13TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 1 - 2, 2019 | BOSTON, MA
presented by ESPN

How Much Do Coaches Matter?

Christopher R. Berry and Anthony Fowler
University of Chicago

Tracks: Other Sports and Open Source
Paper ID: 12549

1. Abstract

Coaches of professional sports teams are often credited or blamed for the success or failure of their teams, and they are compensated as if they are one of the most important features of a franchise. Although we have anecdotal evidence that coaches matter, the sports analytics literature has generally concluded that they do not. We present a new method for estimating coach effects, which we call Randomization Inference for Leader Effects, or RIFLE. We apply RIFLE to the MLB, NBA, NHL, NFL, college football, and college basketball. We detect coaching effects in all sports. Our estimates generally imply that coaches explain about 20-30 percent of the variation in a team's success, although coaching effects vary notably across settings and across various outcomes. For example, baseball managers affect runs allowed more than runs scored. Coaches matters more in college football than in the NFL, but do not meaningfully differ in their use of rushing vs. passing. In addition to estimating average coaching effects, we also discuss the difficult task of assessing the quality of an individual coach.

2. Introduction

The emergence of *sports analytics* in recent years has transformed the way sports teams are managed and, indeed, the way sports are played. In most organizations, coaches are viewed as consumers of sports analytics.[1] Relatively seldom, however, have coaches themselves been the subject of the sort of rigorous empirical analyses regularly applied to players. As a result, we generally know less, from an analytical perspective, about a coach than about virtually any other person on the field at a sporting event.

The small extant empirical literature on the subject has tended toward the conclusion that all coaches are created equal. The claim is not that teams would be just as well off without a coach, but rather that the market for coaches ensures that they are all of roughly equal quality and employ similar strategies, such that changing coaches does not meaningfully change a team's performance on the field.

We contend that existing studies of coaches suffer from methodological limitations that would make them unlikely to accurately estimate the effects of coaches even where such effects exist. Our main contribution is to develop a new methodology for estimating coach effects and to apply it to a variety of professional and collegiate sports. Our method accounts for player quality and strength of

[1] See Alamar and Mehrota (2011) and Alamar (2013) for an historical overview and current survey of the field of sports analytics.

schedule and it is able to separate variation in team performance related to coaches from variation that would be expected by chance. Contrary to most of the prior literature, we find strong evidence of coaching effects. Our analysis also covers a broader range of sports than any previous study of which we are aware, allowing us to make comparisons across sports and between professional and collegiate levels of the same sport.

The paper proceeds as follows. We first review the prior empirical literature on coaches and highlight some methodological limitations that have hampered those studies. We then present our method for estimating coach effects, which is based on randomization inference. We describe the theory underlying our method and present a series of Monte Carlo simulations to demonstrate its performance. The main part of the paper is an analysis of coaching effects in Major League Baseball (MLB), the National Basketball Association (NBA), the National Hockey League (NHL), the National Football League (NFL), as well as NCAA football (CFB) and basketball (CBB). In each case, we analyze several different outcomes to identify where coaches matter most. We conclude by discussing methodological issues in estimating the effects of individual coaches and, relatedly, ranking coaches relative to one another.

Related Literature

Much of the sports analytics literature on coaches is oriented around questions of *leadership succession*, inspired by the influential early work of Grusky (1960, 1963). Most of the studies in this tradition seek to assess whether a team's performance changes significantly following the replacement of a coach. The general conclusion, across a variety of contexts, is that coaching changes have either no effect on team performance, or a slight negative effect. Notable contributions include Gamson and Scotch (1964), who were among the first to find that MLB managers had little impact on team performance. Several subsequent papers support the conclusion that changing managers does not improve MLB team performance (e.g., Canella & Rowe, 1995; Fabianic, 1994; Smart et al., 2008; Smart & Wolfe, 2003). A related body of research on soccer has produced null findings for the Dutch Premier League (Koning, 2003) and the Italian league (De Paola & Scoppa, 2012).

Other studies have concluded that coaching changes actually result in worse performance for the teams in the NBA (Giambatista, 2004), the NHL (Rowe et al., 2005), English professional soccer (Audas, Dobson, & Goddard, 1997), and college basketball (Fizel and D'Itri (1999).

A handful of studies differentiates within-season versus between-season coaching changes, but generally fails to find effects in either situation. Brown (1982) found that NFL teams perform worse after within-season coaching changes, but that between-season coaching switches had little impact on outcomes. In one of the more comprehensive studies to date, McTeer, White, and Persad (1995) found similarly little effect of within-season coaching changes across four sporting leagues: MLB, NBA, NFL, and NHL.

With so many studies having found null effects of coaching changes overall, others have investigated the conditional effects of coaching changes under different circumstances. Among these, Adler et al. (2013) found for college football that while the performance of the worst teams did not change after a coach replacement, mediocre teams actually performed worse after a coaching change. Also studying college football, Dohrn et al. (2015) found that coaching changes

made little difference for the largest programs (by revenue), while smaller programs experienced a temporary improvement after a coach was replaced.

While the aforementioned studies are concerned with estimating the effects of coaching succession in general or in specific contexts, a few studies have attempted to estimate the effects of individual coaches on team or player performance. Using manager dummy variables for MLB, Bradbury (2017) found that managers in general have little effect on the performance of either hitters or pitchers on their team. He suggests that changing managers may be a tactic used by management to excite the team's fanbase, even if the change matters little for outcomes on the field. Berri et al. (2009) employed coach dummy variables to explain variation in the performance of NBA players. Of the 62 coaches they study, 14 had a statistically significant individual impact, although even among these 14, the individual coach coefficients were indistinguishable from one another. A dissenting study in this literature comes from Goff (2013), who uses a hierarchical model with random effects for coaches to study the NFL and MLB. Goff finds that 8.5% of the variance in team win rates is attributable to managers in MLB, and 21% of the variation is attributable to coaches in the NFL.

The preponderance of null results in the extant literature has led many observers to conclude that sports coaches are largely interchangeable. That is, while coaches may be necessary for teams to function, most coaches seem to perform roughly equally. Summarizing the state of knowledge about coaching for the *Freakonomics* blog, Dave Berri writes that the literature, "suggests that coaches in sports are not very different from each other. It may be true (and more than likely very true) that you are better off with a professional coach than with a random person grabbed from the stands (or no one at all). But it doesn't appear that the choice of professional coach matters much."[2]

In the remainder of the paper, we will argue that the conclusion that coaches are interchangeable is unwarranted. Prior studies suffer from methodological limitations that make them unlikely to accurately measure the effects of coaches even where such effects exist. Studies of coaching succession, in particular, are not well designed to detect the effects of coaches on team outcomes. Such studies are designed to measure the *average* effect of coaching changes. If some coaches are better than others—that is, if coaches are not interchangeable—then we would expect some coaching changes to result in better performance and others to result in worse performance. An average effect of zero could be consistent with great heterogeneity in the effects of coaches. Indeed, we would only expect a significantly positive or significantly negative average effect of coaching transitions if teams reliably improved or reliably reduced the quality of their coaches when making a change. The finding that they don't does not imply that coaches are interchangeable.

Another concern with studies of coaching succession is regression to the mean in team performance. That is, if there is some element of randomness in team performance from season to season, and if a coaching transition is more likely to occur after a particularly bad or particularly good year, then we would expect team performance to change in the following year even if there had been no change in coaching. While some studies explicitly account for mean reversion (e.g., (Adler et al. 2013; Cannella and Rowe, 1995; Koning, 2003), many do not.

[2] "Is Changing the Coach Really the Answer?" Posted 12/21/2012. Last accessed 12/7/2018. http://freakonomics.com/2012/12/21/is-changing-the-coach-really-the-answer/.

Finally, standard studies of coaching succession will reflect any disruptive effects that occur early in the tenure of a new coach. If team performance declines in the first year(s) under a new coach, as players learn the new system, then the immediate effect of changing coaches will reflect not only differences in the quality of the coaches but also these transition effects. Transition effects are especially relevant since most of the studies in this literature examine only the first season after a coaching change.

In other words, the standard studies of changes in team performance following coaching changes will reflect the composite effects of: changes in coaching quality; mean reversion; and transition costs. Finding a null effect on average does not necessarily imply that coaches are interchangeable. Note that we are not necessarily making a critique of the literature on coaching succession, which is designed to measure the average effects of coaching changes per se, not the effects of coaches on team performance overall.

