

MIT Open Access Articles

The Effect of Basketball Analytics Investment on National Basketball Association (NBA) Team Performance

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Wang, H., Sarker, A., & Hosoi, A. (2025). The Effect of Basketball Analytics Investment on National Basketball Association (NBA) Team Performance. *Journal of Sports Economics*, 26(6), 668-688.

Published Version: <https://doi.org/10.1177/15270025251328264>

Publisher: SAGE Publications

Permanent Link: <https://hdl.handle.net/1721.1/165085>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: <https://creativecommons.org/licenses/by-nc/4.0/>



The Effect of Basketball Analytics Investment on National Basketball Association (NBA) Team Performance

Journal of Sports Economics
2025, Vol. 26(6) 668-688

© The Author(s) 2025



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/15270025251328264

journals.sagepub.com/home/jse



Henry Wang¹ , Arnab Sarker¹ ,
and Anette Hosoi¹

Abstract

In the National Basketball Association (NBA), basketball data and analytics is an area of significant financial investment for all 30 franchises, despite there being little quantitative evidence demonstrating analytics adoption actually improves team-level performance. This study seeks to measure the return on investment of analytics on NBA team success in a time of great demand for analytical front office personnel. Using a two-way fixed effects modeling approach, we identify the causal effect of analytics department headcounts on regular season wins using 12 years of season-level data for each team. We find a positive and statistically significant effect, suggesting clubs that invest more in analytics tend to outperform competitors when controlling for roster characteristics, injuries, difficulty of schedule, and team-specific and time-specific effects. This research contributes to the body of literature affirming the value of data analytics for organizational performance and supports current investments in analytics being made by NBA teams.

Keywords

basketball analytics, econometric analysis, organizational performance, sports management

¹ Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA.

Corresponding Author:

Henry Wang, Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA.

Email: hwang21@mit.edu

Introduction

The implementation of sports analytics is motivated by the fundamental belief that data-driven decision-making improves performance outcomes. This notion has been researched and proven to be true with respect to financial outcomes for for-profit companies (Baijens et al., 2022; Müller et al., 2018; Shabbir & Gardezi, 2020). Despite this, there is limited quantitative evidence of such phenomena in professional sports. Currently, it is unclear if investment in sports analytics helps professional teams win more games, and if so, to what extent.

In the National Basketball Association (NBA), one of the most prominent sports leagues in the world, a key performance indicator (KPI) for every franchise during the regular season is number of wins. With such fierce competition, team owners make significant investments in players, coaches, training staff, facilities, amenities, and more in order to vie for a competitive advantage. For instance, during the 2022–2023 NBA season, clubs spent on average US\$156 million on player salaries alone. However, franchises face financial constraints; in particular, the league-mandated salary cap restricts roster spending, and maximum salary restrictions limit how much a single athlete can be paid. While it is possible to exceed the salary cap by paying a luxury tax, doing so is difficult, especially for smaller market teams. Ultimately, the allocation of resources, particularly toward player personnel, is a costly and high-stakes task with little room for error and naturally demands well-informed decision-making processes.

In the public eye, sports analytics has become a hot topic amongst fans, and there is well-known anecdotal evidence of significant return on investment (ROI) for team performance. The 2011 film *Moneyball* (Miller, 2011), based on Michael Lewis's 2003 book *Moneyball: The Art of Winning an Unfair Game*, chronicles how the 2002 Oakland Athletics, under general manager Billy Beane, utilized baseball analytics, or sabermetrics, to compete against much wealthier teams in Major League Baseball's (MLB's) unequal payroll system. By relying on data-driven decision-making while competitors engaged in traditional and flawed scouting practices, the athletes were able to consistently punch above their weight class, demonstrating how sports analytics can provide a competitive edge. Today, sabermetrics is pervasive across all levels of amateur and professional baseball and has greatly influenced the operations of front offices (Baumer & Zimbalist, 2014).

In tandem, the NBA has not been immune to the 21st-century data revolution. Fast-paced advances in basketball data collection, distribution, and analysis have present-day teams relying more and more heavily on data-driven decision-making. Naturally, the introduction of richer streams of basketball data has created a demand for technical staff with the skills to generate insights and improve performance. In 2024, all 30 teams have at least one employee who specializes in basketball analytics, but the level of adoption varies amongst clubs. Some organizations embrace newer, data-driven approaches in basketball operations, while others

remain more grounded in traditional practice. Because significant financial resources are dedicated to basketball analytics across the association, understanding its ROI is crucial, and measuring a significant ROI would justify league-wide investment and potentially inform each team's resource allocation process.

The goal of this paper is to quantify how investment in basketball analytics impacts win total during the NBA regular season using team-level results spanning the 2009–2010 and 2023–2024 NBA regular seasons. If we can demonstrate analytics indeed provides a competitive advantage, then we have evidence that the rapid adoption amongst professional sports teams is justified. Conversely, if we find analytics investment does not affect, or even hurts team success, then franchises may consider pivoting resources to other channels, such as coaching, player personnel, and facilities, among others. In this study, using analytics department headcount as a measure of investment and adoption, we estimate the effect of headcount on wins while controlling for roster strength, coaching experience, team continuity, and player health. We also include team-specific effects and time-specific effects to account for unobserved time-invariant confounders across teams, as well as trends and NBA-wide policies that affect all clubs. Under this framework, we find a positive and significant analyst effect that remains stable across multiple model specifications and assumptions.

The remainder of the paper proceeds as follows. Section “Literature Review” briefly summarizes the existing body of literature surrounding data analytics and organizational performance, as well as the landscape of basketball analytics. We then describe the data and variables used in our study in Section “Data” and the methods used in Section “Methods.” Next, Section “Results” presents the results from our empirical tests. Finally, Section “Conclusion” discusses conclusions, implications, and limitations, and offers avenues for continued research.

Literature Review

While research into the impact of data analytics on sports team performance is lacking, its positive effects in managerial and business settings are well-documented. The collection of survey data from business executives has allowed researchers to uncover big data analytics's (BDA's) beneficial effects on decision-making and financial performance. A 2010 MIT Sloan Management Review survey of over 3,000 executives across 108 countries and 30 industries found that approximately half of the respondents reported that enhancing data analytics was a priority for their organization, and high-performing companies were five times more likely to be employing data-driven solutions (LaValle et al., 2010). BDA assets improve firm productivity by 3% to 7%, particularly, in information-intensive and/or highly competitive industries (Müller et al., 2018). Additionally, BDA solutions improve the financial performance of businesses through mediating effects of business value and customer satisfaction (Raguseo & Vitari, 2018). The positive impact of BDA on organizational performance and knowledge management in small and medium enterprises is well

established (Baijens et al., 2022; Shabbir & Gardezi, 2020), as well as its utility in real-time resource allocation and asset exchange (Fosso Wamba et al., 2015). Despite these advances, there is a notable gap in empirical research on BDA's added value, particularly in the social sciences (Maroufkhani et al., 2019).

