

Measuring Quality in College Football *

Allyson Lynch

May 13, 2021

Abstract

This paper looks at the effect of college coaches on player development and in the process explores the development of a model using imperfect measures of quality. I constructed a model that created a continuous measure of player quality based on the ordinal measure of player quality that is available in high school recruit ranks and professional draft ranks. The model also includes a significant correction for attenuation bias. The regression results showed a small effect of college coaches on player development. Overall, this paper addresses the larger questions about how to evaluate value added when performance is mis-measured.

*I would like to thank Giri Parameswaran, Anne Preston, and Kieran Lynch.

1 Introduction

This study was motivated by the fact that college football coaches are highly paid with the top salary in 2019 at over 9 million dollars for the head coach at Clemson University, Dabo Swinney (Berkowitz et al., nd). High salary individuals from college football coaches to CEOs are often discussed in the news. The debate centers around if these individuals are worth what they cost their employer and if their high salary is at the expense of those below them –either their employees or players. However, it is difficult to answer these pressing questions with the data available and take the appropriate action based on the results. This study begins to answer this question by looking at the effect of college football coaches on player development.

In order to look at player development I constructed a model using changes in high school recruit rankings and professional draft rankings to determine the average coaching effect from different college coaches at different schools. However, using ordinal rank as a measure for the continuous variable of quality presents some issues in understanding the true difference in quality. Players at the tails may have large differences in quality whereas players in the middle are likely very close in quality. Thus, the same absolute improvement in quality will result in dramatically different changes in ordinal rankings, depending on where in the distribution the player begins. To fix the problem with ordinal rank as a measure I created a model that uses a log normal distribution to construct a cardinal measure of quality to rank players.

An important part of constructing this model is addressing the attenuation bias that comes from high school rank being an imperfect measure of true ability. This bias will make coaches at top programs seemingly better and coaches at lower ranked programs seemingly worse. Thus, the model I constructed includes a correction for the effects of this attenuation bias to provide consistent estimates of the true effects.

Finally, this study utilizes the model to determine if college football coaches develop the quality of their players. To do this, a regression analysis was conducted to determine the coaching effects on a player when they are drafted into the National Football League (NFL), controlling for where they are ranked when they started as recruits. The results of the show that coaching effects are generally small and negative.

College football coaches were selected as the group of observation for this study for several reasons. In line with the motivation for this study, college football coaches are highly paid.

Colbert and Eckard (2015) also noted that top coaches are often paid well above their university president's pay. Additionally, there has been a large inflation adjusted increase in coaches pay (Colbert & Eckard, 2015). This makes the study of college football coaches' effectiveness an issue of high monetary value. College football presents the ideal group to study because there is a publicized and defined measure of player development available by assessing the difference between the ranking a college recruit is assigned leaving high school to their ranking in the professional draft leaving college. Additionally, because players are required to be three years out of high school before they can be drafted into the NFL, there is plenty of opportunity for player development by coaches during the college years.

The results of this study showed that the effect of college coaches on players after correcting for attenuation bias is small and often negative for schools when compared to Alabama. The attenuation bias factor was large, justifying the correction developed in the model. This significantly increased the effect of high school ability on college ability, holding school constant.

1.1 Literature Review

This study is grounded in two economic theories. The first is the principal-agent relationship described by Tuckman and Tuckman (1976) with regards to university faculty. In this relationship they describe the school administrators as the principals who must incentivize their faculty as the agents to focus on institution enhancing activities rather than personal interests. To do this, faculty are given distinct compensation increases for publications because they are a measurable output of faculty contributing to the institution. When applied to college football, the school administrators are the principals, and the coaches are the agents. This relationship requires school administrators to determine a metric, such as wins, to monitor coaches and ensure that they dedicate time and effort to the desired activities (Tuckman & Tuckman, 1976). Thus, it is difficult to place value on unmeasurable job responsibilities such as player development. This might indicate that college coaches focus on, and thus effect on, player development is low.

The second economic theory is a monopsony as described by Colbert & Eckard (2015) with regards to the NCAA. The NCAA is a monopsony, or single employer, of college players. This market allows colleges to collect economic rents from players which are transferred to schools and athletic departments as well as the coaches (Colbert & Eckard, 2015). This theory explains the fact the college coaches receive such high levels of compensation because

there are high profits left over when players are paid below the wage of a competitive labor market. Furthermore, it provides motivation for the idea that although players are paid below market wages, they may receive non-monetary compensation in the form of development from their coaches.