As we will discuss in more detail below, studies that use fixed or random effects for individual coaches are subject to another set of concerns. First, because of the inherent multiple testing involved, we would expect to find some individual coaches to have statistically significant effects just by chance; that is, we should expect to find 5 percent of the individual coach coefficients to be significant, by chance, at a conventional 5-percent level. If there is serial correlation—that is, persistent differences in performance across teams that are unrelated to coaches—then we would expect to find even more individual coach effects to be significant, as individual coach tenures by chance overlap with unrelated periods of good or bad team performance. On the other hand, even if coaches do on average influence team performance, we might fail to find individual coach effects to be significant if there are relatively few observations for each coach. Thus, if our question is whether *coaches* matter, rather than whether some particular coach matters, we would be more interested in the joint significance of all the coach fixed effects. Such an omnibus test, however, will be subject to the same sorts of concerns about false positives due to serial correlation and over-fitting to random noise. In short, examining individual coach fixed (or random) effects using standard inferential strategies is unlikely to answer the question of whether coaches are interchangeable.

In the next section, we describe a new statistical test for coach effects that overcomes many of the problems just described.

3. Methodology

We call our method Randomization Inference for Leader Effects (RIFLE), and we have previously used a similar approach to estimate the effects of political leaders on various economic and policy outcomes (Berry and Fowler 2018). Our goal here is to test whether coaches matter for particular outcomes of interest. We do not attempt to estimate the effects of each individual coach, which would be very difficult for reasons we will explain, but we can ask whether coaches matter in the aggregate. Are some coaches better than others, such that we can statistically reject the null hypothesis that all are the same with respect for a particular outcome?

There are several methodological challenges associated with estimating coach effects. A key challenge, and the primary focus of this paper, is inference. Suppose we observe an apparent

correlation between coaches and outcomes. We would expect some of that apparent correlation just by chance, and we'd like to account for the idiosyncrasies of luck to determine whether that correlation is indeed statistically significant. If there is serial correlation in team performance—that is, trends over time that are unrelated to coaches—we would observe additional correlation between coaches and outcomes because the tenure of some coaches will happen to overlap with good or bad runs for the team. Because of random noise and serial correlation in coaches and outcomes, standard methods would fail to distinguish coach effects from luck.

Another set of challenges has to do with identification. If we detect a statistically significant relationship between coaches and outcomes, it might be attributable to coach effects or there might be other reasons that outcomes systematically correspond with coaches. For example, if the outcome of interest affects coach retention, this could also generate a correlation between coaches and outcomes. Although we cannot entirely remove these concerns, we can show through simulations and sensitivity analyses that the substantive relevance of these identification concerns is minimal in the settings we study.

Our general strategy involves regressing an outcome on coach fixed effects, recording a summary statistic of fit, and then simulating the distribution of summary statistics that we would expect under the null. As summary statistics of fit, the r-squared, adjusted r-squared, and F-statistic will all produce identical p-values and implied effect sizes in our subsequent analyses because, for a given sample size and number of regressors, these statistics all increase monotonically as the others increase. For the purposes of this paper, we focus on the r-squared statistic, which is familiar to social scientists and has a substantive interpretation as the proportion of variation in the outcome that appears to be explained by the coach fixed effects. However, if subsequent practitioners would prefer using another fit statistic, that is perfectly allowable within our framework.

In and of itself, the r-squared statistic is not particularly informative. A high value could reflect coach effects, but it could also reflect within-team variation over time unrelated to coach effects, or it could suggest that the regression with many independent variables over fit random variation in the outcome. Therefore, we need a strategy for simulating the distribution of r-squared statistics that we would expect under the null hypothesis of no coach effects. To do this, we randomly permute the ordering of coaches within each team, keeping the tenure of each coach the same as in the real data set but varying the order in which each coach served. For each random permutation, we regress the outcome of interest on the artificial coach fixed effects and record the r-squared statistic. We repeat this procedure many times to estimate the distribution of r-squared statistics we would obtain under the null of no coach effects. The proportion of random permutations that produce an r-squared statistic greater than that from the real data is an estimated p-value testing the null hypothesis that all coaches are equally effective with respect to this particular outcome.[3]

Prior to implementing our method, we take several steps to prepare the data for analysis. We typically start with game-by-team level data, but to improve computational efficiency, we aggregate the data up to the season-by-team level. This causes little to no loss of information since the coach of a team rarely changes mid-season.

[3] These hypothesis tests are one sided because there is no reason to expect the real r-squared statistic to be smaller than the expected r-squared statistic under the null. If some coaches are better than others, this will only increase the value of the real r-squared statistic.

Before aggregating up to the season level, we try to remove variation attributable to the quality of a team's opponents as well as home field advantage. This step is not necessary, but it improves statistical precision. When examining game-team level data, each data point corresponds to a game between team *i* and team *j*. For each outcome of interest, we calculate the average value for team *j* across all games that season that were not played against team *i*. Then, we run a regression of the outcome of interest on these measures of opponent quality, year fixed effects, and an indicator for a home vs. away game. We calculate the residuals from this regression, indicating the performance of each team in each game over above what would be expected given the year, home field advantage, and quality of their opponent. We then calculate the average residual across games for each team-season, which becomes our outcome of interest for estimating coach effects.

Having processed the data in this way, we regress our outcome of interest on coach fixed effects and record the r-squared statistic. Then, to simulate the distribution of r-squared statistics that we would expect under the null, we randomly permute the coach identifiers, keeping each coach's tenure together as a block in each permutation, re-run the regression of the outcome on coach fixed effects, and repeat this procedure many times.

Summary of our Procedure

1. Residualize game-team level data by season, home-field, and quality of opponent (optional).
2. Aggregate to season-team (optional).
3. Regress the outcome on coach fixed effects and record the r-squared statistic.
4. Randomly permute coaches within each unit, sampling each coach's tenure as a block.
5. Regress the outcome on permuted coach fixed effects and record the r-squared statistic.
6. Repeat steps 4 and 5 many times, recording the proportion of cases where the r-squared from the permuted data is greater than that from the real data.

The logic of our random permutation tests is as follows. Assume that coach transitions are unrelated to potential outcomes, such that in the absence of any coach effects, there should be no systematic correspondence between coach and outcomes. There are three ways we can get a high r-squared statistic when we regress an outcome of interest on coach indicators. First, there could be coach effects, and this is what we'd like to identify. Second, there could be serial correlation or genuine trends in performance over time within teams even in the absence of coach effects, and the coaches who happened to serve in good (or bad) times will get credit for this in the regression. Third, the coach fixed effects could be over-fit to random, season-to-season fluctuations in performance, further inflating the r-squared statistic. Therefore, in order to test for coach effects, we'd like our random permutation tests to incorporate the last two factors but not (all of) the first.

In our random permutations, the number of fixed effects in each regression is held constant, and the distribution of tenure across coaches is also held constant. This means that the extent of overfitting is the same, in expectation, in the real data and the permuted data.[4] Furthermore, if

[4] This assumes that the researcher does not use the observed data to make specification choices. If a careless researcher modified the above procedure to better fit the observed data, the resulting p-values would be misleading. This is, of course, a concern with virtually all quantitative analyses, although we attempt to mitigate these concerns in this case by specifying a simple and generalizable procedure that will be applied in the same way to different data sets.

there is serial correlation or team-specific time trends unrelated to coaches, this will inflate the r-squared from the permuted regressions in the same way, in expectation, as it inflates the r-squared from the real regression. In either case, some coaches might wrongly receive credit for good times. However, if there are genuine coach effects, this will increase the r-squared in the real regression by more than it increases the r-squared in the permuted regressions. Therefore, if the r-squared from the real data is larger than that from the random permutations, this is an indication that ebbs and flows in performance coincide with the intervals of time in which different coaches served, suggesting that some portion of that r-squared statistic can be attributed to coach effects rather than just serial correlation or chance.

Our practice of sampling each coach as a block and maintaining the same distribution of contiguous periods of service in our permutations is important. If we randomly sampled each season independently, we would account for random noise but not the possibility of serial correlation or team-specific time trends, and as a result, we would likely reject the null even if there are no coach effects.