Given BDA is a proven asset in business and management, its potential in sports, and basketball in particular, is a topic of great interest. Chase (2020) argues that in the 21st century, the ability to harness the power of data through artificial intelligence and cloud computing is key to sustaining a competitive advantage in sports. Beyond basketball, sports such as professional soccer and baseball are already embracing analytics tools (Herberger & Litke, 2021). The wealth of historical sports data has made analytics a critical part of research & development processes. Teams across sports have been leveraging analytics for measuring the value of game states and actions, projecting win probabilities, and measuring team and player strength, among other applications (Baumer et al., 2023).

Basketball has been near the forefront of the sports analytics movement. In 2013, the NBA adopted SportVU technology, installing camera systems capable of “quantifying and recording unprecedented basketball data” in every stadium (Richman, 2013). The service, provided by technology company Stats Perform (previously STATS LLC), produces a live feed of player and ball-tracking metrics that can be analyzed by teams. Since this introduction of spatiotemporal data, basketball analytics has blossomed with novel avenues of research. Tracking data can be used to evaluate player decision-making using concepts such as expected possession value (Cervone et al., 2014; Jutamulia, 2021). Wearables and motion capture technology are helping teams manage injury risk (Bishop, 2023), and data science techniques are helping trainers and physicians understand how specific injuries hurt performance (Sarlis et al., 2021). Innovations in machine learning and artificial intelligence, such as the automation of basketball play classification through neural networks, have streamlined game planning and scouting processes for coaches (Markovic et al., 2020; Wang & Zemel, 2016). The rise of player and ball-tracking technologies in college basketball is enhancing talent identification and informing NBA draft strategies (Patton et al., 2021). Most recently, a new partnership between the NBA and Hawk-Eye Innovations brought cutting-edge skeletal data starting in the 2023–2024 season (NBA and Sony's Hawk-Eye Innovations launch strategic partnership powering next generation tracking technology, 2023). These advances underscore the profound impact of basketball analytics on the way the game is studied and played, from strategy and talent identification to injury reduction and performance optimization. However, despite the impressive breadth of research in advancing sports analytics methodologies, there remains a hole in the literature on how such methodologies actually impact team-level KPIs. While modern-day sports analytics is impressive, its value remains unproven.

Notably, limited previous research on the topic fails to find a causal link between sports analytics investment and team outcomes, but this conclusion comes with limitations. Examination of the four major US sports leagues found that franchises that

have adopted analytics show no competitive advantage (Freeman, 2016). The analysis used categories of analytics adoption published by the Entertainment and Sports Programming Network (ESPN) from the MLB, National Football League, NBA, and National Hockey League for the 2014 season. However, this study was a correlational analysis, and because categories were released for just the 2014 season, the sample is limited to only one year. Another study done on MLB teams from 2014 to 2017 uses ESPN-published categorical analytics adoption groups, as well as research staff headcounts from Baseball America Directories (Chu & Wang, 2019). However, this study does not use an econometric framework and relies on empirical conditional distributions, correlations, simple linear regressions, and decision trees. It finds being a “believer” in analytics is moderately positively correlated and statistically significant with wins for all seasons. However, when controlling for team payroll, multiple regressions show a positive and significant effect for research staff in only 2015. The authors do not control for any other confounding factors across MLB clubs beyond payroll. Additionally, decision trees for predictive classification tasks, unless carefully constructed, do not encode causal relations (Li et al., 2016). Nevertheless, this study demonstrates there is reason to believe an analytics effect may exist in MLB.

Although basketball analytics is pervasive in the modern professional game, we have yet to quantify its long-term effect on NBA teams. This paper seeks to address the lack of conclusive, empirical research on analytics in sports at a time when investment is at a historical high. Given the rapid spread of data analytics in basketball and existing evidence of BDA’s positive impact on firms in business contexts, we seek to determine whether a similar phenomenon exists in the NBA using an econometric approach and several years of data.

Data

Data Collection

All necessary datasets used in this study were publicly available. NBA teams are required to disclose player salaries, which can be found on www.HoopsHype.com. Season results (wins, losses, offensive metrics, defensive metrics, etc.) were obtained from ESPN at www.ESPN.com, and roster data was obtained from www.Basketball-Reference.com. Injury data, which teams must disclose on the NBA Injury Report (IR), was obtained from www.ProSportsTransactions.com. Finally, information about each team’s investment in basketball analytics was acquired from www.NBAStuffer.com, which maintains a head count of each franchise’s analytics department. The NBAStuffer website is dynamically updated, meaning data from previous seasons is not easily accessible, so the Internet Wayback Machine from www.web.archive.org was used to collect data from previous years. If multiple archives were created during a given calendar year, the archive logged closest to the date of the beginning of the regular season was used.

It should be noted that during the time period under consideration, the New Jersey Nets moved to Brooklyn, and the Charlotte Bobcats and New Orleans Hornets rebranded as the Charlotte Hornets and New Orleans Pelicans, respectively. To maintain consistency, the identity of these franchises remained consistent throughout the analysis. That is, all New Jersey Nets season-level observations are in the same group as all Brooklyn Nets observations and so on; we do not consider them as two different organizations.

Both manual annotation and automated scripts were used to scrape online data and merge the various sources into one table. Ultimately, we create a panel dataset consisting of season-level observations for each team, making the unit of analysis a team season. For each of these team seasons, we note the basketball analytics department headcount, relevant time-varying covariates, and team-level performance metrics. The exact variables are defined in the next subsection.

Variables of Interest

Table 1 defines the variables of interest for our experiment. In this section, we discuss the rationale behind the inclusion of each one.

For this study, we require a measure of investment in basketball analytics. Ideally, every franchise would disclose how much money was spent on basketball analytics

Table 1. Variable Definitions.

Variable name	Definition
Independent variables	
<i>Analysts</i>	Number of basketball analytics staff members on the team.
<i>Roster Salary</i>	Sum of the roster's individual player salaries in millions of USD, adjusted for inflation.
<i>Roster Experience</i>	Number of years previously spent in the NBA averaged over each player on the roster.
<i>New Coach</i>	Dummy variable indicating whether or not the team started the season with a new coach or hired a new coach during the season.
<i>Roster Continuity</i>	Percent of a team's regular season minutes that were filled by players from the previous season's roster.
<i>Coach Experience</i>	Number of years previously spent in the NBA by the head coach. If coaching changes were made during the regular season, the average experience across coaches was used.
<i>Player-Games Injured</i>	Total number of player games missed due to placement on the daily NBA injury report over the entire regular season.
<i>Road B2Bs</i>	Number of instances where a team plays away games on consecutive days.
Dependent variables	
<i>Wins</i>	Number of games won during the regular season.
$\text{logit}(p_{\text{win}})$	The logit function evaluated at the win percentage p_{win} .