Understanding coaches' effect of player development begins with identifying potential salary determinants which might motivate coaches to be more, or less, focused on player development. Education presents a useful field of comparison because faculty-like coaches have measurable outcomes but are also focused on developing their students. Gomez-Mejia and Balkin (1992) conducted a regression analysis of salaries for professors of management. They found that publications have a positive relationship on faculty pay, but not teaching (Gomez-Mejia & Balkin, 1992). Tuckman and Tuckman (1976) found similar results of publications on salaries in their study of professors across the country. They extended the analysis to determine that premium salaries are available to professor who have exceptional measurable output in the form of publications. They also determined that teaching could increase salaries after high outcomes in publications are achieved. The researchers postulate that the emphasis on output is because comparisons of school quality are correlated with publications, while teaching success is only measured within a school (Tuckman & Tuckman, 1976). This research indicates that universities do not compensate their faculty for developing students, only measurable outcomes as indicated by the principal-agent relationship. However, there are limitations to comparing college coach with professors as they are allocated funding differently, with faculty often earning much less than highly paid college coaches.

As previously mentioned, another group of individuals who are highly paid and studied for their compensation structure as a result are CEOs. Mixon et al. (2013) reviewed past CEO studies and noted that it was limited to case studies on individuals, a strategy, or team interaction, due to difficulty measuring for variables. To address this limitation, they drew a relationship between CEOs and college coaches. The authors explained that in addition to high salaries both coaches and CEOs are held to very high expectations because of their role and the publicity that it receives, have high turnover rates, and have bonus incentives for their outcome measure. Lastly, they highlight the importance of this issue because previous studies have shown that better CEO incentives result in higher returns (Mixon, Jennings & Wright, 2013). This work highlights the importance in understanding where college coaches are effective and provides precedence for this study to use college coaches to understand managers in the broader labor market.

With this background on salaries in other professions it is then helpful to return to the question of salary determinants of college coaches. As discussed previously, Mixon, Jennings &

Wright (2013) used college coaches to answer questions about CEOs' salaries. Specifically, they wanted to determine the relationship between individual output and pay. They conducted a regression to estimate college football coaches' wages based on football program revenue before the season to establish causality and then estimated worker productivity, measured by wins, based on coach wages. They found that talent as measured by the number of players that make conference teams leads to a higher winning percentage and not the wage that a coach is paid. This indicates that institutions need strong talent development programs rather than paying above market wages. These results are among the few that support the value proposition of player development for college coaches which is important to the positive findings expected in this study. The authors also noted that even college coaches' wages can be based on factors that are not measurable, including fit or personality and connections (Mixon, Jennings & Wright, 2013). This presents limitations to this studies ability to capture coaches effecting player development.

Other studies of college coaches attempt to predict pay based on different variables. Colbert & Eckard (2015) studied changes in coaching staff to establish causality of pay on performance where pay is a proxy for skill. They found that higher coach salaries lead to better performance as measured by an increase in ratings points. Specifically, a one rating point increase for a team is valued at around one million dollars at top schools. However, they also noted diminishing marginal returns to performance for increasing pay. Performance can also be affected by a school's commitment to football measured with variables such as operating expenditures and past success (Colbert & Eckard, 2015). Leeds and Pham (2020) took a different approach using coach performance to determine pay. Like many studies conducted on college players, they estimated the marginal revenue product of college coaches to determine their pay. They found that more bowl game appearances and higher rankings increase salary. In their analysis they noted that part of the reason college coaches are paid so highly is due to their bargaining power. As a result, they attempted to account for salary bargaining power with an equation estimating rehiring based on winning percentage. These studies, like those in the field of education, show that measurable outcomes can and will impact salary. (Leeds & Pham, 2020).

Next, it is helpful to understand how coaches can qualitatively effect player development. This topic is studied in education and sports psychology literature. Naylor (2007) conducted a literature review of previous findings on player development as they age in the field of education. His paper describes the fact that by the time players reach college age they value winning more than simply playing the sport. However, coaches must balance player's feelings of contributing to a game and the emphasis placed on the outcome. Additionally,

players at this age also use observation of other players to improve their skills more than technical coaching (Naylor, 2007). This may indicate that college coaches should in fact place most of their focus on winning because players will develop on their own at that level. Giacobbi Jr et al. (2002) conducted a psychology case study to determine the behavior of hall of fame basketball coach John Wooden in practice. This case study showed that over 50% of Wooden's comments in practice were instructional (Giacobi, et al., 2002). Thus, the literature on player development supports the fact that great coaches incorporate instruction for player development.

