Spring 5-2-2016

"One-and-dones": A Data Science Analysis of the Impact of Leaving College Early for the NBA on a Player's Career

Ruth C. Catlett

"ONE-AND-DONES": A DATA SCIENCE ANALYSIS OF THE IMPACT OF LEAVING COLLEGE EARLY FOR THE NBA ON A PLAYER'S CAREER

An honors paper submitted to the Department of Computer Science
of the University of Mary Washington
in partial fulfillment of the requirements for Departmental Honors

Ruth C Catlett

May 2016

“One-and-dones”: A Data Science Analysis of the Impact of Leaving College Early for the NBA on a Player’s Career

Ruth Catlett

Computer Science Honors Thesis
Faculty Advisor: Stephen Davies

Dedication

To my dad, a fellow Kentucky fan who has enjoyed
this project as much as I have.

To my mom, for always being on my team and
encouraging me to be my best.

Table of Contents

Abstract

College basketball is a highly popular sport and, in the wake of March Madness, one wonders what will happen to key players next year. The National Basketball Association (NBA) currently restricts players from entering the draft until they are 19. This leads some players to enter college simply as a practice and "waiting area" for the NBA. These players - often termed "one-and-dones" - stay for a year and then at their earliest chance, enter the draft. I wondered if staying in college longer allows NBA-bound players more practice and experience playing under pressure, or if players were better off by leaving college early to play while they were still young.

Using R, I gathered individual player data from both college and the NBA. This required significant work in gathering, fusing, and cleansing electronic data from multiple sources into a usable form. I then investigated various accepted performance aggregation metrics, and settled on efficiency (EFF) which is a relatively simple measure that consolidates a player's yearly performance (including points, rebounds, assists, etc.) into a single number. Using machine learning techniques, I divided the players into "clusters" (small groups of statistically "similar" players) based on their freshman-year data and then examined each cluster individually. For each cluster, I analyzed whether there was a significant difference between the one-and-dones and the others. In this way I could examine the likely effect that additional college experience would have had on a player's NBA career. This analysis found little significance between the “one-and-dones” and the “more-and-dones”, meaning perhaps a player's NBA performance is not hurt by coming out early.

Introduction and Basketball Primer

Basketball is a popular sport, especially at the college level where its season culminates in the “March Madness” national tournament. My research was focused on college students who played Division I (DI) basketball which is the highest tier of play in the National Collegiate Athletic Association(NCAA). The goal of this research was to examine how “one-and-done” (OND) players fared in the NBA compared to other students who stayed in the NCAA longer.

When players are ready to proceed to the NBA they must declare their eligibility and desire to enter the draft. Before 2005, high school students could immediately enter the draft. The 2006 draft marked the beginning of a new rule that dictated players must be 19 years old during the draft calendar year and at least one NBA season must have passed since their high school graduation [10]. This restricts the previously NBA-bound high school players into being forced to wait a year. Some of these players choose to play on an international team where earn a large sum of money playing professionally. Other players decide to continue their education, both in the classroom and on the court. Some students attempt to leave college early, after their freshman year, once they have met the draft eligibility requirements. Those students who choose to leave after a year are often referred to as “one-and-done”. For the purpose of my research those students were contrasted with “more-and-dones” (MNDs). The students who chose to continue their education still might leave before graduation, but the research focused on the difference between ONDs because of the draft rule highlighting the division.

Statistics and Measures

Knowing the basic statistics and measures behind the sport itself are important to understanding how players are evaluated. In [2], team based measures and other high level statistics are discussed and modeled. I collected the following stats by season for each individual player:

GP - Games Played
MP - Minutes Played
FG - Field Goals, including made, attempted, and the total percentage
2P - 2-Pointers (baskets), including made, attempted, and total percentage
3P - 3-Pointers including made, attempted, and total percentage

FT - Free Throws (1 point) including made, attempted, and total percentage
PTS - Total Points scored (sum of FG made, 2P made, 3P made, and FT made)
ORB - Offensive Rebounds
DRB - Defensive Rebounds
TRB - Total Rebounds
AST - Assists
STL - Steals
BLK - Blocks
TOB - Turnovers
PF - Personal Fouls

There are plenty of measures on a player's performance and contribution to a specific game or team. Overall I will discuss the main three measures I noted about a player.

Wins Produced

Wage of Wins was published in 2006 and details some common conclusions and misconceptions in multiple sports. The book discuss basketball in multiple chapters, specifically when comparing Shaquille O'Neal and Kobe Bryant [14]. The authors developed the method comparing players' contributions to their teams by wins produced. Wins produced (WP) is a measure of how many wins a season the individual player is responsible for, the higher the better. WP can also be measured by games or adjusted based on the game's pace. WP is calculated with a multi-step process whereby an individual's performance is compared and adjusted based on their team's performance as well as how their position and time contribution has created wins. Steps to calculate include: calculate the value of a player's production (and production per game), adjust for assists and position, and incorporate team defense (specifically defensive rebounds) [14].

Player Efficiency Rating

An alternate to WP is the Player Efficiency Rating (PER) which was developed by John Hollinger in 2006 while he was an analyst and writer for ESPN [7]. PER attempts to aggregate a player's per-minute performance into one number. PER ensures that the average over all players in a given year is 15 so that two players, years apart, may be compared equally. This allows two players who never played against each other and who played at different times to be compared to see who is better or more productive to a team. PER attempts to take into account the pace of the games and tries to treat no two games in a season the same. To calculate a player's PER certain stats are multiplied by factors and adjusted by the team and league averages. While PER is widely used today, there are problems with it just like all statistical measures. Hollinger himself notes that while PER does not favor a player's defensive contribution, the "[t]wo important things to remember about PER are that it's per-minute and is pace-adjusted" [7].

Efficiency

In the end, the data I collected was not as in depth as was needed for the WP and PER calculations, which are often used as part of the official NBA statistics. The measure of Efficiency (EFF) is still usable and was better suited to my purposes. While Hollinger and *Wage of Wins* created their own measures of individual performance on a smaller scale, they had attributed different weights to different collected stats. The EFF measure took all contributions a player made and averaged them over the number of games played and could easily be adjusted to a per minute basis if needed. The specific formula is seen here:

$$\text{EFF} = (\text{PTS} + \text{REB} + \text{AST} + \text{STL} + \text{BLK} - \text{Missed FG} - \text{Missed FT} - \text{TO}) / \text{GP}$$

While all measures are known to have shortcomings and errors including, on the part of EFF, being too limited in the view of player's contribution, since the focus of the research was to examine a player's impact on themselves, I felt that the EFF was best suited to examining how a player's performance is affected.