Note. NBA = National Basketball Association; B2B = back-to-back.

staff and resources each season, but this data is not publicly available. Instead, we use analytics department headcounts as a proxy under the assumption that more analytics personnel is an indicator of greater investment. We note that using a headcount metric is subject to limitations, which we detail in Section “Conclusion.” However, we find a similar approach is used in the business management literature, where human IT assets or the number of IT employees is used to analyze the effect of IT investment on firm performance (Sabherwal & Jeyaraj, 2015). Thus, the independent variable of interest is *Analysts*, which represents the observed analytics department headcount for each team season. According to NBAStuffer, headcounts were sourced by checking X (formerly Twitter) and LinkedIn profiles, annual media guides, front office staff directories, press releases and news, and communications with NBA insiders. It is important to note executives and nominally nontechnical individuals are considered as analytics staff by NBAStuffer if they are known to have an analytical background or embrace data-driven approaches in basketball operations. For example, Daryl Morey, current Philadelphia 76ers president and known basketball analytics aficionado, is included in the 76ers’ analytics department headcount. We recognize that these choices may be prone to subjectivity, but presently represent the most suitable proxy available to the authors.

The dependent variable is team performance, which we measure using the *Wins* variable, which denotes each team’s regular season win total. *Wins* is each team’s KPI assuming the ultimate goal of every club is to win as many games as possible. We exclude postseason wins for consistency since not all teams qualify.

To mitigate omitted variable bias, we collected several time-varying covariates to account for differences in observable team-level variables. These controls can be bucketed into two categories: (1) roster controls and (2) player health and fatigue controls. For category 1, the first control is *Roster Salary*, which is inflation-adjusted and calculated by summing each player’s earnings for the given season. This can be thought of as a weak proxy for roster strength because better players command higher salaries, and wealthier teams can afford better players and more analytics staff, both of which may increase winning odds. Next, we control for *Roster Experience*, calculated as the mean number of years each team member has previously played in the NBA. Intuitively, we expect more experienced teams to perform better than their younger counterparts. We note that we do not account for a turnover during past seasons due to trades and/or free agency as rosters were scraped from Basketball-Reference in January 2024. Next, because the quality of coaching may be an important determinant of team success, *Coach Experience*, the number of seasons of experience the head coach of the team possessed, was added. It is important to mention teams sometimes make coaching changes in the middle of the season. When this occurred, we took the average experience of all coaches who were active for that team season.

We note that an important factor of performance in sports is team coordination or team chemistry (Araújo & Davids, 2016; Eccles & Tenenbaum, 2004). Because coaching changes can interrupt the continuity, morale, and chemistry of a team,

we include a dummy *New Coach* that indicates whether the team had a new head coach or experienced a mid-season firing. Additionally, we control for *Roster Continuity*, defined as the percent of a team's regular season minutes that were filled by players from the previous season's roster. This serves as a proxy for team chemistry, which reflects the degree to which team members interact positively and effectively on and off the court, and how familiar teammates are with each other's play styles. Clubs with greater chemistry are thought to exhibit more teamwork, which could yield a competitive advantage.

For category 2, we include *Player-Games Injured*, which is the total number of times a team member was placed on the daily NBA IR for one game. Teams that suffer more injuries are unable to utilize their best players, which impairs winning chances. Next, we add the number of road back-to-backs played by the team that season, denoted as *Road B2Bs*. Playing consecutive games away from home is notorious for being demanding on the body because it entails shortened recovery windows and irregular sleep patterns due to travel and time zone differences, which are all factors that impact how fresh players are at tip-off and hurt performance (McHill & Chinoy, 2020).

A potential concern with our variable selection is how to disentangle the effects of analytics from *Roster Salary*, which may be viewed as a direct reflection of roster strength. While analytics staff contribute to building a strong roster, this objective does not necessarily align with achieving a high *Roster Salary*. Analytics is particularly effective at identifying undervalued metrics and players, optimizing roster performance relative to cost (Gavião et al., 2020; Harrison & Salmon, 2024; Li, 2021). In contrast, *Roster Salary* is largely driven by external factors, such as ownership priorities, market size, and broader economic conditions, which are beyond the control of analytics staff. To evaluate potential overlap, we conducted a variance inflation factor (VIF) analysis. The VIF values for *Analysts* (1.62) and *Roster Salary* (2.20) were below the threshold of 5, suggesting these variables capture distinct dimensions of team outcomes. Ultimately, we argue *Roster Salary* reflects financial and market-driven factors, while the *Analysts* measures internal capacity for decision-making and strategy, justifying the inclusion of both variables in our model.

Sample Restrictions and Characteristics

The sample size was primarily limited by the availability of information on analytics staff, which spans from the 2009–2010 NBA regular season to the 2023–2024 season, with the exception of 2018–2019. We excluded the anomalous 2011–2012 season, which was shortened due to a player lockout. The final balanced panel dataset consists of 12 seasons of data for all 30 teams, yielding a sample size of $N = 360$. We ignored results from playoff games since only a subset of teams participate.

Table 2 reports the descriptive statistics of the sample. By definition, teams win half of their games on average; the reason the mean *Wins* value is 40.11 and not

Table 2. Descriptive statistics ($N = 360$).

Variable	Mean	SD	Min	Max
<i>Wins</i>	40.11	12.24	10	73
<i>Analysts</i>	2.24	2.01	0	10
<i>Roster Salary (US\$, million)</i>	119.52	31.30	63.37	202.51
<i>Roster Experience</i>	4.53	1.34	1.65	8.65
<i>Coach Experience</i>	6.71	6.24	0	30
<i>New Coach</i>	0.13	0.34	0	1
<i>Roster Continuity</i>	63.69	15.73	14.00	98.00
<i>Player-Games Injured</i>	26.25	24.45	0	211
<i>Road B2Bs</i>	10.63	4.1	1	20

Note. B2B = back-to-back.

41 is that during the 2019–2020 season, the schedule was shortened from 82 games due to COVID-19. Additionally, the Boston Celtics and Indiana Pacers only played 81 games during the 2012–2013 season due to a canceled contest following the tragedy at the Boston Marathon. For roster salaries, clubs pay their players nearly US\$120 million each season, although this figure has trended upward in recent years even when controlling for inflation. The mean Roster Experience is 4.5 years and coaches have between six to seven years of experience on average. For 13% of observations, the team had a new head coach. The average Roster Continuity is 63.7% and 26.25 is the mean number of games missed by players due to injury per team. Finally, the average club must play around 11 road back-to-backs each year. Notably, the number of club technical staff has increased substantially over time as shown in the comparative box plot found in Figure 1. At the start of the 2009–2010 NBA season, only 11 analysts existed across all teams. In October 2022, the league boasted a technical staff count of 132, more than a 10-fold increase. In the time window of the data, the average annual growth rate of basketball analytics headcount was 29%.