The final area of research explored for this study was recruiting in college football as this constructs the variable of interest for this study—player development. Caro (2012) conducted a regression analysis of the effect of college recruit rankings on a team. Rankings as a measure of player development is the same that will be used to answer the research question in this study. He found that successful recruiting—as measured by higher average recruit rankings—leads to successful play—as measured by winning percentage. Increasing a team winning percentage is important because this increases revenue for a school and their athletics (Caro, 2012). Like Mixon, Jennings & Wright (2013), this paper provides support for the fact that player development is important in improving a team's winning percentage, which in turn is important to the institutions' goals of making profit in athletics. This bolsters the hypothesis that coaches effect player development.

2 Model

In order to look at the effect of college coaches on player development, each player needs to be assigned a quality measure. The high school recruit and draft rankings provide a proxy in understanding the quality order of players. However, they are ordinal measures and thus cannot reflect true quality differences among all players. I developed a model that defines a player's true quality in terms of a normal distribution. I then relate each player's true quality percentile, which is a cardinal measure, to the rank data that is available, which is an ordinal measure. The model I created also implies a simple linear relationship between perceived quality in high school and perceived quality in college. As a result, I am able to run an OLS regression on the constructed cardinal measures. Importantly, I include a correction for attenuation bias which occurs because in the model high school effects are correlated with the error term.

First, I establish a definition for perceived quality in high school and college. There are N

players. Each player eventually attends one of S colleges. Player i attended high school h and college s . Thus, each variable is identified by is for player i in college s . Each player has innate ability θ_{is} (which may include the player's intrinsic motivation/effort/training). This can be interpreted as innate player ability for player i in college s . Suppose $\theta_{is} \sim \text{Lognormal}(\mu_\theta, \sigma_\theta^2)$, where $\mu, \sigma > 0$. The log normality assumption implies that log-theta is has a normal distribution $N(\ln \mu_\theta, \sigma_\theta)$. The log-normal distribution is used assuming there is a skew in the distribution of quality, with the best players being very good. The log-normal distribution is also useful in constructing the model because the product of log-normal distributions is log-normal distributed. Additionally, a normal distribution can be used in calculations and converted to log-normal by taking the log. Since there is no natural scale for player talent and the goal is to determine relative changes rather than absolute values, set $\mu_\theta = 1$ such that $\theta_{is} \sim \text{Lognormal}(1, \sigma_\theta^2)$.

2.1 High School

During high school, each player develops a perceived quality based on their history of play. For simplicity, I assume that this perceived quality has two components: the player's innate ability, and an idiosyncratic external shock that may cause the player to appear better or worse than they truly are. For example, the perception of college recruiters may depend on a player's performance during a few select games, which may be affected by weather, illness, particularities of the opponent team, etc. During high school, player i has quality defined by:

1. Innate ability — $\theta_{is} \sim \text{Lognormal}(\mu_\theta, \sigma_\theta^2)$
2. Idiosyncratic External Shock — $\epsilon_{is}^h \sim \text{Lognormal}(\mu_h, \sigma_h^2)$ is a multiplicative shock that captures all of the factors that might cause a high school player to appear stronger or weaker than their true talent. These are perception shocks because the talent observed is not the true talent. Note that ϵ_{is}^h might include high school coach effects. Again, because the goal is to determine relative rather than absolute changes, set $\mu_h = 1$ such that $\epsilon_{is}^h \sim \text{Lognormal}(1, \sigma_h^2)$.

A player's perceived quality at the end of high school, being recruited to college is both innate ability and external factors:

$$V_{is} = \theta_{is} \epsilon_{is}^h$$

I assume ε_{is}^h and θ_{is} are independent. Given the multiplicative nature of shocks and the properties of the Lognormal distribution,

$$\ln V_{is} \sim N(0, \sigma_\theta^2 + \sigma_h^2)$$

Notice that each player's perceived quality is independent of the college they attend.

