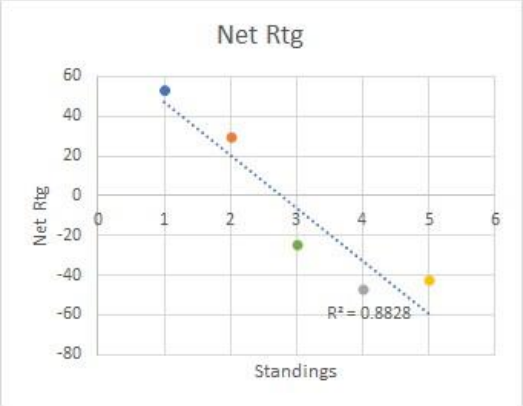




As a whole, by combining both the defensive and offensive efficiencies into net rating, it can be seen that deficiencies on offense or defense for higher ranked teams are made up on the opposite end. This is because the R squared value of 0.8828 illustrates an almost linear

relationship between net rating and team standing. Thus, the team with the highest net rating was the highest ranked team. This shows that advanced statistics accurately represent the abilities of a team. More than that, all adjusted statistics were also recorded for each game each team plays.

This shows in depth how the defense and offense were in each game.



After I had compiled my data, I sent out a survey to all coaches of the team that I analyzed to gain insight on the accuracy and possible implementation of these advanced statistics in girls' high school basketball. Out of the five coaches, four responded. These coaches came from team 1, 2, 3, and 5. They stated that their current statistics are comprised of those available on MaxPreps and Hudl, both of which are focused on basic statistics. Their responses to the information in the graphs illustrated that they accurately represented their team. They further stated that new information could be revealed by using advanced statistics to represent their team. Finally, they all concluded that advanced statistics should be implemented at the high school girls' basketball level to gain additional game insight.

Implications

The scatter graphs from the unadjusted offensive and defensive ratings proved to be relatively linear, as seen by their high R squared values. Thus, as a control, the two advanced

statistics formulas performed accurately. This is because the offensive efficiencies went down about -10 points for each lower standing. This was expected due to the fact that lower ranked teams lost more games because they had less offensive impact on the game. As for defensive efficiencies (DE), the opposite occurred and as the standings decreased, the defensive efficiencies went up about half as much as the offensive efficiency did at 6 points for each lower standing. This positive trend line was also expected due to the fact that a lower DE means a team performed better on defense because they allowed their opponent to perform worse on them. The slowly increasing defensive efficiencies for this trend however, illustrates that there is much less of a discrepancy between defense of teams and their standing for teams in this league.

The graphs for adjusted offensive and defensive efficiency showed greater discrepancy when teams were compared to their opponents. The adjusted offensive ratings illustrated a steady decrease of about -20 points in efficiency as the standings decreased. This shows that offense is where higher ranked teams are more efficient. With the adjusted defensive ratings, there seemed to be no correlation with standings. This has to do with the fact that unadjusted defensive efficiencies were relatively close. Further, by looking at the individual game statistics, it can be determined that some lower ranked teams had better defensive games against higher ranked teams and still lost to them. These discrepancies may be reason for the varying defensive efficiencies.

Upon evaluation of the official NBA advanced statistics, it can be seen that larger discrepancies between ranking and rating occurred for defensive rating, thus, making it most likely the case for my skewed results. Additionally, it can be seen that net rating had the ability to more accurately represent a team in the terms of standings.

Limitations

There were some discrepancies between ranking and the efficiency measuring statistics. Not because of the pace of the game, because pace is reflected in the adjusted statistics, but because of the largely different style of play that occurs in the girls' high school level of basketball. These statistics have more prominent results in the NBA due to the fact that teams do not have largely varying standings throughout the season. Another factor that may have impacted my results comes from the length of the season. The Southwestern League only plays 10 league games as opposed to the 82 games in the NBA and the 30 games in the NCAA. Thus, if more games were played, we would be able to see more stable and accurate results.

From the coaches' responses to my results, it can be concluded that this method of manually applying advanced statistics would be too difficult for a high school coach to manage. This is because all of the coaches in the league focus on statistics provided to them from websites such as MaxPreps and Hudl, as per their responses to the question of what statistics do they currently use. Thus, creating or implementing the formulas into a website, such as MaxPreps, where the coaches simply have to update their box score statistics, would be the only way coaches would use the advanced statistics.

Future Studies

To further future studies on sports statistics as a whole, it is important that more scholarly research is published to increase the body of knowledge on the subject. This alone would allow for more support for basketball statistics. As for my study, perspectives from NBA and NCAA level coaches about the accuracy and significance of advanced statistics could provide further insight as to whether its implementation would be beneficial at the lower level.