Methods

We employed various econometric approaches to estimate the causal effect of analysts on win total. We begin with an ordinary least squares (OLS) regression with controls and robust standard errors. Next, we use two-way fixed effects models to capture team-specific and time-specific effects on both *Wins* and $\text{logit}(p_{\text{win}}) = \log[p_{\text{win}}/(1 - p_{\text{win}})]$, where $p_{\text{win}} \in [0, 1]$ is the win percentage. To account for within-team correlations and heteroskedasticity, team-level clustered standard errors were used across all fixed effects models for reliable significance tests. We also experimented with nonlinear regression using quadratic terms for *Analysts*, *AvgExperience*, and *CoachExperience* but found the nonlinear coefficients

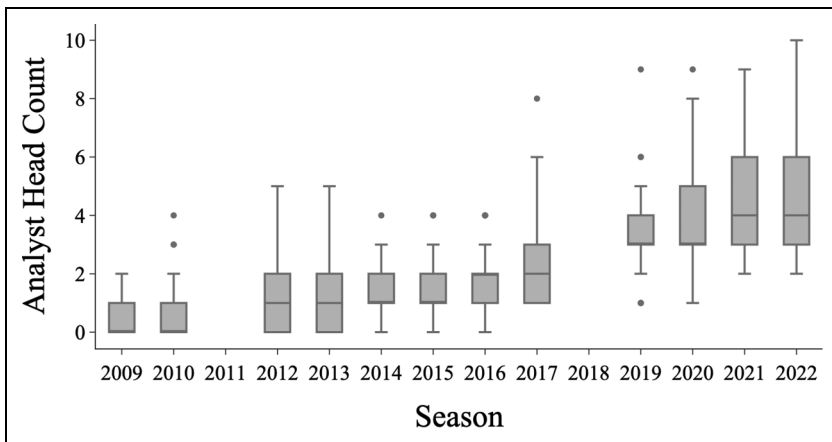


Figure 1. Box plots of NBA analytics department headcounts over time according to NBAStuffer. The years 2011 and 2018 are omitted owing to the lockout in 2011 and missing data in 2018. The median headcount is monotonically increasing from 2009 to 2022. Note. NBA = National Basketball Association.

to not be statistically significant, prompting us to use the simpler linear specification. In addition, we estimated a model that used total back-to-backs instead of road back-to-backs as a measure of schedule difficulty, but this did not affect our results. Finally, we experimented with an ordered probit model with outcomes for failing to qualify for the playoffs, making the playoffs, winning the Eastern or Western Conference, and winning the NBA Finals. However, this failed to yield any significant results, potentially due to the relatively small number of playoff games and reduced heterogeneity of observable outcomes.

The inclusion of team and time-fixed effects is a crucial part of our analysis. Team fixed effects, α_i , capture unobserved heterogeneity across teams that may influence both wins and our covariates. For example, historical team success, market size, and long-standing organizational culture can affect both an organization's attitude and approach toward analytics and its win totals. These fixed effects allow us to separate the time-invariant, unique characteristics of each team from the true impact of analytics.

Additionally, time-fixed effects, γ_t , address variations over time that could bias results if omitted. Various league-wide changes occur over time and impact all clubs. We previously mentioned the 2019 and 2020 seasons were affected by COVID-19, which our time-fixed effects capture by allowing for differences in season length, scheduling disruptions, and other league-wide changes during that period to be accounted for in our analysis. Other examples include fluctuations in the salary cap, league scheduling policies, and rule changes. Coaching strategies and playing styles also evolve and impact the nature of the game. Updates to the

collective bargaining agreement between the NBA Player's Association and the league alter salary structures and free agency rules. By including γ_t , we can account for these league-wide trends and changes that may impact the relationship between our independent and outcome variables.

In summary, we estimate parameters in the following four models, where α_i are team-specific effects and γ_t are time-specific effects. We start with a simple OLS regression in model (1) with category 1 controls, which are *RosterSalary*, *AvgExperience*, *CoachExperience*, *NewCoach*, and *RosterContinuity*.

$$\begin{aligned} Wins_{i,t} = & \beta_0 + \beta_1 Analysts_{i,t} + \beta_2 RosterSalary_{i,t} + \beta_3 AvgExperience_{i,t} \\ & + \beta_4 CoachExperience_{i,t} + \beta_5 NewCoach_{i,t} + \beta_6 RosterContinuity_{i,t} + \epsilon_{i,t} \end{aligned} \quad (1)$$

In model (2), we introduce team and time-fixed effects.

$$\begin{aligned} Wins_{i,t} = & \beta_0 + \beta_1 Analysts_{i,t} + \beta_2 RosterSalary_{i,t} + \beta_3 AvgExperience_{i,t} \\ & + \beta_4 CoachExperience_{i,t} + \beta_5 NewCoach_{i,t} + \beta_6 RosterContinuity_{i,t} \\ & + \alpha_i + \gamma_t + \epsilon_{i,t} \end{aligned} \quad (2)$$

Next, we include the category 2 controls, *Player-Games Injured* and *RoadB2Bs*, in model (3) to control for factors impacting player health and performance.

$$\begin{aligned} Wins_{i,t} = & \beta_0 + \beta_1 Analysts_{i,t} + \beta_2 RosterSalary_{i,t} + \beta_3 AvgExperience_{i,t} \\ & + \beta_4 CoachExperience_{i,t} + \beta_5 NewCoach_{i,t} + \beta_6 RosterContinuity_{i,t} \\ & + \beta_7 PlayerGamesInjured_{i,t} + \beta_8 RoadB2Bs_{i,t} + \alpha_i + \gamma_t + \epsilon_{i,t} \end{aligned} \quad (3)$$

Finally, model 4 allows for nonlinearity by using $\text{logit}(p_{\text{win}})$ as the dependent variable.

$$\begin{aligned} \text{logit}(p_{\text{win}})_{i,t} = & \beta_0 + \beta_1 Analysts_{i,t} + \beta_2 RosterSalary_{i,t} + \beta_3 AvgExperience_{i,t} \\ & + \beta_4 CoachExperience_{i,t} + \beta_5 NewCoach_{i,t} + \beta_6 RosterContinuity_{i,t} \\ & + \beta_7 PlayerGamesInjured_{i,t} + \beta_8 RoadB2Bs_{i,t} + \alpha_i + \gamma_t + \epsilon_{i,t} \end{aligned} \quad (4)$$

For all regressions, the coefficient of interest is β_1 , which represents the effect of an additional data analyst on the dependent variable. The error term for team i in season t is notated by $\epsilon_{i,t}$. For a discussion on modeling assumptions and validity, please see Appendix A.