V_{is} is not observed, but what the econometrician would like to measure. College coaches who recruit players may have some sense of V_{is} that they have developed in the recruiting process from watching the player. However, the only data observed is the player's rank in the high school draft is observed. I assume that a players' rank in the draft perfectly correlates with their perceived quality. If N is large, then by the law of large numbers, the empirical distribution of the players' perceives qualities should match the theoretical distribution that I derived above. Hence, knowing each player's rank in the distribution, I can assign a likely quality measure commensurate to that rank. This quality measure is simply the Z score that corresponds to the relevant percentile in the normal distribution. Let n_{is}^h be player i (who ends up at college s) 's rank, where 1 is the highest rank, and N is the lowest. Then player i 's perceived quality V_{is} is given by:

$$\begin{aligned} \Pr[V \geq V_{is}] &= \frac{n_{is}^h}{N} \\ \Phi\left(\frac{\ln V_{is}}{\sqrt{\sigma_\theta^2 + \sigma_h^2}}\right) &= 1 - \frac{n_{is}^h}{N} \\ V_{is} &= e^{\sqrt{\sigma_\theta^2 + \sigma_h^2} \Phi^{-1}\left(1 - \frac{n_{is}^h}{N}\right)} \end{aligned}$$

Thus, V_{is} is a true measure of quality. Define η_{is}^h as the cardinal measure of V_{is} inferred from the ordinal rank assigned to a player.

$$\eta_{is}^h = e^{\sqrt{\sigma_\theta^2 + \sigma_h^2} \Phi^{-1}\left(1 - \frac{n_{is}^h}{N}\right)}$$

Here V_{is} is proportional to η_{is}^h . η_{is}^h is the Z score associated with a player assigned by knowing their rank and determining their percentile on a normal distribution.

2.2 College

During college, each player develops a new perceived quality based on their history of play in college. I assume that this perceived quality has three components: the player's innate

ability which is the same as it was in high school, an idiosyncratic external shock that is like the high school effect but unique to the college experience, and a coaching effect that is the result of the college the player attends. Thus, the perception of NFL recruiters will take into account a player's development in college as a result of their coach and any other factors that may affect the college game conditions a recruiter happens to see. Player i with innate ability $\theta_{is} \sim \text{Lognormal}(\mu_\theta, \sigma_\theta^2)$ is recruited by college s . During his time at college s , player i 's perceived quality changes. There are two sources of change:

1. Idiosyncratic External Shock $\varepsilon_{is}^S \sim \text{Lognormal}(\mu_s, \sigma_s^2)$ is similar to the high school multiplicative shock. This encapsulates all factors of the player's growth that is independent of the college he attends. Set $\mu_s = 1$ such that $\varepsilon_{is}^S \sim \text{Lognormal}(1, \sigma_s^2)$.
2. Coach Effect $c_s \sim \text{Lognormal}(\mu_c, \sigma_c^2)$ is a common multiplicative shock that all players at college s receive from their coach. This captures the role of the college/coach in bringing out each player's potential. Again, because the goal is to determine relative rather than absolute changes, set $\mu_c = 1$ such that $c_s \sim \text{Lognormal}(1, \sigma_c^2)$.

For a given school, s , a player's perceived quality at the end of college is made up of their innate ability, external shocks, and coaching effects:

$$W_{is} = \theta_{is} \cdot \varepsilon_{is}^S c_s$$

Given the multiplicative nature of shocks:

$$\ln W_{is} \sim N(0, \sigma_\theta^2 + \sigma_s^2 + \sigma_c^2)$$

Again, W_{is} is not observed, but it is likely that NFL coaches have some sense of W_{is} from watching the players in college. Instead each player's rank in the NFL draft is observed. For a large N I can determine the score assigned to a player from their rank as I did for high school. Let n_{is}^c denote the player's ranking at the college level draft. Using the same approach as before:

$$\begin{aligned} \Pr[W \geq W_{is}] &= \frac{n_{is}^c}{N} \\ \Phi\left(\frac{\ln W_{is}}{\sqrt{\sigma_\theta^2 + \sigma_h^2 + \sigma_s^2}}\right) &= 1 - \frac{n_{is}^c}{N} \\ W_{is} &= e^{\sqrt{\sigma_\theta^2 + \sigma_s^2 + \sigma_c^2} \Phi^{-1}\left(1 - \frac{n_{is}^c}{N}\right)} \end{aligned}$$

Thus, W_{is} is a true measure of quality. Define η_{is}^c as the cardinal measure of W_{is} inferred from the ordinal rank.

$$\eta_{is}^c = e^{\sqrt{\sigma_\theta^2 + \sigma_s^2 + \sigma_c^2} \Phi^{-1} \left(1 - \frac{n_{is}^c}{N} \right)}$$

Here W_{is} is proportional to η_{is}^c . η_{is}^c is the Z score associated with a player assigned by knowing their rank and determining their percentile on a normal distribution.