To understand how our method performs in different scenarios, let's consider how varying features of the data generating process will influence the r-squared statistic in the real and permuted data sets. Recall that all of the regressions run under RIFLE will include a set of coach fixed effects. By definition, $r^2 \equiv 1 - \frac{RSS}{TSS}$, where RSS is the residual sum of squares and TSS is the total sum of squares. In our regressions, the TSS is identical for both the real data and the permuted data sets where the ordering of coaches is randomly shuffled. Therefore, to think about how our method works, we need to think about how coach effects, time effects, and random noise influence the RSS. Random noise increases the RSS, and it increases the RSS in the same way, in expectation, in the real data set and the permuted data sets. If the noise was expected to affect the RSS differently in the real data, it wouldn't be random. Similarly, time effects that are unrelated to coaches' tenures will also increase the RSS, and they will increase the RSS the same way, in expectation, in the real and the permuted data sets. This is why our method of permuting coaches' tenures accounts for noise and time trends unrelated to coaches.

How do coach effects influence the RSS? A constant effect for each coach should have no effect on the RSS in the real data set. In this context, $RSS \equiv \sum_i \sum_t (Y_{it} - \bar{Y}_i)^2$, where i denotes coaches and t denotes seasons within each coach. In other words, the RSS is the sum of squared deviations of each data point from the mean for each coach. A constant effect for each coach would mean that the outcome is shifted by the same amount for all observations within each coach, such that each Y_{it} would be shifted by the same amount as each $\bar{Y}_i$, and the RSS would be unchanged by coach effects. However, in the permuted data sets, coach effects would increase the RSS. For each permuted coach tenure that overlaps with multiple actual coach tenures, coach effects will shift observations by different amounts within each permuted coach, thereby increasing the RSS. This means that in the presence of genuine coach effects, we expect the RSS to be lower for the real data set than in the permuted data sets, meaning that the r-squared will be higher.

To fix ideas, consider the simplest possible example where our test would allow us to say something about coach effects. Suppose there is 1 team with 2 coaches across 3 seasons. Suppose Coach A served during the first two periods, and Coach B served during the last period. In this simple example, there are only two ways to permute the coaches. We can assign Coach A to the first two periods—as in the real world, or we can assign her to the last two periods. If Coach A is better

than Coach B, or vice versa, we would expect the outcome from the first two periods to be more similar to each other than they are to the value from the third period, and the real data will give a higher r-squared statistic. If there are no coach effects but there is random noise or serial correlation, either permutation is equally likely to give a higher r-squared.

This simple example illustrates several features and limitations our approach. First, identification comes from coaches who serve different periods of time. If there were 4 periods and each coach served two periods, both permutations would yield the same r-squared. Next, our procedure behaves poorly when there are few coaches per team. In the example above, the p-value can only take one of two possible values, but asymptotic refinement improves quickly with more teams or more coaches per team, so long as there is variation in lengths of service. Furthermore, our approach does not require us to hypothesize that one particular coach is better than another. For the purposes of this study, we are agnostic about which coaches are better. We test whether some coaches are different from others in ways that matter for various forms of team performance.

It is worth emphasizing how our method implicitly accounts for player quality. Some prior studies have explicitly controlled for player quality variables when estimating coaching effects, but this is a thorny question. If we believe that coaches have a lot of influence over their roster—through recruiting and drafting new players or by better developing the players they inherited, for instance—then we would not want to control for player quality when estimating coach effects, because player quality itself is an outcome attributable at least in part to coaches. In other words, player quality would be a post-treatment variable. On the other hand, if coaches don't influence player quality, we would want to control for it in some way. But controlling for player quality directly would be difficult, requiring good measures of the quality of potentially every player on the team. Complicating matters further, we may not even know how much influence a coach has over the roster, and the level of influence may vary from team to team even within the same sport.

RIFLE addresses all of these issues implicitly because we can think of that component of player quality that is outside the control of the coach as a source of serial correlation. To the to the extent that player quality varies for reasons unrelated to the coach, it should be equally correlated with the permuted coach effects as the true coach effects, in expectation. But to the extent that coaches do influence the quality of players on their team, player quality will be more highly correlated with the true coach fixed effects than the permuted coach fixed effects. Thus, RIFLE accounts for player quality without our having to make any assumptions about the extent to which coaches control their roster and without our needing to include player-level covariates. RIFLE appropriately attributes to coaches only the variation in player quality that coincides with their tenures.

We have developed a Stata package that will allow future researchers to easily apply RIFLE to many different contexts and outcomes in order to better understand where, when, and why coaches matter.[5]

The procedure described above allows us to statistically test whether coaches meaningfully differ from one another. In other words, we can test the sharp null hypothesis that all coaches are equal to

[5] The package can be downloaded by typing "ssc install rifle" within Stata. Afterward, specific instructions can be obtained by typing "help rifle".

each other in terms of their influence on a particular outcome of interest, and if we reject that sharp null, then we would conclude that some coaches are indeed different from others. But this says nothing about the substantive size of the differences. What if we want to know the proportion of variation in a particular outcome that's attributable to coaches as opposed to other factors?

To say something about substantive effect sizes, we compare the r-squared from the original regression to the average r-squared from the permuted regressions. This difference, in and of itself, doesn't tell us much about substantive effect sizes, but it increases monotonically with the effect size. Therefore, to translate this observed difference to a substantively meaningful number, we conduct simulations in which we vary the proportion of variation attributable to coaches and see how each effect size corresponds to the expected difference in r-squared statistics. Then, we can compare our estimated difference to the simulated differences to see which effect size is most consistent with the observed result.

4. Results

We apply our method for estimating coach effects to the high-stakes settings in the U.S. in which coaches are highly compensated—the MLB, NBA, NHL, NFL, college football, and college basketball. All data was provided by Sports Reference, Inc. The outcome that is most readily observable and arguably most important is whether a game is won or last. Furthermore, we have data on the scores of each game, so we can also examine points scored, points allowed, and the point margin. In some sports, there are reasons to think that coaches might have more ability to affect points scored vs. points allowed, or vice versa, so we separately examine both outcomes. For some sports, we have also collected additional data and conducted our test on other specific outcomes of interest for that sport.

4.1. Power and Effect Size Simulations

Before showing our results, we first demonstrate the statistical power of our tests in each setting, and we conduct simulations that will later allow us to estimate substantive effect sizes. We conduct these simulations by utilizing the same data sets that we use to produce our results, but we simulate new outcome variables with known coach effects.

Specifically, we assume that each coach has an effect that is drawn from a standard normal distribution, we also simulate noise from a standard normal distribution, and we vary the extent to which a hypothetical outcome of interest is influenced by both coach effects and noise. For example, if we simulate the outcome as the coach effect plus the noise, then coach effects explain 50 percent of the variation in outcomes. If the outcome is the coach effect plus 9 times the noise, then coach effects explain 10 percent of the variation.

The results of our power analyses are shown in Figure 1. We have the greatest statistical power in the context of college basketball and football, presumably because we have data from many teams over many seasons. If coach effects explain 10 percent of the variation of an outcome of interest in this setting, we should statistically detect them about half the time, and if coach effects explain 20 percent of the variation, then we are virtually guaranteed to detect them. We have slightly less power when studying the MLB, NBA, NHL, and NFL. In these settings, we are likely to detect 20 percent effects and virtually guaranteed to detect 30 percent effects.

These simulations also allow us to assess substantive effect sizes by observing the extent to which the difference between the real r-squared and the average permuted r-squared increases with the true effect size. Figure 2 shows these results, using the same color coding as in Figure 1. As expected, higher effect sizes correspond with higher differences in the r-squared statistic, although this relationship is non-linear. Having observed particular difference, we can refer back to this graph to say something about the proportion of variation in an outcome of interest that is likely attributable to coach effects.

4.2. Baseball

Table 1 shows the results of our analyses for MLB managers, using data from 1871 through 2016. As with all subsequent sports, we estimate the effect of managers on runs scored, runs allowed, point margin, and victories. The table reports the r-squared statistic from the real data, the average r-squared statistic arising from the permuted data sets, the difference between the two, the estimated p-value, and also an estimate of the proportion of variation attributable to coach effects. To compute this last number, we take the estimates from Figure 2, and we use linear interpolation to obtain a point estimate.

We find evidence that MLB managers matter for all of these outcomes, although they appear to matter more for runs allowed than for runs scored. For runs scored, the estimate is not statistically significant at conventional levels ($p = .058$) and the difference in r-squared is substantively small. But for runs allowed, the p-value is strongly statistically significant ($p < .001$) and the difference in r-squared suggests that managers explain 28 percent of the variation within teams and across seasons in runs allowed.