One-and-Dones' Impact

I wondered if the age requirement mattered to the players themselves, since they spend one year studying and playing with other teenagers only to leave and be thrust into an environment where people are earning millions of dollars playing a sport they love. When compared to other sports, basketball has one of the higher minimum age requirements and the NBA Commissioner Adam Silver wants to raise it to 20 years old [12]. The Professional Golfers' Association, the National Hockey League, and Major League Baseball have a minimum age requirement of 18; however, all have age waivers if other qualifications are met. According to the Women's Tennis Association a player can become a pro when they are 14 years old, although they are limited in number of competitions until they are 18 [1]. Some basketball players and coaches, such as Gary Williams, felt that players who stayed longer had “stayed in school and learned how to play” [9].

Thesis Questions

Since the OND scenario seems to be a focus of the sport and more a result of the age restriction change, I wanted to evaluate:

- Does leaving college early affect OND’s NBA careers?
- Do ONDs have a better NBA performance, as measured by EFF overall, or do they get a slight edge in the beginning?
- Does leaving after one year giving ONDs better draft positions or higher salaries?
- Does a player’s position a factor in whether or not more time at college helps his career?

Statistical Tools

For this project, I chose to use R and its integrated development environment (IDE), RStudio. It was chosen for the research since I had some familiarity with it. A GitHub repository was established so that my mentor would have access to the data and code as I worked on it and could then assist with any questions or comments.

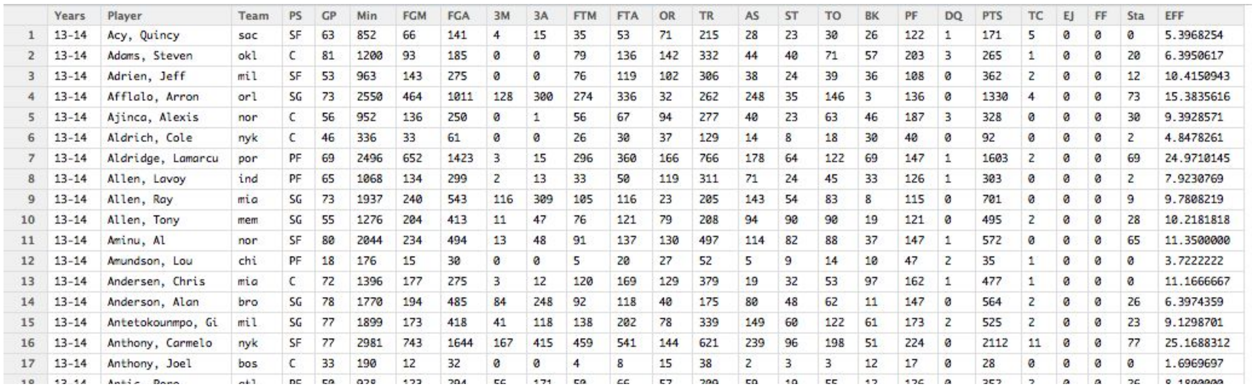
Data Collection and Fusion

Collection

Since the research was focused on the players’ college careers, I needed both NCAA data and NBA data. At first, I attempted to get the data from the associations themselves. Other researchers collected their data from third-party sources including dougstats.com and basketball-reference.com (I only needed NCAA stats from this site, so I used their child site, sports-reference.com). It has been estimated by data scientists and experts that 50 to 80 percent of time working on a data project is spent “mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets” [6]. That was certainly true for this project; collecting the data and ensuring the data import to R was done properly took up more than half of this project’s time.

NBA Data

Doug’s Stats has data for every season from 1988 divided by player. This was exactly what I needed for the NBA data, and Doug was even helpful responding to an email asking about an unlabeled column in data collection. Each row of the data has all the collected statistics that corresponds to one season of one player’s career. In Figure 2 below, the first row is Quincy Acy’s 2013-2014 season and all his collected stats for that year. After collecting all the files, the resulting data set had almost 10,000 rows, one for each year a player spent in the NBA.