Results

Model Estimates

Table 3 reports estimates for the effect of *Analysts* on *Wins* and $\text{logit}(p_{\text{win}})$ for all four regressions, along with *R*-squared values and *F*-statistics to test the joint significance of our estimates. Under the simple OLS model, there is no clear relationship between the analytics headcount and win total, highlighting the importance of our fixed effects approach. However, the effect is positive and statistically significant at the 5% level for all fixed effects models (regressions 2 through 4). Additionally, we find all *F*-statistics reject the null at the 0.01% level, indicating coefficients are jointly significant. Model (2), which only controls for roster characteristics, suggests an additional analyst is worth around 1.19 wins. When adding controls for player health and fatigue in model (3), the effect remains similar at 1.25 wins. Finally, the nonlinear model reveals a positive and statistically significant effect of 0.067, suggesting the presence of diminishing marginal returns in headcount.

Table 3. Regression Results.

	(1) Wins	(2) Wins	(3) Wins	(4) $\text{logit}(p_{\text{win}})$
<i>Analysts</i>	0.501 (0.306)	1.185* (0.569)	1.253* (0.546)	0.067* (0.030)
<i>Roster Salary</i> (US\$ million)	0.021 (0.017)	0.112* (0.052)	0.108* (0.052)	0.005 (0.003)
<i>Roster Experience</i>	3.838** (0.383)	3.933** (0.482)	3.822** (0.478)	0.206** (0.027)
<i>Coach Experience</i>	0.016 (0.087)	0.054 (0.122)	0.180 (0.099)	0.009 (0.005)
<i>New Coach</i>	-1.408 (1.584)	-2.658* (1.345)	-2.373 (1.493)	-0.123 (0.083)
<i>Roster Continuity</i>	0.249** (0.030)	0.176** (0.041)	0.149** (0.039)	0.008** (0.002)
<i>Player-Games Injured</i>	-0.066** (.019)		-0.118** (0.023)	-0.006** (0.001)
<i>Road B2Bs</i>	0.476** (.123)		0.215 (0.231)	0.011 (0.012)
Time-Fixed Effects?	No	Yes	Yes	Yes
Team-Fixed Effects?	No	Yes	Yes	Yes
Observations	360	360	360	360
<i>R</i> -squared	0.95	0.40	0.45	0.44
<i>F</i> -statistic	7658.5	35.305	31.798	31.105
Prob > <i>F</i>	0.000	0.000	0.000	0.000

Note. * $p < .05$, ** $p < .01$. B2B = back-to-back.

Robust, clustered standard errors are in parentheses. Each column shows a regression, with the dependent variable in the top row. The *Analysts* effect is significant at the 0.05 level in all fixed effects models.

In summary, the coefficient on the number of analysts in our linear fixed effects models remains stable between 1.1 and 1.3. Implications and practical takeaways are further discussed in Section “Conclusion,” where we discuss results from models (3) and (4).

Conclusion

From the results of our fixed effect models, we conclude the effect of analytics department headcount on regular season win total is positive and statistically significant for NBA teams. Interpreting the coefficients of model (3) from Table 3 yields a number of insights. First, the directions of all statistically significant estimates are consistent with intuition. More technical analysts, more expensive players, more experienced personnel, and greater team chemistry all improve winning chances as indicated by the positive regression coefficients. The negative coefficient on *Player-Games Injured* demonstrates higher injury frequency decreases wins as expected. Next, we examine the absolute and relative magnitudes of our estimates. In model (3), we find that one analyst is worth an additional 1.25 wins. While one game in an 82-game season may seem insignificant, it can be the determining factor for making the playoffs or earning home-court advantage. Also, given the *Roster Salary* coefficient, $\hat{\beta}_{RosterSalary} = 0.108$, we can estimate the cost of an additional win to be approximately US\$9.3 million in Roster Salary. This suggests that, in the current environment, investing in basketball analytics is a much less costly, yet equally effective tool for improving performance compared to buying players.

While model (3) provides interpretable effects within the observed range of the data, it is important to note that extrapolation beyond the range of 0 to 10 analysts would be invalid. For example, teams cannot achieve undefeated seasons simply by increasing analyst headcount or Roster Salary indefinitely. Additionally, model (4) demonstrates that the *Analysts* variable, while positive and significant, exhibits diminishing marginal returns.

There are a few limitations to this study. The analytics staff data from NBASTuffer is manually collected, meaning there is a potential measurement error in the *Analysts* variable due to human mistakes. Although we have no reason to suspect this, there is a possibility that this error is nonrandom if, for instance, a certain team intentionally conceals their analytics staff directory, or if NBASTuffer systematically under-researches one team. If this is the case, then our estimates may be unreliable. Because the headcount figures are collected from publicly available sources, it is possible there are additional technical personnel that are unaccounted for in the data, which is beyond our control. In addition, there were instances where the most recent analyst headcount archive was several months before the beginning of that year’s regular season, at times even during the previous season. In these cases, we assumed there were no significant changes made to the staff during the off-season.

Notably, excluding the years without recent headcounts (2010, 2012, and 2013) from the sample did not alter the results significantly. With respect to headcounts, we assume more analysts indicate a higher degree of analytics adoption, which may not be true for all teams. It may also not be true that increasing headcount necessarily improves productivity. Other measures, particularly financial data on research and development expenditures for each club, would serve as better indicators of analytics investment, but are not publicly available to researchers. Finally, while our fixed effects capture a multitude of unobserved confounders, there remain many that cannot be included in our models. For instance, we cannot account for day-to-day fluctuations in player fatigue, team morale, or on-court playing conditions, which can all impact team performance.

A design choice we considered when performing our empirical tests was whether or not the time-fixed effects allow for the inclusion of the 2011–2012 lockout season. However, due to the steep reduction in the number of games and uncertainty around how variables relating to player injury and fatigue behave in a shortened competition window, we still decided to omit it. Additionally, there is reason to be concerned about the accuracy of analytics department headcount data during this season, as analytics operations are likely to have been significantly reduced due to the financial constraints and operational disruptions caused by the lockout. Nevertheless, we remark the *Analysts* coefficient remains statistically significant at the 10% level ($p = .059$) when including the 2011–2012 season under the specification of model (3).

As for the next steps, it may be worth exploring mediating effects through structural equation modeling, which could elucidate the mechanisms through which analytics improves team outcomes. Additionally, interactions between analytics investment and other team descriptors may be revealed with larger sample sizes in future studies. Also, there are a variety of more granular performance outcomes beyond wins that are prone to less randomness and may be impacted by basketball analytics such as offensive and defensive ratings, player efficiency ratings, assists-to-turnovers ratio, pace, player health, and more. Studying potential interaction effects with these performance metrics may shed light on areas where analytics is most useful. Because professional sports teams are also business units, similar analyses can be performed using financial outcome variables such as ticketing revenue or net profits. Models from economics or management theory describe how productivity scales with human capital can be incorporated. Finally, the scope of this project was focused on the NBA in the United States, and more research is needed to understand if the same effect exists in other sports and countries. With technology trickling down to college and high school athletics, it would be interesting to learn how the effect varies with skill level, or if it even exists at all beyond the professional ranks.