One potential explanation for this discrepancy is that managing defense in baseball requires more strategic decisions than managing offense. For the most part, the job of the manager on offense is to put the best hitters in the lineup in the best order, and most managers would probably make similar decisions with the same team. However, on defense, the manager must efficiently utilize their pitchers without wearing out their arms. Some managers may be better than others at determining when a starter has thrown too many pitches to be effective, when to use a reliever, and which reliever to use in a particular situation.

One common notion is that a big part of an MLB manager's job involves the allocation of scarce resources across games. Each team plays 162 games in the regular season, and each pitcher can only throw so many pitches per week. Therefore, some might argue that the most important job of the manager is not to increase runs scored or reduce runs allowed but instead to efficiently allocate runs across games. If the outcome of one game is a forgone conclusion, a manager might as well save their best pitchers for the next game. To test whether some managers are better at this than others, we also utilize wasted runs as one of our outcomes of interest. Wasted runs are measured as the margin of victory when a team wins and the number of runs scored when a team loses, and we might expect efficient coaches to reduce the number of wasted runs—allowing their teams to win more games with the same numbers of runs scored and allowed. Interestingly, we find little evidence that managers affect wasted runs. Our estimated effect of coaches on wasted runs in substantively small and statistically insignificant ($p = .240$). One potential explanation for this result is that it may not be easy to ex ante predict which runs will be wasted or not, and therefore, managers are unable to effectively decide when to save their pitchers for the next game.

To illustrate our inferential strategy, Figure 3 shows the distribution of r-squared statistics across 1,000 permuted data sets for both runs allowed and wasted runs. The graph also plots (in red) the r-squared from the original data set. In the case of wasted runs, we see that the real r-squared is notably higher than any of those from the permutations, consistent with large and genuine coach effects. In the case of wasted runs, the real r-squared falls in the middle of the permuted distribution, suggesting that we have no statistical evidence that coaches matter for wasted runs.

4.3. Football

Table 2 shows the results of our analyses for NFL coaches. We look at virtually the entire history of the NFL from 1922 through 2016, although the results are similar if we just focus on the modern era. As with baseball, we analyze points scored, points allowed, point margin, and victories. NFL coaches clearly affect points scored and the point margin. The estimates imply that coaches explain 18 to 25 percent of within-team, between-season variation in points allowed and point margins. The estimated effect on points scored is slightly smaller, although this is not true if we just focus on the modern era. The estimated effect on victories is also not statistically significant if we only focus on the modern era, presumably because football teams play few games per season so the power of this particular test is low.

We also have season-level data for some other outcomes of interest for the NFL during the modern era of 1970 through 2016. Specifically, we examine fumbles per game, penalties committed per game, opponents' penalties per game, and the proportion of offensive plays on which a team passes. Interestingly, we find that coaches matter a lot for fumbles and for the penalties a team commits. Coach effects explain about 30 percent of the within-team, between-year variation in these variables, with some coaches apparently doing a much better job preventing fumbles and penalties than others. Interestingly, coaches appear to have little effect on penalties committed by opponents, perhaps revealing that there's not much a team can do to systematically induce penalties by their opponents. And quite surprisingly, coaches don't appear to meaningfully differ in their use of passing versus rushing. Clearly, coaches could simply force their teams to pass or run more often, but we don't find much evidence that coaches systematically differ from one another on this dimension. Perhaps most coaches are following the same rules of thumb and are getting their teams close to the optimal share of passing versus rushing.

Table 3 shows our results for college football coaches from 1900 through 2016. We include data from all Division 1-A teams after 1978—when Division 1 was subdivided—and all Division 1 teams before 1978. Interestingly, the estimated effects are larger for college football than for professional football. One potential explanation is that in addition to managing practices and games, college football coaches also play a crucial role in recruiting. Furthermore, because we have so much data for college football, the results are extremely statistically significant ($p < .001$) for all outcomes.

4.4. Basketball

Tables 4 and 5 show our results for the NBA and for Division 1 college basketball, respectively. In both cases, the estimated effects are substantively quite large. Coaches explain about 30 percent of the variation in points scored and allowed. One initially surprising result is that in college basketball, coaches matter more for points scored and allowed than they do for the point margin. One potential explanation is that coaches differ from each other in their preferences for fast- versus slow-paced games, with the fast-paced coaches both scoring and allowing more points. To explicitly test this hypothesis, we also test whether coaches matter for the total points scored in the game, and here, we detect a huge effect, confirming this hypothesis about different coaching styles.

4.5. Hockey

Lastly, Table 6 shows results from NHL coaches from 1918 through 2017. As with the other sports, coaches matter and the results are statistically significant, although coaches appear to matter much more for goals allowed than for goals scored. We do not have a clear explanation for this result as we do in the case of baseball, and we defer to hockey experts who might have a good explanation for this and might also suggest additional outcomes of interest in this context.

5. Addressing Endogenous Retention

Our method accounts for the fact that some coaches might appear to look good simply because of good luck or the fact that they happened to serve when a team had really good players. A threat to identification would have to come from performance coinciding with coaches' tenures for reasons other than the effects of the coaches. In our view, the most concerning such possibility is that the tenures of coaches are influenced by their past performance. Coaches who have performed poorly are more likely to be replaced, and this could make it look like coaches matter even if they don't. In Berry and Fowler (2018), we present a theoretical model and Monte Carlo simulations that show how endogenous retention could bias our test. However, the bias is typically small, and it can go in either direction.

To assess the likely extent and direction of the bias in this context, we have conducted additional Monte Carlo simulations. The goal of these simulations is to suppose that there are no coach effects but there is serial correlation in a team's success (perhaps because of good players) and coach retention is endogenous (perhaps because the team's management believes that coaches matter and they're learning about each coach's ability based on their prior performance). How would this bias the results of our test?

To implement these simulations, we start with the actual data sets used to generate our main results, keeping the number of time periods and teams as they are. We simulate the outcome of interest and the coach identifiers according to a known process. Specifically, the outcome is drawn irrespective of coaches but with serial correlation. Performance in each year is a weighted combination of last year's performance and a new random draw. The weight given to last year's performance is chosen based on the actual serial correlation of outcomes observed for the sport in question. Specifically, we regress residualized victories on the lagged residualized victories, and we use that coefficient to determine the level of serial correlation for the simulations within that setting.

To generate the simulated coach identifiers, we assume that the probability of turnover varies linearly with previous performance. The intercept comes from simply measuring the average probability of turnover in our sample, and the slope is estimated by standardizing residual victories and then regressing coach turnover on this measure of lagged performance and team fixed effects. So the extent of endogenous turnover in our simulations is determined by the observed extent of endogenous turnover in each setting.

Using these simulated data sets, we implement our method to see if we reject the null hypothesis (i.e., $p < .05$). We repeat this many times to estimate the false discovery rate of our method assuming no coach effects but the level of endogenous retention observed for that sport. Ideally, we would obtain a false discovery rate of .05, and to the extent that our results deviate from that, we can learn the extent to which endogenous retention leads us to over- or under-reject the null.

The results of these Monte Carlo simulations are in Table 7. For each setting, we report the average probability of coach turnover (Intercept), the extent to which the probability of turnover corresponds with a standard deviation increase in performance (Slope), and the extent to which performance corresponds with lagged performance (Serial Corr). And importantly, the last column of the table presents the estimated false discover rate (FDR) using these estimated parameters and the data sets for each setting.

Although the combination of serial correlation and endogenous retention can, in principle, bias our test, the implications of this bias are small given the extent of these phenomena observed in these settings. According to our Monte Carlos, our false discovery rates are all close to the ideal of .05. For the professional sports settings, we conducted 1,000 iterations and the estimated false discovery rates range from .049 to .053. For college football and basketball we have more data, so these simulations take more time, and we accordingly only did 100 iterations. Even still the estimated false discovery rates are very close to the theoretical ideal—.05 for football and .07 for basketball. These results suggest that the endogenous retention of coaches does not meaningfully bias our results.

6. What Can We Say about Individual Coaches?

Our method allows us to estimate the extent to which variation in performance is attributable to coaches as opposed to luck and other factors outside coaches' control. We think it's useful to know how much coaches matter and for what outcomes they matter most. Nonetheless, our method does not allow us to say which coaches are particularly effective or ineffective.