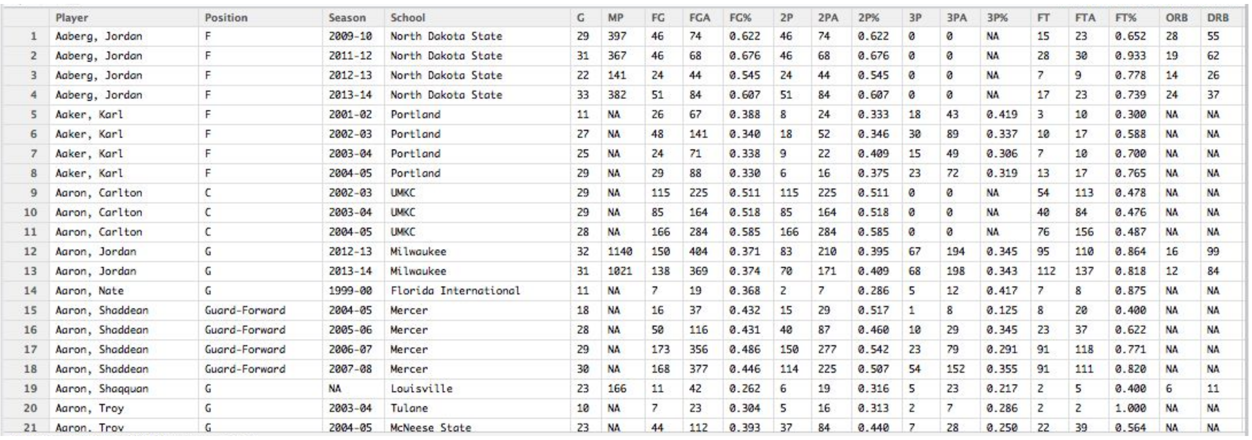


	Years	Player	Team	PS	GP	Min	FGM	FGA	3M	3A	FTM	FTA	OR	TR	AS	ST	TO	BK	PF	DQ	PTS	TC	EJ	FF	Sta	EFF
1	13-14	Acy, Quincy	sac	SF	63	852	66	141	4	15	35	53	71	215	28	23	30	26	122	1	171	5	0	0	0	5.3968254
2	13-14	Adams, Steven	okl	C	81	1200	93	185	0	0	79	136	142	332	44	40	71	57	203	3	265	1	0	0	20	6.3950617
3	13-14	Adrien, Jeff	mil	SF	53	963	143	275	0	0	76	119	102	306	38	24	39	36	108	0	362	2	0	0	12	10.4150943
4	13-14	Afflalo, Arron	orl	SG	73	2550	464	1011	128	300	274	336	32	262	248	35	146	3	136	0	1330	4	0	0	73	15.3835616
5	13-14	Ajinca, Alexis	nor	C	56	952	136	250	0	1	56	67	94	277	40	23	63	46	187	3	328	0	0	0	30	9.3928571
6	13-14	Aldrich, Cole	nyk	C	46	336	33	61	0	0	26	30	37	129	14	8	18	30	40	0	92	0	0	0	2	4.8478261
7	13-14	Aldridge, Lamarcu	por	PF	69	2496	652	1423	3	15	296	360	166	766	178	64	122	69	147	1	1603	2	0	0	69	24.9710145
8	13-14	Allen, Lavoy	ind	PF	65	1068	134	299	2	13	33	50	119	311	71	24	45	33	126	1	303	0	0	0	2	7.9230769
9	13-14	Allen, Ray	mia	SG	73	1937	240	543	116	309	105	116	23	205	143	54	83	8	115	0	701	0	0	0	9	9.7808219
10	13-14	Allen, Tony	mem	SG	55	1276	204	413	11	47	76	121	79	208	94	90	90	19	121	0	495	2	0	0	28	10.2181818
11	13-14	Aminu, Al	nor	SF	80	2044	234	494	13	48	91	137	130	497	114	82	88	37	147	1	572	0	0	0	65	11.3500000
12	13-14	Amundson, Lou	chi	PF	18	176	15	30	0	0	5	20	27	52	5	9	14	10	47	2	35	1	0	0	0	3.7222222
13	13-14	Andersen, Chris	mia	C	72	1396	177	275	3	12	120	169	129	379	19	32	53	97	162	1	477	1	0	0	0	11.1666667
14	13-14	Anderson, Alan	bro	SG	78	1770	194	485	84	248	92	118	40	175	80	48	62	11	147	0	564	2	0	0	26	6.3974359
15	13-14	Antetokounmpo, Gi	mil	SG	77	1899	173	418	41	118	138	202	78	339	149	60	122	61	173	2	525	2	0	0	23	9.1298701
16	13-14	Anthony, Carmelo	nyk	SF	77	2981	743	1644	167	415	459	541	144	621	239	96	198	51	224	0	2112	11	0	0	77	25.1688312
17	13-14	Anthony, Joel	bos	C	33	190	12	32	0	0	4	8	15	38	2	3	3	12	17	0	28	0	0	0	0	1.6969697

Figure 2: The resulting NBA data set with almost 10,000 rows.

NCAA Data

Collecting the NCAA data was a more complicated process which took time but was a great learning experience since I was able to learn a lot about R and its extended libraries. There are R libraries which allow for screen scraping and pulling the HTML code of a page. The NCAA data was pulled from sports-reference.com by examining the list of players, checking for duplicate players and removing them (to assist with removing possible name collisions. While this does limit the data set I felt that this was a better option instead of working around collisions). In order to screen scrape the data I wrote a spider program to iterate through the player list and access their individual page with their college data. Using the R library XML, I was able to grab the data for each player’s college career from the HTML table and correctly parse it. This resulted in a set with over 78,000 rows of data, one for each year a player spent at college.



	Player	Position	Season	School	G	MP	FG	FGA	FG%	2P	2PA	2P%	3P	3PA	3P%	FT	FTA	FT%	ORB	DRB
1	Aaberg, Jordan	F	2009-10	North Dakota State	29	397	46	74	0.622	46	74	0.622	0	0	NA	15	23	0.652	28	55
2	Aaberg, Jordan	F	2011-12	North Dakota State	31	367	46	68	0.676	46	68	0.676	0	0	NA	28	30	0.933	19	62
3	Aaberg, Jordan	F	2012-13	North Dakota State	22	141	24	44	0.545	24	44	0.545	0	0	NA	7	9	0.778	14	26
4	Aaberg, Jordan	F	2013-14	North Dakota State	33	382	51	84	0.607	51	84	0.607	0	0	NA	17	23	0.739	24	37
5	Aaker, Karl	F	2001-02	Portland	11	NA	26	67	0.388	8	24	0.333	18	43	0.419	3	10	0.300	NA	NA
6	Aaker, Karl	F	2002-03	Portland	27	NA	48	141	0.340	18	52	0.346	30	89	0.337	10	17	0.588	NA	NA
7	Aaker, Karl	F	2003-04	Portland	25	NA	24	71	0.338	9	22	0.409	15	49	0.306	7	10	0.700	NA	NA
8	Aaker, Karl	F	2004-05	Portland	29	NA	29	88	0.330	6	16	0.375	23	72	0.319	13	17	0.765	NA	NA
9	Aaron, Carlton	C	2002-03	UMKC	29	NA	115	225	0.511	115	225	0.511	0	0	NA	54	113	0.478	NA	NA
10	Aaron, Carlton	C	2003-04	UMKC	29	NA	85	164	0.518	85	164	0.518	0	0	NA	40	84	0.476	NA	NA
11	Aaron, Carlton	C	2004-05	UMKC	28	NA	166	284	0.585	166	284	0.585	0	0	NA	76	156	0.487	NA	NA
12	Aaron, Jordan	G	2012-13	Milwaukee	32	1140	150	404	0.371	83	210	0.395	67	194	0.345	95	110	0.864	16	99
13	Aaron, Jordan	G	2013-14	Milwaukee	31	1021	138	369	0.374	70	171	0.409	68	198	0.343	112	137	0.818	12	84
14	Aaron, Nate	G	1999-00	Florida International	11	NA	7	19	0.368	2	7	0.286	5	12	0.417	7	8	0.875	NA	NA
15	Aaron, Shaddean	Guard-Forward	2004-05	Mercer	18	NA	16	37	0.432	15	29	0.517	1	8	0.125	8	20	0.400	NA	NA
16	Aaron, Shaddean	Guard-Forward	2005-06	Mercer	28	NA	50	116	0.431	40	87	0.460	10	29	0.345	23	37	0.622	NA	NA
17	Aaron, Shaddean	Guard-Forward	2006-07	Mercer	29	NA	173	356	0.486	150	277	0.542	23	79	0.291	91	118	0.771	NA	NA
18	Aaron, Shaddean	Guard-Forward	2007-08	Mercer	30	NA	168	377	0.446	114	225	0.507	54	152	0.355	91	111	0.820	NA	NA
19	Aaron, Shaqquan	G	NA	Louisville	23	166	11	42	0.262	6	19	0.316	5	23	0.217	2	5	0.400	6	11
20	Aaron, Troy	G	2003-04	Tulane	10	NA	7	23	0.304	5	16	0.313	2	7	0.286	2	2	1.000	NA	NA
21	Aaron, Troy	G	2004-05	McNeese State	23	NA	44	112	0.393	37	84	0.440	7	28	0.250	22	39	0.564	NA	NA