This work contributes to the growing body of literature uncovering the value of data and information technology for organizational success. Unlike previous research, this study demonstrates that with an econometric approach and sufficient time horizon, the utility of analytics found in business contexts also exists in

professional sports. This finding suggests basketball analytics is a legitimate source of competitive advantage for clubs and has a tangible impact on team success, independent of roster composition, coaching experience, team chemistry, injuries, time-specific effects, and unobserved team differences. For both the NBA and its franchises, the evidence we present supports the significant league-wide investments being made in basketball data, whether it be contracting new data providers or hiring analytics personnel. It is clear that the adoption of basketball analytics is necessary to be competitive, and teams who are slow movers risk being left behind by the competition. Finally, our findings suggest that for sports leagues and sanctioning bodies hoping to foster a competitive and egalitarian landscape, making data and analytical tools available to all franchises and/or members is a worthy initiative.

Acknowledgments

The authors would like to thank NBAStuffer.com for their efforts in collecting and maintaining analytics department headcounts around the NBA. We also thank Professor Anna Mikusheva of the MIT Department of Economics for their advice on this project.


Declaration of Conflicting Interests


The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: For this project, the first author was supported by the MIT Sports Lab Pro Sports Consortium. The second author was supported by the ONR 2016 Vannevar Bush Faculty Fellowship from the Office of the Under Secretary of Defense.

ORCID iDs

Henry Wang  <https://orcid.org/0000-0002-0468-4257>

Arnab Sarker  <https://orcid.org/0000-0003-1680-9421>

References

- NBA and Sony's Hawk-Eye Innovations launch strategic partnership powering next generation tracking technology. 2023. <https://pr.nba.com/nba-sony-hawk-eye-innovations-partnership/>.
- Araújo, D., & Davids, K. (2016). Team synergies in sport: Theory and measures. *Frontiers in Psychology*, 7, 1449. <https://doi.org/10.3389/fpsyg.2016.01449>
- Baijens, J., Helms, R., & Bollen, L. (2022). Data analytics and SMEs: How maturity improves performance. In *2022 IEEE 24th conference on business informatics (CBI)* (vol. 1, pp. 31–39). Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/CBI54897.2022.00011>

- Baumer, B. S., Matthews, G. J., & Nguyen, Q. (2023). Big ideas in sports analytics and statistical tools for their investigation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(6), e1612. <https://doi.org/10.1002/wics.1612>
- Baumer, B., & Zimbalist, A. (2014). *The sabermetric revolution: Assessing the growth of analytics in baseball*. University of Pennsylvania Press.
- Bishop, E. (2023). The science of injury prevention in the NBA: How teams are using technology to keep players healthy. <https://www.sportskeeda.com/basketball/the-science-injury-prevention-nba-how-teams-using-technology-keep-players-healthy>
- Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2014). Predicting points and valuing decisions in real time with NBA optical tracking data. In *Proceedings of the MIT Sloan sports analytics conference*, Boston, MA. <https://www.semanticscholar.org/paper/POINTWISE%3A-Predicting-Points-and-Valuing-Decisions-Cervone-Bornn/f4b381f4482586dbdd15fc92bee81ce68bcb6898>
- Chase, C. (2020). The data revolution: Cloud computing, artificial intelligence, and machine learning in the future of sports. In *21st century sports. Future of business and finance* (pp. 175–189). Springer International Publishing. https://doi.org/10.1007/978-3-030-50801-2_10
- Chu, D. P., & Wang, C. W. (2019). Empirical study on relationship between sports analytics and success in regular season and postseason in Major League Baseball. *Journal of Sports Analytics*, 5(3), 205–222. <https://doi.org/10.3233/JSA-190269>
- Eccles, D. W., & Tenenbaum, G. (2004). Why an expert team is more than a team of experts: A social-cognitive conceptualization of team coordination and communication in sport. *Journal of Sport and Exercise Psychology*, 26(4), 542–560. <https://doi.org/10.1123/jsep.26.4.542>
- Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165, 234–246. <https://doi.org/10.1016/j.ijpe.2014.12.031>
- Freeman, L. A. (2016). The impact of analytics utilization on team performance: Comparisons within and across the US professional sports leagues. *Journal of International Technology and Information Management*, 25(3), 137–160. <https://doi.org/10.58729/1941-6679.1322>
- Gavião, L. O., Sant'Anna, A. P., Lima, G. B. A., & de Almada Garcia, P. A. (2020). Evaluation of soccer players under the Moneyball concept. *Journal of Sports Sciences*, 38(11–12), 1221–1247. <https://doi.org/10.1080/02640414.2019.1702280>
- Harrison, W. K., & Salmon, J. L. (2024). War accumulation differential: A statistic for identifying undervalued players. In J. S. Dong, M. Izadi & Z. Hou (Eds), *Sports analytics* (pp. 60–74). Springer Nature Switzerland.
- Herberger, T. A., & Litke, C. (2021). The impact of big data and sports analytics on professional football: A systematic literature review. In *Digitalization, digital transformation and sustainability in the global economy: Risks and opportunities* (pp. 147–171). Springer International Publishing.
- Jutamulia, I. C. (2021). *Expected possession value: An evaluation framework for decision-making, strategy, and execution in basketball* [Master's thesis, Massachusetts Institute of Technology]. <https://dspace.mit.edu/handle/1721.1/1139205?show=full>

- LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2010). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*. <https://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/>
- Li, Y. (2021). *When Moneyball meets the beautiful game: A predictive analytics approach to exploring key drivers for soccer player valuation* [Master's thesis, Brock University]. <https://dr.library.brocku.ca/handle/10464/15088>
- Li, J., Ma, S., Le, T., Liu, L., & Liu, J. (2016). Causal decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 29(2), 257–271. <https://doi.org/10.1109/TKDE.2016.2619350>
- Markovic, S., Cuk, I., & Živkovic, A. (2020). The impact of information technologies on the scouting process in sports games. In *Sinteza 2020—international scientific conference on information technology and data related research* (pp. 240–245). <https://doi.org/10.15308/Sinteza-2020-240-245>
- Maroufkhani, P., Wagner, R., Wan Ismail, W. K., Baroto, M. B., & Nourani, M. (2019). Big data analytics and firm performance: A systematic review. *Information*, 10(7), 226. <https://www.mdpi.com/2078-2489/10/7/226>; <https://doi.org/10.3390/info10070226>
- McHill, A. W., & Chinoy, E. D. (2020). Utilizing the National Basketball Association's COVID-19 restart “bubble” to uncover the impact of travel and circadian disruption on athletic performance. *Scientific Reports*, 10(1), 21827. <https://doi.org/10.1038/s41598-020-78901-2>
- Miller, B. (2011). *Moneyball* [Film]. <https://www.imdb.com/title/tt1210166/>
- Müller, O., Fay, M., & vom Brocke, J. (2018). The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics. *Journal of Management Information Systems*, 35(2), 488–509. <https://doi.org/10.1080/07421222.2018.1451955>
- Patton, A. N., Scott, M., Walker, N., Ottenwess, A., Power, P., Cherukumudi, A., & Lucey, P. (2021). Predicting NBA talent from enormous amounts of college basketball tracking data. In *Proceedings of the MIT Sloan sports analytics conference (virtual)*. https://www.researchgate.net/publication/354131580_Predicting_NBA_Talent_from_Enormous_Amounts_of_College_Basketball_Tracking_Data
- Raguseo, E., & Vitari, C. (2018). Investments in big data analytics and firm performance: An empirical investigation of direct and mediating effects. *International Journal of Production Research*, 56(15), 5206–5221. <https://doi.org/10.1080/00207543.2018.1427900>
- Richman, P. (2013). NBA 2013-14 season: New SportVU system in arenas to improve player tracking. <https://bleacherreport.com/articles/1764278-nba-to-install-sportvu-system-in-arenas-for-2013-14-and-improve-player-tracking>.
- Sabherwal, R., & Jeyaraj, A. (2015). Information technology impacts on firm performance. *MIS Quarterly*, 39(4), 809–836. <https://doi.org/10.25300/MISQ/2015/39.4.4>
- Sarlis, V., Chatziilias, V., Tjortjis, C., & Mandalidis, D. (2021). A data science approach analysing the impact of injuries on basketball player and team performance. *Information Systems*, 99, 101750. <https://doi.org/10.1016/j.is.2021.101750>
- Shabbir, M. Q., & Gardezi, S. B. W. (2020). Application of big data analytics and organizational performance: The mediating role of knowledge management practices. *Journal of Big Data*, 7(1), 47. <https://doi.org/10.1186/s40537-020-00317-6>

Wang, K. C., & Zemel, R. (2016). Classifying NBA offensive plays using neural networks. In *Proceedings of the MIT Sloan sports analytics conference*, Boston, MA. <https://www.cs.toronto.edu/zemel/documents/1536-Classifying-NBA-Offensive-Plays-Using-Neural-Networks.pdf>

Appendix A

Appendix A.1. Modeling Assumptions

In order to produce a consistent and asymptotically Gaussian estimate of the effect of basketball analytics staff on team performance, we must satisfy the following conditions for all proposed regressions. Let $X_{i,t}$ be the vector of regressors for team i in season t . Let $Y_{i,t}$ be the dependent variable for team i in season t . Finally, $\epsilon_{i,t}$ is the error term for team i in season t .

- (A1) We must have exogeneity. That is, $E[\epsilon_{i,t} | \alpha_i, \gamma_t, X_{i,2009}, \dots, X_{i,2022}] = 0 \forall \text{team } i$.
- (A2) Observations $(X_{i,2009}, \dots, X_{i,2022}, Y_{i,2009}, \dots, Y_{i,2022})$ are independent and identically distributed (i.i.d) across teams.
- (A3) No outliers: Fourth moments of all regressors and errors are bounded.
- (A4) Absence of perfect multicollinearity.

We must consider that using *Wins* as an outcome variable results in a necessary violation of (A2) because winning is a zero-sum result. That is, for any game, one team must win and the other must lose meaning outcomes are not independent. However, we argue this is overcome by the structure of the NBA regular season schedule and the large number of teams and games. Each of the 30 teams plays 82 games every year, for a total of 1230 regular season matches. Additionally, every team plays all other 29 teams, and never plays the same team more than four times. While outcomes are not strictly independent, the impact of this interdependence on the model assumptions is diluted by the volume and distribution of games. In Subsection Appendix A.2, we empirically measure the extent to which our regression adheres to all four modeling assumptions.

Appendix A.2. Empirical Tests of Model Assumptions

In this subsection, we present empirical justification that model (3), our preferred specification, is appropriate for this analysis with respect to the assumptions in Subsection Appendix A.1. The Q-Q plot in Figure 2(a) illustrates residuals are normally distributed as evidenced by the tight alignment of points along the red line. We calculate the Pearson correlation coefficient to be effectively zero ($\rho = 7.1 \times 10^{-17}$; $p = .99$) and observe no correlation between fitted values and

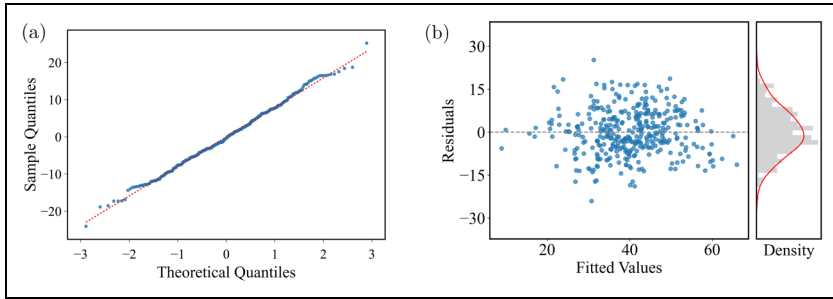


Figure 2. Analysis of residuals for preferred model (3). (a) Q-Q plot for observed residuals of model 3. Markers hug the diagonal line, which suggests residuals are approximately normally distributed, indicating that the assumption of normality required for OLS is reasonably satisfied. (b) Model 3 residuals plotted against fitted values. Markers appear random, indicating that exogeneity is satisfied. The vertical red bell curve represents the Gaussian kernel-density estimate (KDE), accompanied by the underlying empirical distribution of residuals.

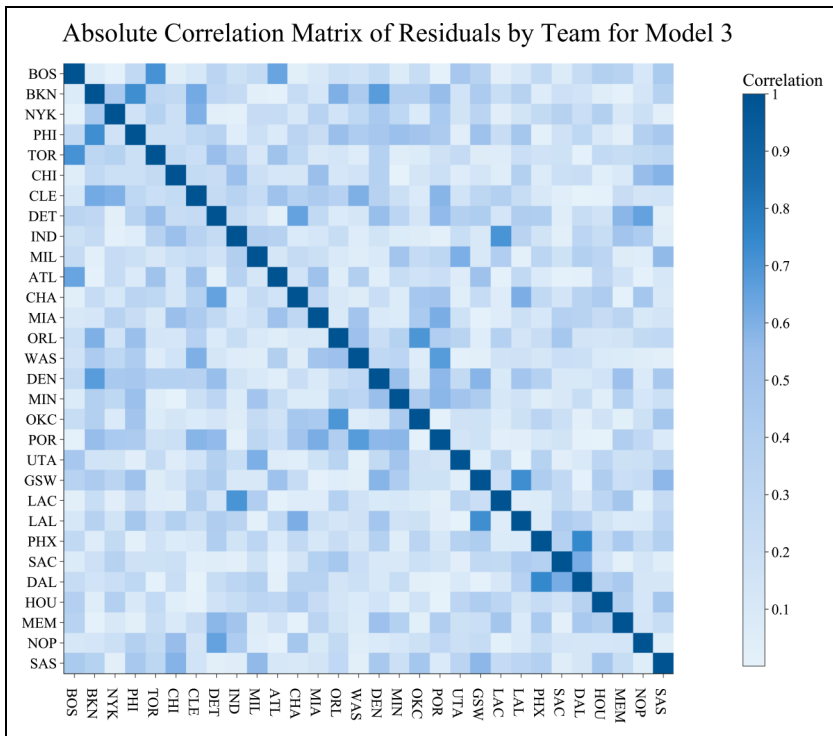


Figure 3. Absolute correlation matrix of team residuals for model (3). Note that correlations adjacent to the diagonal are slightly darker due to axes being grouped by NBA division. However, for the vast majority of pairwise comparisons, residuals are not highly correlated. To emphasize this, note the color scale increases nonlinearly, with a sharper transition rate beginning at 0.55. Note. NBA = National Basketball Association.