Analysts will naturally want to determine which coaches are most effective, and although our method is not suited for answering that question directly, our basic approach to inference can be useful for this question as well. Often, however, careful analysts will find that it's difficult to confidently assess the quality of an individual coach.

In particular, analysts and team stakeholders would like to be able to use a coach's past performance to predict their future performance, but this is naturally difficult when they have only served a few seasons. Our estimates imply that coaches often explain 20-30 percent of the variation in a team's success. Substantively, that's a large effect, and it's well worth investing in a good coach

if you can identify one. But if a coach has only served a few seasons, a good record doesn't provide much information about their quality. More likely than not, the other factors that explain the remaining 70-80 percent of the variation were working in that coach's favor, and we'll expect regression to the mean in future years.

Acknowledging this inferential problem, the best coaches are not necessarily the ones with the best records. But a coach who has served several seasons and has a high record given their tenure is worth investigated as a potentially exceptional coach. Figure 4 shows the average residual victory across seasons coached for every coach-by-team in the history of the NFL. The coaches with the best averages tend to be those that only coached one or two seasons, just like the coaches with the worst averages tend to have only coached one or two seasons. A few coaches have averages above .25, meaning they were 25 percentage points more likely to win a game than an average coach—conditional on the quality of their opponent and home field advantage. This is a remarkable record. But nobody who coached more than 2 seasons maintained such a high record. And furthermore, even if coaches didn't matter, we'd probably expect a few coaches to have records like this merely by chance.

To identify the coaches that are genuinely likely to be much better than average, we'd want to look at those who coached more seasons, and we'd want to look at those coaches on the upper frontier who have high records given their number of seasons in the league.

Some analysts believe that Bill Belichik is one of the greatest NFL coaches of all time, and his remarkable tenure with the New England Patriots is filled in and colored red in Figure 4. Through the 2016 season for which we have data available, Belichik had served 17 seasons with the Patriots and his average residual victory was .18. Other coaches have higher averages, but an average that high after coaching so many seasons is extremely unusual. Only Paul Brown's 17-year tenure with the Cleveland Browns starting in 1946 exceeds Belichik's run. As Bears fans, we'd like to point out that arguably the most impressive record in Figure 4 is that of George Halas, who coached far more seasons than any other coach and nevertheless was 9 percentage points more likely to win a given game than an average coach—controlling for opponent quality and home-field advantage.

How can we test whether an outlier like Belichik is genuinely great or whether he could have achieved his success by luck? Similar in spirit to the Monte Carlos conducted above, we could simulate outcomes in a world in which all coaches are equally effective but there's random noise, serial correlation, and endogenous retention, and we could see how often someone who looks as good as Bill Belichik arises. Importantly, the right question is not the odds that a randomly selected coach will look as good as Bill Belichik. Belichik might just be the luckiest coach, and we're focusing on him because of his impressive record. Instead, the right question is about the odds that *any* coach could arise with a record as good as Belichik's even in a world where coaches don't matter.

Using the same simulations described in the previous section, and using the parameters estimated for the NFL, we can compute the average performance in their first 17 seasons for all simulated coaches that served that long, and we can record the best such average across all coaches. We can then compare this to Belichik's actual performance, and we can see how likely such a record is to arise by chance. The answer is *very unlikely*.

The results of this exercise are shown in Figure 5. If we standardize the season-level performance measure, we see that Belichik's season-level average is 1 standard deviation above the mean. When

we simulate 10,000 hypothetical NFL's with no coach effects and with serial correlation and endogenous retention comparable to what we observe in the real data, there are only 5 cases in which a coach served 17 seasons and had such an impressive record. In that sense, we can strongly reject the null hypothesis that Belichik is no better than an average coach (p = .0005).

To illustrate the difficulty of assessing a coach's quality early in their careers, we have conducted the same kinds of simulations but only using data from an early point in a coach's career. The best average residual victory for any coach in their first season was .412, achieved by Adam Walsh who coached the Rams in 1945. In our simulations with no coach effects, at least one coach has a first season as good as this one 62 percent of the time. The null result for Walsh may be appropriate since he had a lackluster season in 1946 and then never coached at the professional level again.

Similarly, the best first two seasons for any coach were achieved by George Seifert with the 49ers in 1989 and 1990. In our simulations with no coach effects, at least one coach exceeded Seifert in their first two seasons 57 percent of the time. This might be a false negative in the case of Seifert who won two Super Bowls and only missed the playoffs once in his eight seasons with the 49ers. However, as coach of the Panthers, Seifert had two lackluster seasons and one atrocious season before being fired, so we may have been right not to draw overly strong conclusions from two great seasons.

Using a conventional threshold of .05 for statistical significance, we cannot reject the null hypothesis that any NFL coach is better than average if we only use data from their first four seasons. We have to wait until they have served 5 seasons before we are able to find statistically significant evidence for any coach. The highest average by any coach in their first 5 seasons was achieved by Paul Brown—mentioned above, and for him, we obtain a p-value of .008 at that point in his career. Sure enough, Brown went on to have a long, successful career with Cleveland and Cincinnati, maintaining a better-than-average record throughout.

Future analysts could adapt this procedure to their setting of interest and apply this logic to determine whether we should be confident or not that a particular coach is truly high quality. The method cannot explicitly say who is best, but it can assess which coaches' records are more or less likely to have arisen by chance, which is informative for forecasting their likely success in future seasons.

7. Summary and Conclusion

Our analysis challenges the prevailing view in sports analytics that coaches are interchangeable. Using a new method that overcomes methodological limitations of previous studies, we show that coaches significantly affect outcomes in every sport we studied. Some of our most notable findings are as follows.

- MLB managers affect runs scored, runs allowed, point margin, and victories. They matter more for runs allowed than for runs scored. They do not matter for wasted runs.

- NFL coaches affect points allowed and the point margin. They significantly affect the number of fumbles and penalties a team commits. Coaches don't meaningfully differ in their choice of passing versus rushing, perhaps because of in-game constraints.

- Coaches matter even more in college football than in the NFL. They significantly affect points scored, points allowed, the point margin, and the number of victories in a season.

- Coaches are highly significant for team outcomes in the NBA and Division 1 college basketball. They influence points scored, points allowed, the point margin, and the number of victories. In college basketball, coaches also affect the total number of points scored in a game.

- NHL coaches matter, although they matter much more for goals allowed than for goals scored.

- We discussed a method for evaluating individual coaches, which is informative for understanding whether a given coach has performed better than would be expected by chance. This method is superior to conventional tests of statistical significance for individual coach fixed effects and could be applied by team stakeholders in evaluating coaching candidates.

Future research could extend this study in any number of fruitful directions. Most obviously, an expert in a particular sport might have ideas for other relevant outcomes to investigate. In addition, RIFLE can be used to help understand more about why and how coaches matter. For instance, if an expert had hypotheses suggesting that coaches matter more in some contexts than others, she might subset the data to test differences in coaching effects across such contexts.

Analyses using RIFLE need not be confined to team-level outcomes. One might be interested in whether coaches matter for the performance of players at particular positions. For instance, RIFLE could be used to test whether NFL coaches matter for quarterback play specifically. Nor do analyses need to be confined to head coaches. An analyst might want to know whether offensive or defensive coordinators in the NFL, or pitching coaches in MLB, matter for some specific outcomes under their purview. However, to the extent that head coaches influence the hiring of these additional coaches, their effects are included in our estimates.

Others might be interested in understanding whether some particular attribute, such as experience or education, is associated with coach effects. One could control for coach characteristics in the residualizing regression. If the estimated coach effects under RIFLE diminish after controlling for the characteristic in question, this would constitute evidence that the characteristic is associated with coaching effects (though not necessarily that the relationship is causal).

Our goal in this paper has not been to deliver an exhaustive or definitive study of coaching. Rather, we offer a method and associated software that analysts can use to study the effects of coaches on any measurable outcome in any sport.

References

Adler, E. S., Berry, R. J., & Doherty, D. (2013). Pushing "reset": The conditional effects of coaching replacements on college football performance. *Social Science Quarterly, 94*(1), 1-28.

Alamar, B. (2013). *Sports analytics: A guide for coaches, managers, and other decisionmakers*. Coumbia University Press: New York.