Figure 2. The NCAA data set

Fusion

After collecting the data, the next step was fusing the two sets (NCAA and NBA) together so that they could be compared in a meaningful way. Columns and measures had to be added to the frames and calculated to fill in missing comparisons. I also had to ensure entity resolution (that I could match one player's NCAA stats to their NBA stats). This was made complicated since the two sets had capitalized and formatted the player's names differently; for example, one set had "O'neal, Shaquille" and the other "Shaquille O'Neal". Unnecessary players were then eliminated such that players who went straight from high school or those who went to play internationally would not be included. After reformatting and removing players, I had a resulting list of 1400 players between the years 1989 and 2014. The more recent NBA data was not uploaded at that time; it has since been updated but I felt that this range was acceptable since it would include players who had retired and players still involved in the professional league.

Comparison Issues

After the data was collected and cleansed the process of analysis could begin and I could examine how to approach answering some of my questions. When beginning to look at how ONDs performed related to their peers how could I look at their differences beyond the years spent at college. How could I examine what a OND may have looked like if they were instead a “two-and-done”? There is no clear way to compare how a OND would have done or what a OND’s stats would have looked like if they had stayed more years in college since there is no way to rewind time. The other issue was that players are often categorized by the position they play on the court, mainly center, guard, or forward. I suspected that perhaps college seasoning would help only certain types of players, beyond the types defined by position. A player is much more than their position, though those are starting points to describe their playing style. The solution to separate players not only needed to be able to, in some way, turn back the clock on a player’s history, but also categorize itself into “types” beyond the accepted position categories.

Clustering

Defining Clustering

Clustering allows researchers to group data into smaller subgroups. Each subgroup has elements that are closely related to each other in comparison to the overall set. There are many clustering algorithms that can be used to rank and describe the dissimilarity between two objects and create the subgroups, also called clusters. In this case, I used hierarchical clustering because of its final cluster structure. Hierarchical clustering gives not only a final cluster set based on the desired number of clusters, but it also allows for an overall ranking of every object, putting those objects which are more similar closer together and creating a tree structure to show the clusters.

The tree structure, represented frequently by a dendrogram, shows the relationships between the objects with respect to their similarity. With this structure the overall number of clusters can be altered at any point in time simply by shifting where the tree is cut. When trimming the clusters, one finds the spot on the trunk of the tree where a cut results in the desired number of clusters and all resulting trees are the subgroups. In this case, eight clusters were used for the analysis because that ensured every cluster had at least one OND.

In the example dendrogram below, I have chosen to exhibit four clusters, R has allowed for a representation of this by coloring each cluster in a distinct color such that the objects are colored to match their respective clusters.

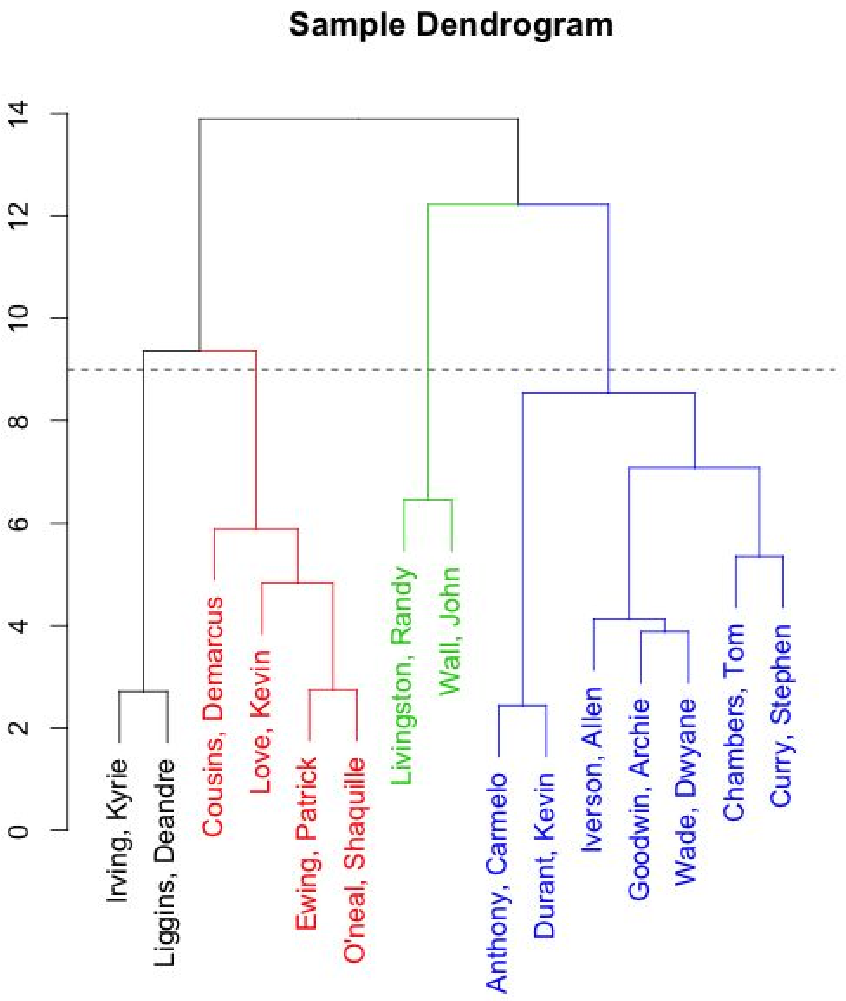


Figure 3: A sample dendrogram with 15 players and 4 clusters

Defining the Distance Metric

When R hierarchically clusters a set of data, it uses a distance metric to compare how close an object, in this case a player, is to another. This distance metric can be predefined as the Euclidian distance or the manhattan distance between the points weighing all statistic measures equally. It can also be defined by multiplying the measures to weight them higher or lower than the others after standardizing the measures themselves. In this way I was able to weight the importance of some statistics in a player's performance were more important than others and vice versa.