Due to the low correlation I received for my adjusted defensive efficiencies, I would further my research by focusing on this one statistic. I would go into depth on the factors that contributed to this, with some of them being external factors. These factors could include defense type played at this level (most play simple man or zone) and steals. Thus, to research this, I could do a qualitative analysis on the teams in this league by watching their game play or film and then dissecting the defensive elements of it that have been neglected by the defensive efficiency rating statistic.

To increase the depth of my study, further research can be done on other statistics such as individual player efficiencies, as suggested by one of the coach participants in my study. This would give coaches insight on which players to play in their games. And although such is enticing, it would only be beneficial if this type of calculation was easily accessible by a coach due to the fact that the information would be needed right away because players on high school teams are constantly changing. As a result, team statistics are much more plausible for high school levels, especially when looking at cumulative results.

To make advanced statistics plausible for coaches, it is important that they are easily accessible for coaches. This means that they must be formulated into websites such as MaxPreps and Hudl. Here, coaches would only need to put in their box score statistics to find their efficiencies. The limitation to this however, is the fact that the calculations are quite complex due to the fact that they take into account the pace of the game, thus, the opposing teams' statistics would have to be matched up with their opponent to find the results, making them unlikely to be implemented in the near future.

Conclusion

The results from this study proved to show that advanced statistics could be used at the high school level for girls' basketball to improve game insight for each team. One coach even claimed that implementing advanced statistics should be used "by every coach on every team and it should be mandatory". This shows a strong inclination by coaches to use advanced statistics if they were available to them.

From this, it can be illustrated that there are many team benefits from using advanced statistics. Specifically, as compared to the opponents within the league. In addition, one of the coaches brought to light that these statistics could even be useful for college recruiting. Because of this, advanced statistics are more important at the amateur level than I had thought before; especially because playing at the collegiate level for these athletes is dependent on exposure both in person and through statistics.

Further research pertaining to the use of advance statistics in the WNBA and women's college basketball could cause its significance to spread down to the high school level. Because of this, it is important that the higher levels are first targeted to implement higher level statistics for results to be possible at the lower levels.

If advanced statistics were implemented in the high school level. It can prove to provide additional game insight to teams and their coaches. All five coaches from the teams I analyzed supported this claim. And because these coaches also stated they can improve their game strategies as a result of the information they received, implementing advanced statistics could even increase competitiveness in women's' high school basketball.

From this research, it can be seen that advanced statistics are an accurate representation of girls' high school basketball. Because of this, they should be implemented into this level of basketball as another tool for teams to better understand their team. Specifically, this would give coaches a new understanding of how their team fits offensively and defensively against other teams in their league. Thus, they can adjust their strategies based on this new information.

References

Birnbaum, P. (n.d.). A Guide to Sabermetric Research. *Society for American Baseball Research*. https://sabr.org/sabermetrics

Gough, C. (2019). North American Sports Market Size from 2009 to 2023. *Statistia.* https://www.statista.com/statistics/214960/revenue-of-the-north-american-sports-market/

Hollinger, J. (2005). *Pro Basketball Forecast*. Potomac Books.

Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. (2007). A Starting Point for Analyzing Basketball Statistics. *Journal of Quantitative Analysis in Sports*, 3(3). doi: https://doi.org/10.2202/1559-0410.1070

Martin, B. (2016). *A New Level of Analysis: Into to Advanced Stats.* WNBA. https://www.wnba.com/news/new-level-analysis/

McGoldrick, K., & Voeks, L. (n.d.). "We Got Game!": An Analysis of Win/Loss Probability and Efficiency Differences Between the NBA and WNBA. *Journal of Sports Economics*, *6*(1), 5–23. https://doi.org/10.1177/1527002503262649

Millington, B., & Millington, R. (2015). "The Datafication of Everything": Toward a Sociology of Sport and Big Data. *Sociology of Sport Journal*, *32*(2), 140–160. https://doi.org/10.1123/ssj.2014-0069

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Raab, M. (2013). Data Mining in Elite Sports: A Review and a Framework. Measurement in Physical Education & Exercise Science, 17(3), 171–186. https://doi.org/10.1080/1091367X.2013.805137

Oliver, D. (2004). *Basketball on Paper*. Potomac Books.

Perin, C., Vuillemot, R., Stolper, C. D., Stasko, J. T., Wood, J., & Carpendale, S. (2018). State of the Art of Sports Data Visualization. *Computer Graphics Forum*, *37*(3), 663–686. https://doi.org/10.1111/cgf.13447

Silver, N. (2012). *The signal and the noise. The art and science of prediction*. London: Penguin.

Zimmermann, A. (2016). Basketball predictions in the NCAAB and NBA: Similarities and differences. *Statistical Analysis & Data Mining*, *9*(5), 350–364. https://doi.org/10.1002/sam.11319