residuals in Figure 2(b). In addition, the Pearson correlation coefficient between *Analysts* and the fitted values is 8.7×10^{-17} ($p = .99$). This collection of evidence suggests Assumption 1 is satisfied. We also note the absence of any outliers and perfect multicollinearity, satisfying Assumptions 3 and 4.

Finally, we check Assumption 2, which stipulates team-season observations should be i.i.d. Recall this assumption is necessarily violated because the total number of wins across all teams for a given year must always sum to the number of games played that year, since wins are a zero-sum outcome. Despite this, we assert the high number of teams and games makes the effect of violating this assumption minimal. First, we observe the correlation matrix of residuals by team in Figure 3 exhibits no evidence of highly correlated teams. The axes are ordered by division, so some cells along the diagonal are darker because divisional games occur slightly more often. However, the vast majority of pairwise correlations are weak or near zero.

Algorithm 1 Robustness simulation for necessary violation of A2

```

for  $T \in \{2, 6, 10, \dots, 30\}$  do
  Randomly sample subset of teams  $S$  s.t.  $|S| = T$ 
  Select observations  $\{X_{i,t} : i \in S\}$  from data
  for team  $i \in \{1, 2, \dots, T-1\}$  do
    Predict  $\widetilde{Wins}_{i,t} \leftarrow X_{i,t}^T \widetilde{\beta}$ 
  end for
  Deterministically calculate  $\widetilde{Wins}_T$  for final team  $T$ 
  for simulation  $n \in \{1, \dots, 100\}$  do
    Estimate  $\widetilde{\beta}_{T,n}$  using model (3) specification
    Compute bias w.r.t original estimate  $b(\widetilde{\beta}_{T,n}) \leftarrow \widetilde{\beta}_{T,n} - \widehat{\beta}_{Analysts}$ 
  end for
end for
Observe distribution of  $b(\widetilde{\beta}_{T,i}) \forall T, i$ 

```

Additionally, we run simulations using synthetic data, varying the number of teams competing in an 82-game season between 2 and 30. These simulations, detailed by the pseudocode found in Algorithm 1, show how the distribution of our estimates changes with different levels of dependence among team outcomes. We randomly select T teams without replacement (and thus $T \times 12$ observations because there are 12 years of data) and use the vector of original estimates from model (3), notated here as $\widetilde{\beta}$, to predict a win total for all but one team. For the last team's 12 observations, we deterministically calculate their win totals by subtracting the total number of existing wins in the data from the total number of games played in the hypothetical seasons. This ensures the number of games won equals the number of games played across the entire simulated league schedule. We then use the same model (3) specification to estimate the *Analysts* coefficient on the synthetic data, and over $N = 100$ simulations, observe the distribution of

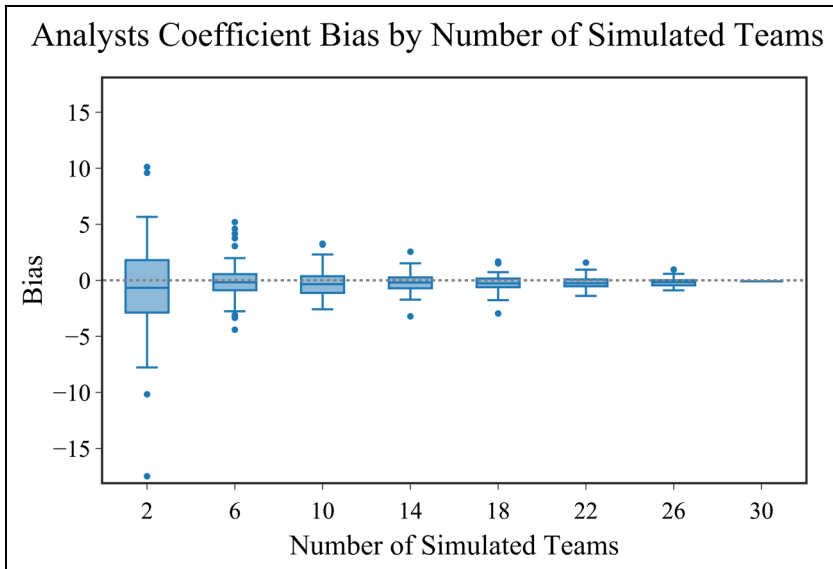


Figure 4. Distribution of the bias of simulated estimates of the *Analysts* effect as the number of simulated teams increases. With a small number of teams, the estimate is unstable because win totals exhibit higher dependence than expected. However, as the number of teams increases to the true value (30), the bias converges to -0.09 . This bias is negative and over an order of magnitude smaller than the estimated β_1 in model (3), suggesting that the estimated effect size is potentially a slight underestimate of the true effect.

these 100 simulated beta values $\tilde{\beta}_{T,1}, \dots, \tilde{\beta}_{T,100}$. We repeat for $T \in \{2, 6, 10, 14, 18, 22, 26, 30\}$. For each $\tilde{\beta}_{T,i}$, we calculate the bias from the true estimate:

$$b(\tilde{\beta}_{T,i}) = \tilde{\beta}_{T,i} - \hat{\beta}_{Analysts}$$

where $\hat{\beta}_{Analysts} = 1.25$ from Table 3. Ultimately, we show in Figure 4 that as the number of teams T increases to 30, the distribution of the bias of the simulated estimate converges to near zero, suggesting 30 teams playing 82 games is enough to dilute any bias the dependence of the *Wins* measure between clubs may have on our regression coefficients. In fact, the bias converges to -0.09 , suggesting that the estimated effect size is potentially a slight underestimate of the true effect.

Ultimately, we argue that the extent to which the data violates Assumption 2 does not affect the validity of the fixed effects regressions, and hence the model is well-specified and produces a consistent and asymptotically Gaussian estimate of the effect of basketball analytics staff on wins. Thus, we assert the significance tests that underlie our conclusions are reliable.