To define my distance metric, I talked with two domain experts, a women's' basketball coach and player. These experts were more intimately knowledgeable about the specific statistics that they might use to compare the similarities between two players. For example, they suggested measuring minute per game and weighting that as an important statistic. I also approached defining the metric with my own opinions and intuitive pairings for which players may be similar or dissimilar. To compare how different weights affect the data I chose a small sample of 60 players I was familiar with and changed the distance measure to reflect both sets of opinion. The distance measure even weights all player stats, by changing the weights of stats one at at time I was able to view the changes using tanglegrams to track the progression of players through the clusters. When comparing a change to the distance metric, I actually compared two different dendrograms with a tanglegram.

Tanglegram

A tanglegram is a graph created in R that tracks where objects are in two dendrograms. The graph puts two dendrograms end to end such that a line can be drawn for each object in the trees. When the line is colored, the object is paired with exactly the same player in both dendrograms. The tanglegram also includes the coloring of clusters and the overarching tree structure. In the tanglegram the tree structure itself is compared such that the exact same branches are solid in both trees but dissimilar branches are instead dotted.

Below see an example of a tanglegram with the set of players from Figure 3. In the tanglegram notice how there are two pairs that remain the same in each dendrogram, and while others stay similarly related, the overarching tree structure is altered.

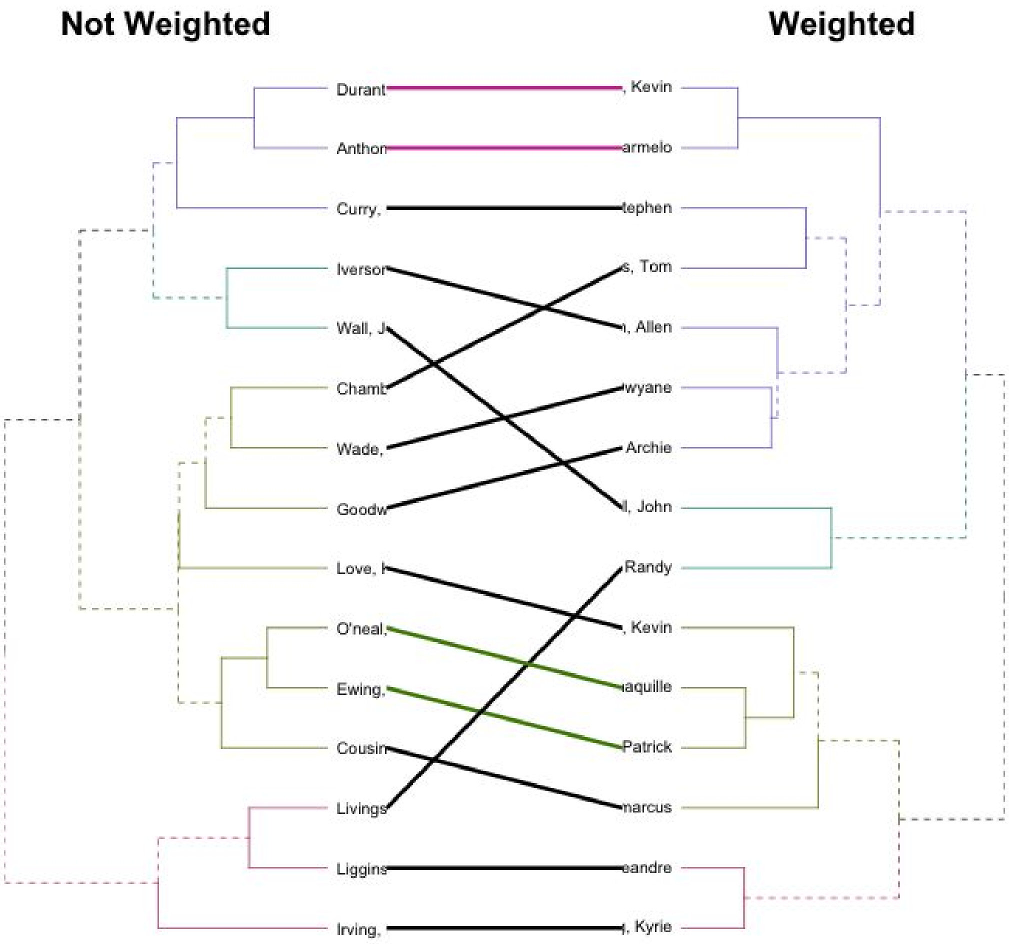


Figure 4: A sample tanglegram

Tracking Changes

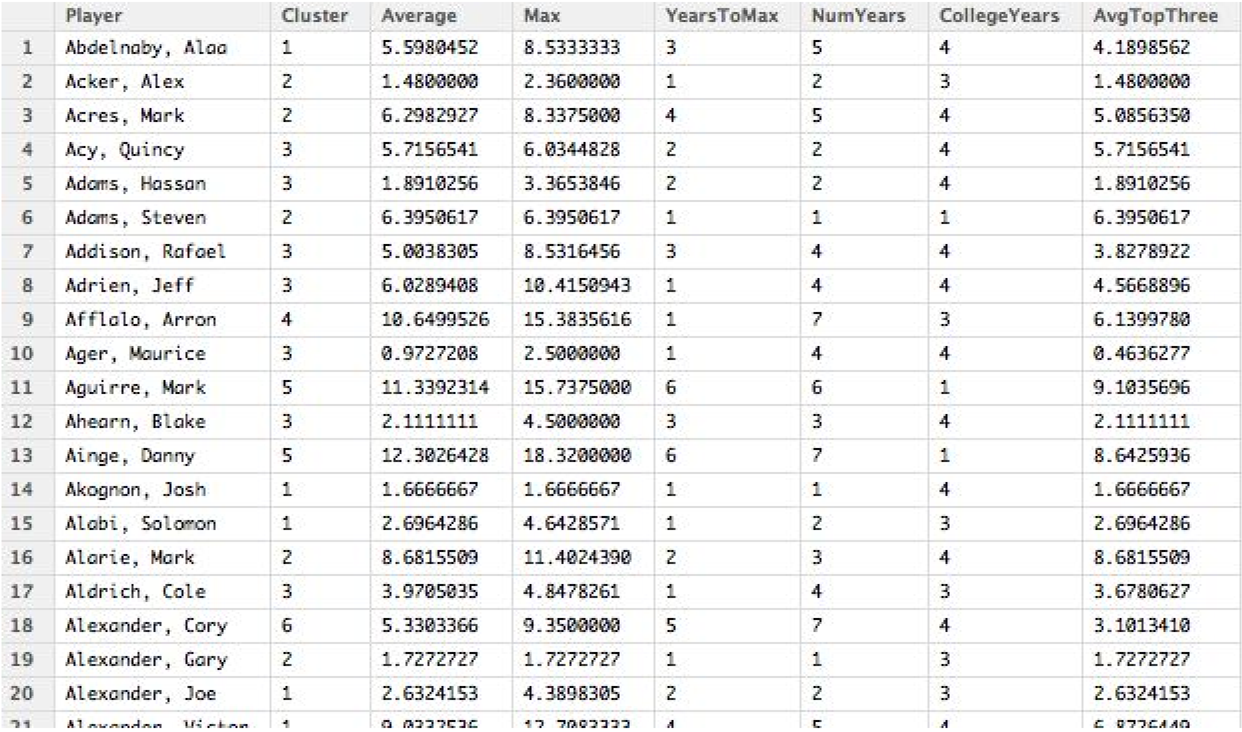
Within the small sample, I changed one stat's weight at a time before deciding whether to keep the change in the accepted metric or not. For example, when discarding the shooting percentages for all players, 27 players changed clusters; this count does not include players who were in the same cluster, it just shifted position in the overall ranking. The next alteration was weighting assists (AST) twice as much, such that assists were more valuable when comparing players. To this end, only 18 players changed clusters; however, the hierarchical ranking shifted in a positive direction to match more closely with expected clusters. In this shift, John Wall and Stephen Curry went from having 3 clusters between them, to only having 2 and Shaquille O'Neal and DeMarcus Cousins moved into the same cluster. Total rebounds (TRB) was similarly added (final weights at 2 times other stats) and changed 12 player positions. The other stats were more complicated to add because their additions made contrary changes. When points (PTS) weight was doubled, 17 players changed; however, the overall groups shifted around in a more favorable position. In an effort to keep the cluster order but get the players more aligned the weight was shifted from twice, to 1.5 times the stats, the result was the same change overall; however, the overall change affected 19 players and certain players were closer to their desired similar players. For example, when weighting PTS twice as much, Shaquille O'Neal and DeMarcus Cousins were in separate clusters; however, when I weighted PTS 1.5 times, they were in the same cluster and close within that cluster - this was a more favorable grouping. Adding the minutes per game (MPG) a player played had a dramatic effect on the clusters - 34 players changed positions. While some players were