Alamar, B., & Mehrota, V. (2011). Beyond 'Moneyball': Rapidly evolving world of sports analytics. *Analytics Magazine* September/October.

Audus, R., Dobson, S., & Goddard, J. (1997). Team performance and managerial change in the English Football League. *Economic Affairs, 17*, 30-36.

Berri, D. J., Leeds, M. A., Leeds, E. M., & Mondello, M. (2009). The role of managers in team performance. *International Journal of Sport Finance*, *4*, 75–93.

Berry, C. R., & Fowler, A. (2018). Leadership or Luck? Randomization Inference for Leader Effects. Working Paper: The University of Chicago.

Bradbury, J. C. (2017). Hired to be fired: The publicity value of managers. *Managerial and Decision Economics*, *38*, 929–940.

Brown, M. C. (1982). Administrative succession and organizational performance: The succession effect. *Administrative Science Quarterly, 27*(1), 1-16.

Cannella Jr., A. A., & Rowe, W. G. (1995). Leader capabilities, succession, and competitive context: A study of professional baseball teams. *The Leadership Quarterly, 6*(1), 69-88.

De Paola, M., & Scoppa, V. (2012). The effects of managerial turnover: Evidence from coach dismissals in Italian soccer teams. *Journal of Sports Economics, 13*, 152-168.

Dohrn, S., Lopez, Y. P., Reinhardt, G. (2015). Leadership succession and performance: An application to college football. *Journal of Sport Management, 29*, 76-92.

Fabianic, D. (1994). Managerial change and organizational effectiveness in Major League Baseball: Findings for the eighties. *Journal of Sport Behavior, 17*(3), 135-147.

Fizel, J. L., & D'Itri, M. P. (1999). Firing and hiring of managers: Does efficiency matter? *Journal of Management, 25*(4), 567-585.

Gamson, W. A., & Scotch, N. A. (1964). Scapegoating in baseball. *American Journal of Sociology, 70*(1), 69-72.

Giambatista, R. C. (2004). Jumping through hoops: A longitudinal study of leader life cycle in the NBA. *The Leadership Quarterly, 15*, 607-624.

Goff, B. (2013). Contributions of managerial levels: Comparing MLB and NFL. *Managerial and Decision Economics*, *34*, 428–436.

Grusky, O. (1960). Administrative succession in formal organizations. *Social Forces, 39,* 105-115.

Grusky, O. (1963). Managerial succession and organizational effectiveness. *American Journal of Sociology, 62,* 21-31.

Koning, R. H. (2003). An econometric evaluation of the effect of firing a coach on team performance. Applied Economics, 35, 555-564.

McTeer, W., White, P. G., & Persad, S. (1995). Manager/coach mid-season replacement and team performance in professional team sport. *Journal of Sport Behavior, 18*(1), 58-68.

Rowe, W. G., Cannella Jr., A. A., Rankin, D., & Gorman, D. (2005). Leader succession and organizational performance: Integrating the common-sense, ritual scapegoating, and vicious-circle succession theories. *The Leadership Quarterly, 16*(2), 197-219.

Smart, D., Winfree, J., & Wolfe, R. (2008). Major League Baseball managers: Do they matter? *Journal of Sport Management, 22*(3), 303-319.

Smart, D., & Wolfe, R. (2003). The contribution of leadership and human resources to organizational success: An empirical assessment of performance in Major League Baseball. *European Sport Management Quarterly, 3*, 165-188.