not in their desired groups yet, their overall rankings were still improving. After shifting MPG weight, it was finally weighted at three times the rest of the stats. The free throw stats (attempts and made shots, FTA and FT) required similar attention. Free throws would be weighted lower than the other stats and in testing whether it would be ½ or ¼ of its weight I discovered that between ½ and ¼ changed 26 players in both weights. Overall there were a few players who moved further from their groups when FT and FTA were weighted at ½ the base. Therefore, the final weightings were:

Statistic	**Weight**
FT, FTA	0.25
FG, 2P, 3P, STL, BLK, TOB, PF	1
PTS	1.5
AST, TRB	2
MPG	3

These final weights in the full 1400 player set made me feel that 8 clusters would be appropriate to examine the ONDs in contrast to MNDs. The clusters were created based on each player's freshman year in order to have a fair comparison between the NCAA players. It also allowed me to say which players were similar at the beginning of their college career and would allow for future examinations of comparisons between the ONDs and a finer granularity in MNDs (for example looking at "two-and-dones").

Analysis

For the research EFF was used to measure “success”, I wanted to have one number to measure a player’s career such that their overall success could be measured. In the future I would want to take a more in depth look at how their college career influence their NBA stats over time, but for now I wanted to just combine into one value. I aggregated a player’s NBA career into different “career metrics” including their max EFF, average EFF and the average of their top three efficiencies (in this case only players who had more than four years in the NBA were considered). Using this data, along with the player's cluster number, I was able to create a new table of the data with their cluster, number of years in college, and different “career metrics”.



	Player	Cluster	Average	Max	YearsToMax	NumYears	CollegeYears	AvgTopThree
1	Abdelnaby, Alaa	1	5.5980452	8.5333333	3	5	4	4.1898562
2	Acker, Alex	2	1.4800000	2.3600000	1	2	3	1.4800000
3	Acres, Mark	2	6.2982927	8.3375000	4	5	4	5.0856350
4	Acy, Quincy	3	5.7156541	6.0344828	2	2	4	5.7156541
5	Adams, Hassan	3	1.8910256	3.3653846	2	2	4	1.8910256
6	Adams, Steven	2	6.3950617	6.3950617	1	1	1	6.3950617
7	Addison, Rafael	3	5.0038305	8.5316456	3	4	4	3.8278922
8	Adrien, Jeff	3	6.0289408	10.4150943	1	4	4	4.5668896
9	Afflalo, Arron	4	10.6499526	15.3835616	1	7	3	6.1399780
10	Ager, Maurice	3	0.9727208	2.5000000	1	4	4	0.4636277
11	Aguirre, Mark	5	11.3392314	15.7375000	6	6	1	9.1035696
12	Ahearn, Blake	3	2.1111111	4.5000000	3	3	4	2.1111111
13	Ainge, Danny	5	12.3026428	18.3200000	6	7	1	8.6425936
14	Akognon, Josh	1	1.6666667	1.6666667	1	1	4	1.6666667
15	Alabi, Solomon	1	2.6964286	4.6428571	1	2	3	2.6964286
16	Alarie, Mark	2	8.6815509	11.4024390	2	3	4	8.6815509
17	Aldrich, Cole	3	3.9705035	4.8478261	1	4	3	3.6780627
18	Alexander, Cory	6	5.3303366	9.3500000	5	7	4	3.1013410
19	Alexander, Gary	2	1.7272727	1.7272727	1	1	3	1.7272727
20	Alexander, Joe	1	2.6324153	4.3898305	2	2	3	2.6324153
21	Alexander, Victor	1	9.0333536	13.7083333	4	5	4	6.8726449

Figure 5: The data set used to compare the EFF measures

Using this data set I was able to compare the ONDs in a cluster to the rest of the cluster and compared all ONDs to the whole set of NCAA players. To evaluate how different OND’s were I used the t.test to examine difference of means and examined the p-value and confidence interval to determine how unalike the two sets were. This allowed for examining each cluster individually and seeing whether each measure of the EFF comparisons showed any difference between the ONDs and MNDs.

Results

In evaluating the eight clusters, only seven were able to be used in the analysis since one ended up with only a single OND. Of the seven clusters, for each career metric, only two showed any significant difference between ONDs and MNDs. Yet the two clusters had conflicting results; in one cluster ONDs fared better, while MNDs fared better in the other group. As seen below, clusters 2, 3, and 4 have large differences; however, cluster 4 did not have enough ONDs to support any conclusion. With cluster 2, the ONDs had a higher career metrics of EFF (max, average, rookie year) and peaked earlier in their careers. Interestingly, the other cluster that showed statistical difference had the opposite result, cluster 3, had lower career metrics for the ONDs when compared to the MNDS. Overall, there was a difference between ONDs and MNDs for all career metrics, regardless of cluster which suggested that ONDs were better NBA players than the MNDs. This makes sense considering that ONDs are normally star players who are considered destined for the NBA since high school, so it should be expected that on average ONDs have better NBA careers than the MNDs since they are expected to be playing at a higher level to start.

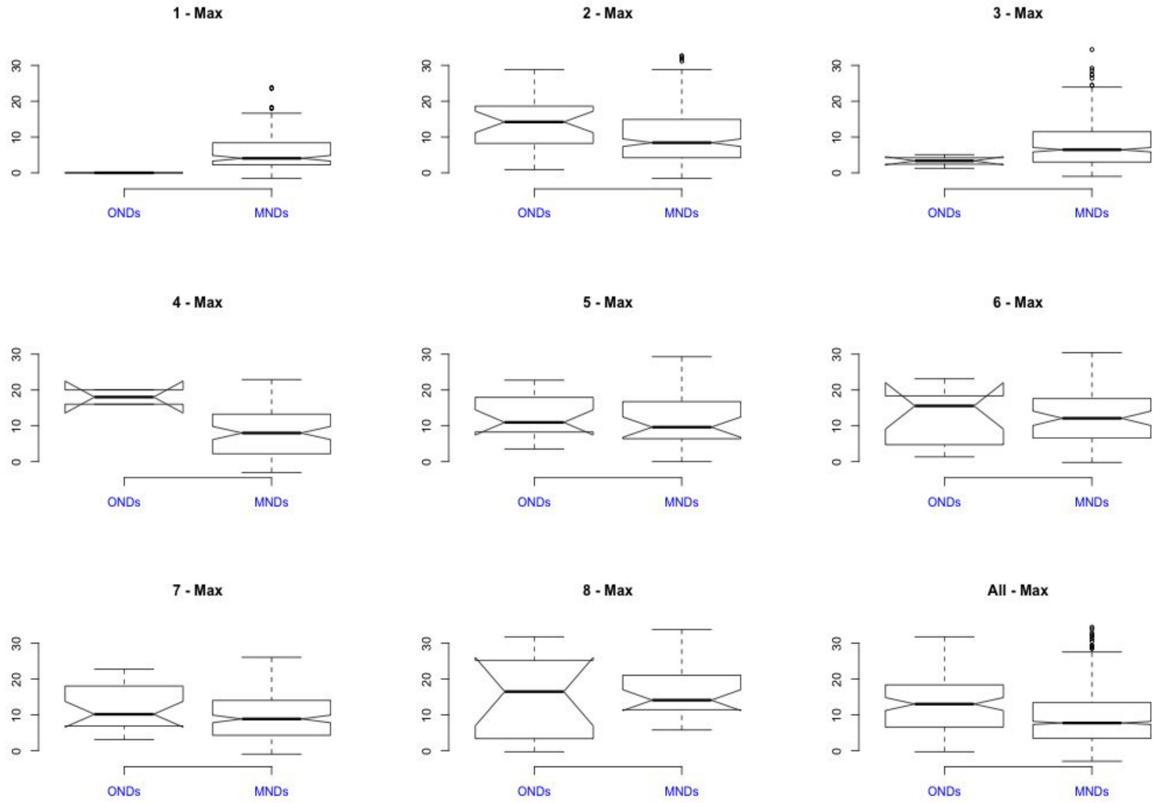


Figure 6: The Max NBA EFF in each cluster comparing ONDs and MNDs.

Simpson’s paradox

The fact that most clusters saw no difference between means while the full set saw a difference is an example of Simpson’s paradox [14]. Simpson’s paradox says that not only is it possible that a statistical trend in subgroups is reversed when all groups are combined, but it is common. In the case of ONDs and their NBA EFF, most clusters did not see a statistical difference; yet the whole set did.

Player Types

While overall ONDs seemed to perform better in the NBA, the type of player you are does seem to make a difference. Depending on the type of player you are, staying at college extra years can help or hurt your career. Only clusters 2 and 3 saw a statistical difference in their means between ONDs and MNDs; players in cluster 2 performed better as ONDs while players in cluster 3 performed worse as ONDs. Players in cluster 2, who performed better as ONDs than other types of players, had lower 3-pointer stats (fewer 3P, 3PA, and lower 3P%), lower free throw percentage, and fewer assists and steals. ONDs from this cluster include Andre Drummond and DeAndre Jordan, while MNDs include Patrick Ewing and Alan Henderson. Players in cluster 3 were similar to cluster 2 but had better 3-pointer stats (more 3P, 3PA, and higher 3P%) and more assists and steals than cluster 2. ONDs from cluster 3 include Josh Selby and Daniel Orton, while MNDs from this cluster include DeAndre Liggins and Quincy Acy.

To summarize: the data suggest that if a player is a big bruiser type (who does not shoot 3-pointers or free throws well, and does not handle the ball particularly well), he is better off leaving college early, but if he is a little better with the 3-pt shooting and a better ball handler, he is better off with the extra seasoning at college.

Future Work

Because I compared the NBA performance based on a career metric, one number representing their whole career, in the future I would examine their career over time. Boiling a player's career and efforts down to a single number necessitates a simplification that is not fair to the players or the game. Future examination of the data would expand to look at not just the efficiency of the NBA career but perhaps examine their whole career to examine whether ONDs had earlier or later peaks in their career.