Figure 1. Statistical Power

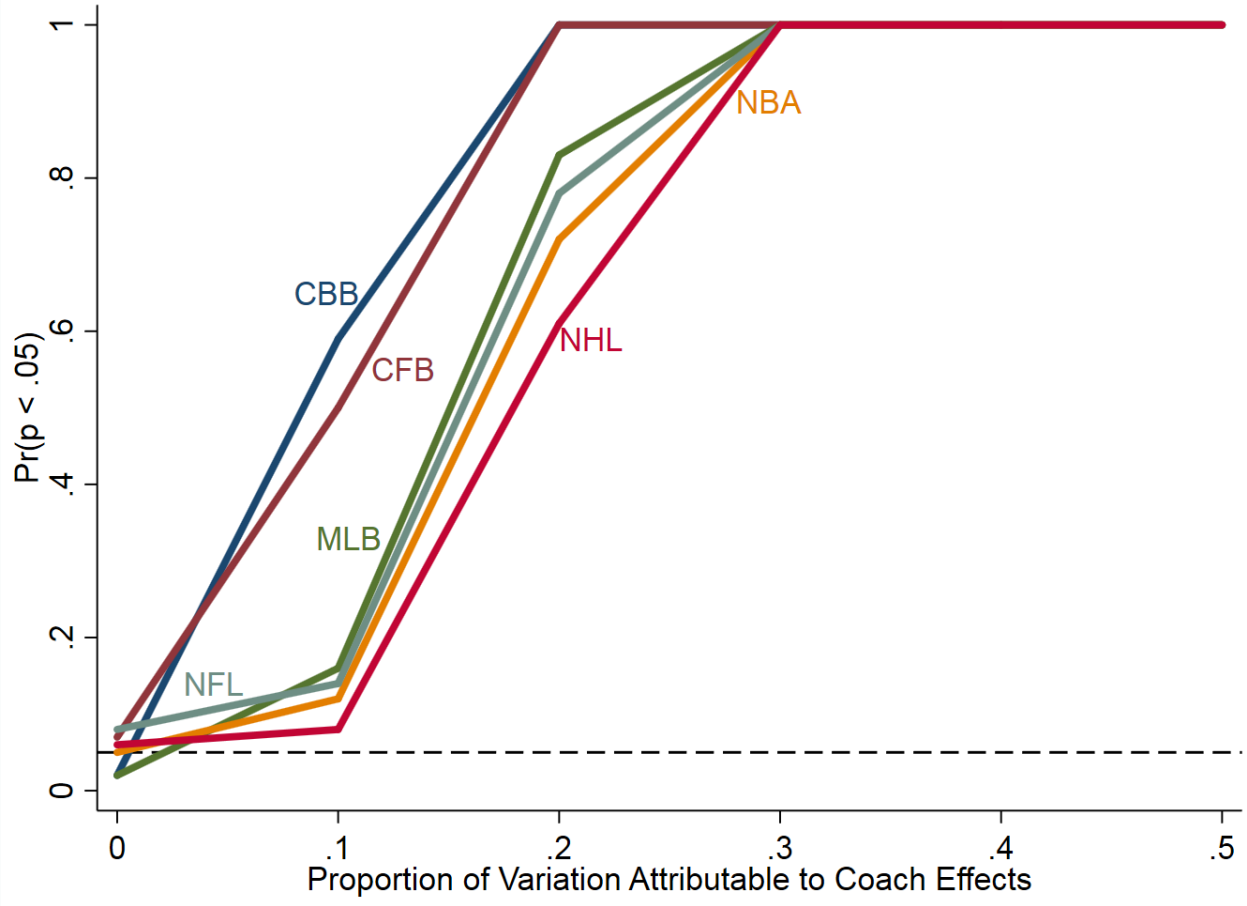


Figure 2. Interpreting Effect Sizes

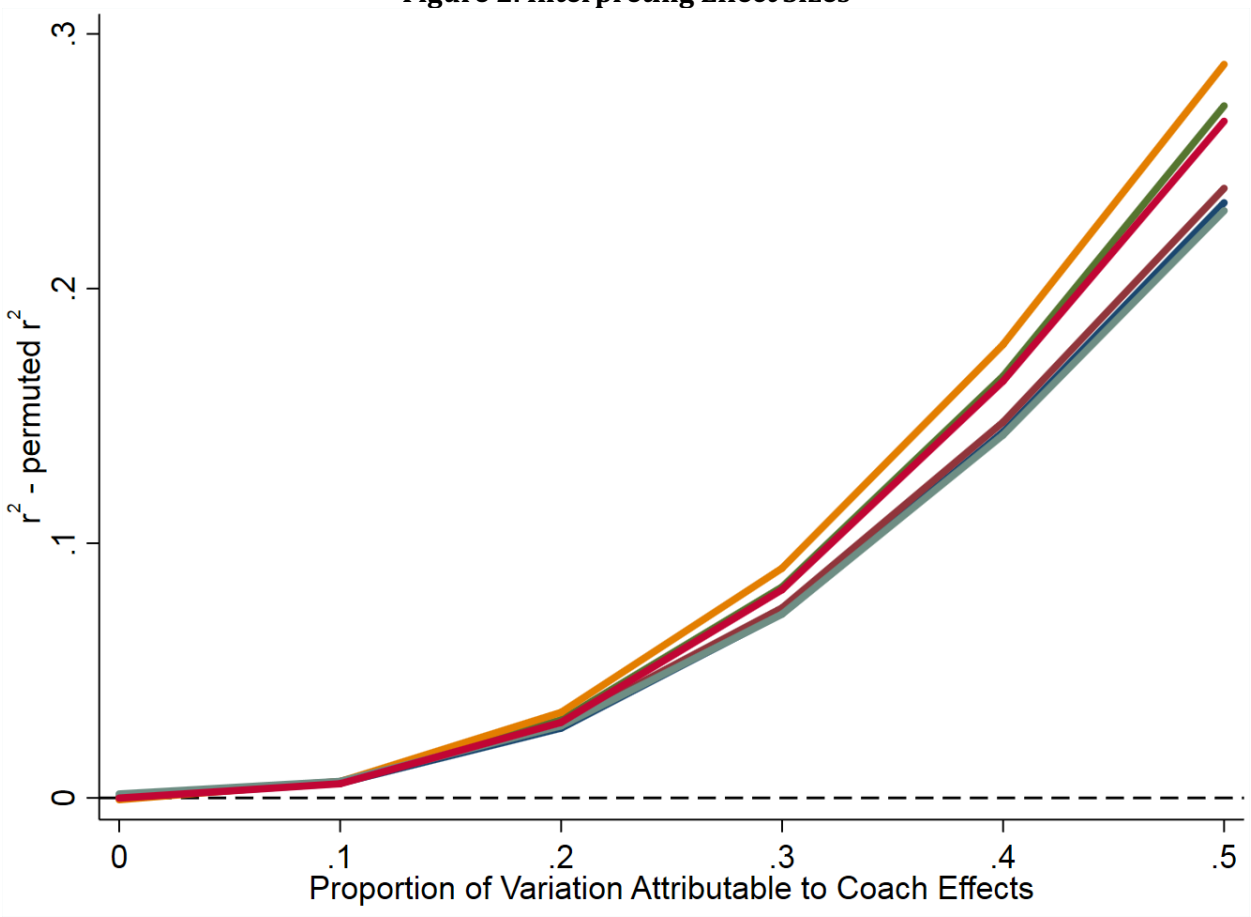
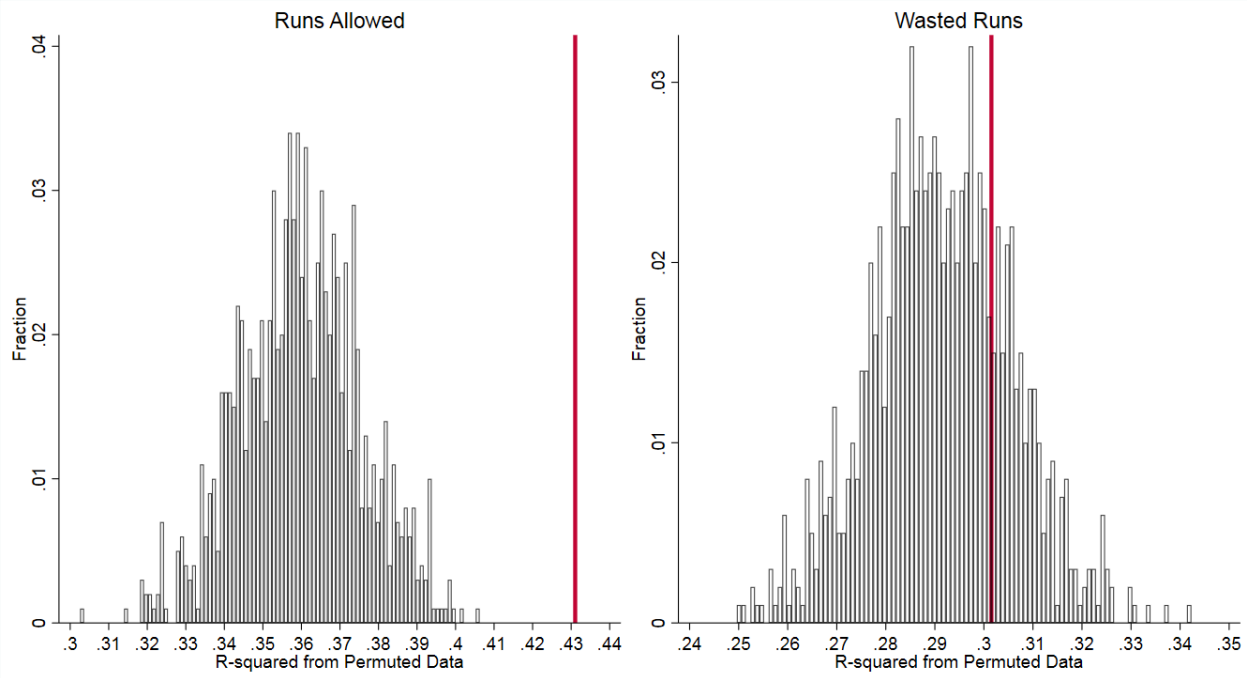
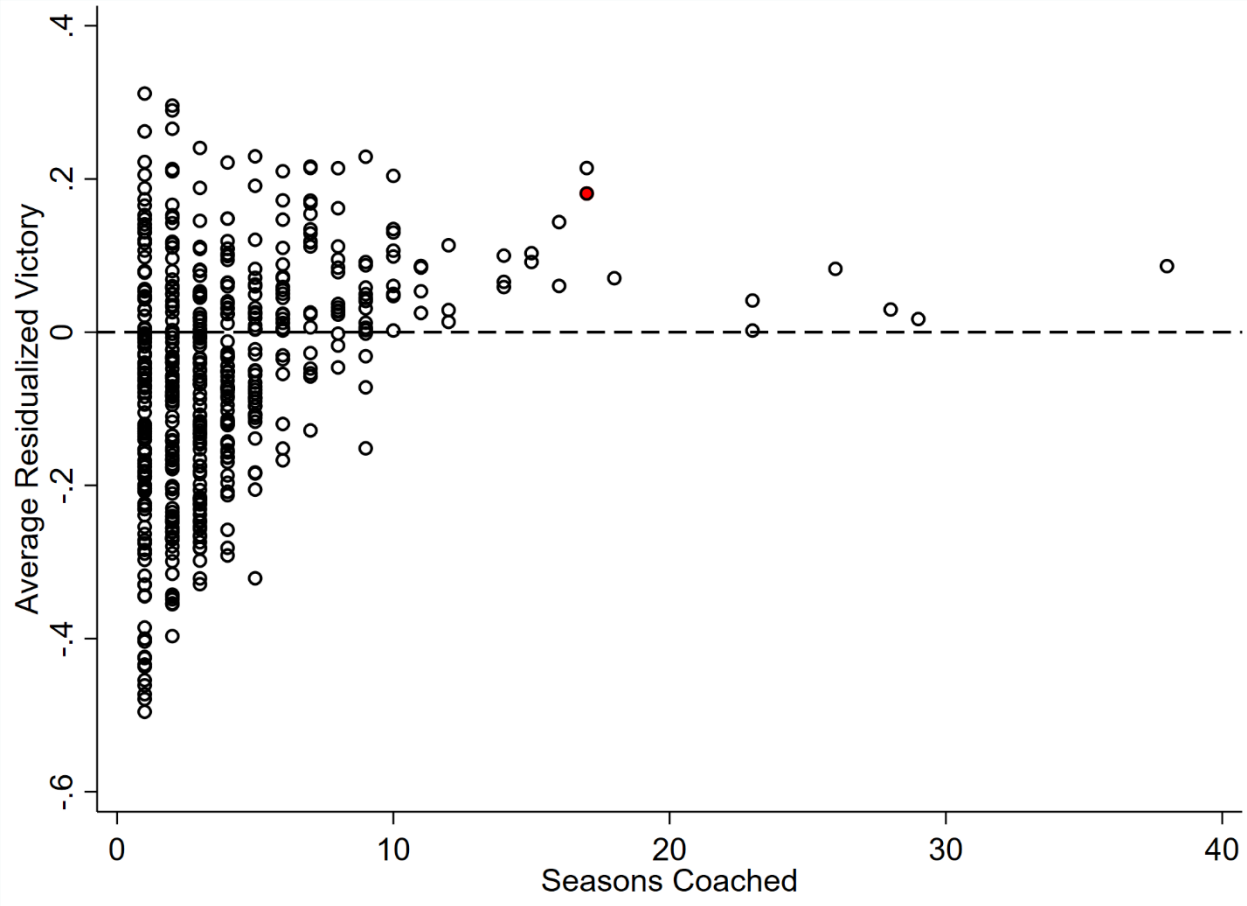