All analysis was done based on the NCAA player's freshman year, creating a standardized basis for the college data. This was effective for examining the ONDs' differences; however, by using all the players' data this would allow for comparing, as they were, "two-and-dones", "three-and-dones", and so on. In this way one could better evaluate at what point in the college career the extra coaching did not add to the player's performance. By separating each cluster, the effect each additional year had on the players could be further highlighted.

Alternatively, examining the MNDs in each cluster could also allow for some predictive modeling. By knowing what type of player each cluster describes, or even being able to fit new players into the clusters, one could evaluate how much time at college would best benefit an individual player. Since the site used to collect NBA data has been updated with the most recent season, one could theoretically go and examine the NCAA students who were drafted for the 2014-2015 season and see if one could evaluate how successful their rookie season would go and whether or not those students would benefit leaving early or not. This method of evaluation would also allow for additional data for the evaluation of existing NBA players.

While all aggregate stats have their advantages and disadvantages, EFF is a measure that has been improved upon by both WP and PER. By collecting more data for each season the other aggregate measures could be used and compared to see how each measure affects the differences between ONDs and MNDs.

While the goal of this research was to evaluate how a player's statistics were affected by the length of their college career, it should be noted that is not the only thing that affects a player's career. Adding a player's starting salary, age, college team, professional team ranking, etc. could allow for a greater depth of comparison. The draw of a million dollar contract is an easy sell compared to the wait of another year of college.

Conclusion

This research project started because I love math and computer science and as I was preparing to graduate, the data science minor was created and it seemed like the best mix of my two interests. I am also a big Kentucky basketball fan and was curious about ONDs since that particular team seems to have a higher percentage of ONDs than other teams. With Dr. Davies's encouragement and guidance, I figured out how I could combine my love of basketball and data science with some great questions to investigate. The whole thesis process has been a great learning experience, both about the research topic and the skills I needed to do the research.

Using R for the project was an interesting challenge; I had learned the basics in CPSC 219 and from there had to learn on my own and with Dr. Davies's help. This project helped me learn how to use R for gathering data from reading files to scraping data from websites. I also discovered the plethora of libraries R has dealing with statistics and analytics and had a great time learning what the dendrogram libraries were and how to read the graphs in R. Not only do dendrograms and tanglegrams have useful coloring patterns, they also were really helpful in my analysis and a complete foreign concept when I began.

I also learned a lot about basketball in the process, though perhaps not as much as I expected to based on the results. I was so entrenched with basketball when March Madness came this year I threw up my hands because I was already to overwhelmed with basketball. I enjoyed my research but I will be relieved to get to relax on the couch and watch the Wildcats play this season.

Acknowledgements

I could not have made it through this project without:

Stephen Davies, my mentor and advisor, without your support, encouragement, and teaching I would never have picked up my Data Science minor at the last minute. You helped guide me throughout this process between teaching me R and supporting my ideas on how to cluster players.

Jeanne Campbell, your constant reassurances that I could do this will always be appreciated and I will miss our morning talks.

Alex Priest, this paper would be a mess without you and your constant willingness to read and edit it. Your love and support during this have been greatly needed.

Computer Science Department, from the professors who taught and believed in me; to my peers who welcomed and encouraged me. I'm so glad I ended up in this department with you amazing people.

References

[1]

Greg Bianchi, “Age Requirement in Professional Sport,” *The Sport Journal*, 03-Mar-2006. .

[2]

Justin Kubatko, Dean Oliver, Kevin Pelton, and Dan T. Rosenbaum, “A Starting Point for Analyzing Basketball Statistics,” *Journal of Quantitative Analysis in Sports*, vol. 3, no. 3.

[3]

A. Putterman and A. Putterman, “By the numbers: How Kentucky’s one-and-dones have fared in NBA,” *SI.com*. [Online]. Available: http://www.si.com/nba/2015/04/11/karl-anthony-towns-kentucky-draft-one-and-dones-john-calipari-success-anthony-davis. [Accessed: 30-Apr-2016].

[4]

B. Lovell, “Eighteen Years Old and Ready for Driving, Cigarettes and War, but not Basketball: Why the NBA is Committing a Foul on the Age Eligibility Rule,” *J. C.R. & Econ. Dev.*, vol. 26, p. 415, 2013 2011.

[5]

NCAA, Ed., “Estimated Probability of Competing in Athletics Beyond the High School Interscholastic Level.” 24-Sep-2013. [Online]. Available: https://www.ncaa.org/sites/default/files/Probability-of-going-pro-methodology_Update2013.pdf. [Accessed: 30-Apr-2016].

[6]

S. Lohr, “For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights,” *The New York Times*, 17-Aug-2014.

[7]

“Hollinger: What is PER? UPDATED,” *ESPN.com*, 08-Aug-2011. [Online]. Available: http://espn.go.com/nba/columns/story?id=2850240&columnist=hollinger_john. [Accessed: 30-Apr-2016].

[8]

“How to calculate Wins Produced | The Wages of Wins Journal.” [Online]. Available: http://wagesofwins.com/how-to-calculate-wins-produced/. [Accessed:

30-Apr-2016].

[9]

John Feinstein, “It’s time to be done with one-and-done,” *Washington Post, The*, Feb. 2015.

[10]

“NBA Player’s Association,” 27-Feb-2008. [Online]. Available: https://web.archive.org/web/20080227065646/http://www.nbpa.com/cba_articles/article-X.php. [Accessed: 30-Apr-2016].

[11]

J. C. Weber, “One-and-Done: An Academic Tragedy in Three Acts,” *College & University*, vol. 85, no. 2, pp. 57–62, Fall 2009.

[12]

Brian Windhorst, “Silver: Making NBA age limit 20 top priority,” *ESPN.com*, 18-Apr-2014. [Online]. Available: http://espn.go.com/nba/story/_/id/10803355. [Accessed: 30-Apr-2016].

[13]

E. H. Simpson, “The Interpretation of Interaction in Contingency Tables,” *Journal of the Royal Statistical Society*, vol. 13, no. 2, pp. 238–241, 1951.

[14]

D. J. Berri, M. B. Schmidt, and S. L. Brook, *The Wages of Wins: Taking Measure of the Many Myths in Modern Sport*. Stanford, CA: Stanford Business Books, 2007.