Figure 3. Graphical Illustration of MLB Results



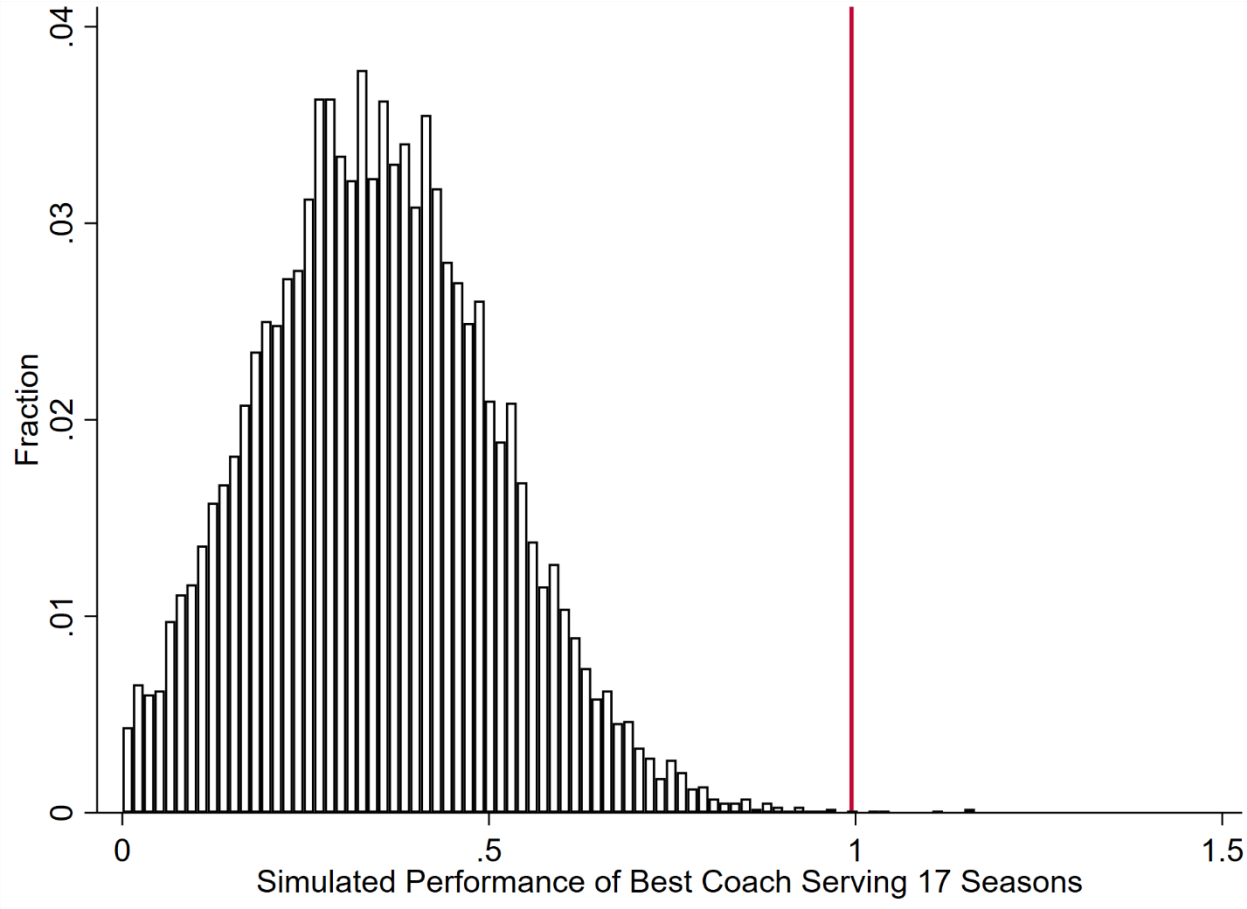
Note: The red lines denote the r-squared from the real data. Unshaded bars show the distribution of r-squared values from the permuted data.

Figure 4. NFL Coach Performance across Seasons Coached



Note: The red dot represents Bill Belichik.

Figure 5. Hypothesis Test of Bill Belichik's Effectiveness



Notes for Tables 1 to 6. *The outcomes referenced in the first column are residualized for opponent quality, home field advantage, and year, as described in the text. The column "r2" is the r-squared in the real data. The column "avg" is the average r-squared in the permuted data sets. The column "difference" is the difference in r-squared between the real data and the average from the permuted data. The column "p-value" is the p-value of the difference, which is the proportion of the permuted data sets with a higher r-squared than the real data. The column "prop" is the proportion of variance in the residualized outcome explained by coach effects, according to the analyses presented in Figure 2.*

Table 1. Results for MLB Managers, 1871-2016

outcome	r^2	avg	difference	p-value	prop
Runs Scored	.360	.334	.027	.058	.185
Runs Allowed	.431	.360	.071	.000	.277
Point Margin	.440	.373	.067	.000	.270
Victory	.383	.328	.055	.000	.247
Wasted Runs	.302	.291	.010	.240	.116

Table 2. Results for NFL Coaches, 1922-2016

outcome	r^2	avg	difference	p-value	prop
Points Scored	.353	.328	.025	.051	.185
Points Allowed	.362	.313	.049	.000	.247
Point Margin	.392	.349	.042	.026	.231
Victory	.350	.317	.033	.000	.211
Fumbles	.471	.395	.076	.000	.305
Own Penalties	.416	.347	.069	.000	.293
Opponent Penalties	.274	.280	−.006	.744	0
Prop. Off. Plays Passing	.450	.446	.004	.513	.049

Table 3. Results for College Football Coaches, 1900-2016

outcome	r^2	avg	difference	p-value	prop
Points Scored	.360	.306	.054	.000	.255
Points Allowed	.412	.338	.074	.000	.299
Point Margin	.422	.351	.071	.000	.292
Victory	.370	.316	.054	.000	.255

Table 4. Results for NBA Coaches, 1947-2017

outcome	r^2	avg	difference	p-value	prop
Points Scored	.396	.289	.107	.000	.319
Points Allowed	.387	.310	.077	.001	.277
Point Margin	.422	.315	.107	.000	.319
Victory	.415	.317	.099	.000	.310

Table 5. Results for College Basketball Coaches, 1938-2017

outcome	r^2	avg	difference	p-value	prop
Points Scored	.374	.271	.102	.000	.341
Points Allowed	.428	.304	.124	.000	.372
Point Margin	.340	.286	.055	.000	.261
Victory	.301	.254	.047	.000	.243
Total Points	.426	.287	.139	.000	.393

Table 6. Results for NHL Coaches, 1918-2017

outcome	r^2	avg	difference	p-value	prop
Points Scored	.404	.358	.046	.016	.231
Points Allowed	.458	.375	.083	.000	.301
Point Margin	.464	.387	.077	.000	.291
Victory	.432	.372	.059	.004	.256

Notes for Table 7. *The column "Intercept" is the average annual coach turnover rate in the data. The column "Slope" is the coefficient from a regression of a dummy variable coded as 1 if the team has a new coach in season t on the team's residualized victories in season t – 1. The column "Serial Corr" is the estimated serial correlation in the data based on a regression of the residualized victories in season t on the residualized victories in season t – 1. The column "FDR" is the false discovery rate in a simulation based on the listed values of intercept, slope, and serial correlation. Details of the simulations are provided in the text.*

Table 7. Monte Carlo Simulations with Endogenous Retention

Setting	Intercept	Slope	Serial Corr	FDR
MLB	.306	−.083	.502	.052
NFL	.234	−.069	.416	.053
CFB	.222	−.063	.417	.05
NBA	.299	−.078	.603	.049
CBB	.156	−.049	.464	.07
NHL	.352	−.083	.543	.053