2.3 Comparison

By construction, the perceived quality of the average player does not change between high school and college - although it will change from player to player. These changes may be due to both idiosyncratic effects and development by college level coaches. The basic idea reflected in the derivation below, is that the average change in perceived quality among players from a given school can be attributed to that school's coaching program.

$$\ln W_{is} = \ln V_{is} + \ln c_s + (\ln \epsilon_{is} - \ln \epsilon_{ih})$$

In order to isolate the coach effects c_s for a player from high school h to college s divide high school ability V_{is} from college ability W_{is} . This will net out θ . Player i (at school s)'s perceived improvement during College is:

$$\frac{\epsilon_{is}^s}{\epsilon_{is}^h} c_s = \frac{W_{is}}{V_{is}} = e^{\sqrt{\sigma_\theta^2 + \sigma_s^2 + \sigma_c^2} \Phi^{-1} \left(1 - \frac{n_{is}^c}{N} \right) - \sqrt{\sigma_\theta^2 + \sigma_h^2} \Phi^{-1} \left(1 - \frac{n_{is}^h}{N} \right)}$$

Taking logs of both sides gives:

$$\sqrt{\sigma_\theta^2 + \sigma_s^2 + \sigma_c^2} \Phi^{-1} \left(1 - \frac{n_{is}^c}{N} \right) - \sqrt{\sigma_\theta^2 + \sigma_h^2} \Phi^{-1} \left(1 - \frac{n_{is}^h}{N} \right) = \ln c_s + (\ln \epsilon_{is}^s - \ln \epsilon_{is}^h)$$

This implies:

$$\Phi^{-1} \left(1 - \frac{n_{is}^c}{N} \right) = \sqrt{\frac{\sigma_\theta^2 + \sigma_h^2}{\sigma_\theta^2 + \sigma_s^2 + \sigma_c^2}} \Phi^{-1} \left(1 - \frac{n_{is}^h}{N} \right) + \frac{1}{\sqrt{\sigma_\theta^2 + \sigma_s^2 + \sigma_c^2}} \ln c_s + \frac{1}{\sqrt{\sigma_\theta^2 + \sigma_s^2 + \sigma_c^2}} (\ln \epsilon_{is}^s - \ln \epsilon_{is}^h)$$

Finally, normalize the standard deviations such that $\sigma_\theta^2 + \sigma_s^2 + \sigma_c^2 = 1$ and substitute η_{is} .

This gives:

$$\eta_{is}^c = \alpha \eta_{is}^h + \ln c_s + u_{is}$$

where $\eta_{is}^c = \frac{W_{is}}{\sqrt{\sigma_\theta^2 + \sigma_s^2 + \sigma_c^2}} = \Phi^{-1} \left(1 - \frac{n_{is}^c}{N} \right)$ and η_{is}^h is similarly defined, $u_{is} = \ln \varepsilon_{is}^s - \ln \varepsilon_{is}^h$, and $\text{var}(u_{is}) = \sigma_h^2 + \sigma_s^2$.

This equation defines observable player quality in college η_{is}^c in terms of a constructed coefficient α and observable player quality in high school η_{is}^h plus coaching effects and error from external factors.

To estimate the model, note that η_{is} is observable with the distribution of perceived qualities and the ordinal rank of each player, so a Z score can be assigned to each player. Now this is a simple linear estimation, of η_{is}^c on η_{is}^h with fixed effects.

By construction, the weight average value of $\ln c_s$ (across different schools) will be zero. Hence, each fixed effect $\ln c_s$ can be interpreted as the contribution of each coach/school. This will give the likely bump in both perceived quality and ranking in the draft that come from having a one standard deviation increase in coaching effectiveness.

2.4 Correction for Attenuation Bias

With attenuation bias the naive OLS coefficients and standard errors will be mis-estimated.

The model is:

$$\begin{aligned} \eta_i^s &= \alpha \eta_i^h + c_s + e_i \\ &= \alpha(\theta + \varepsilon_i^h) + c_s + (\varepsilon_i^s - \varepsilon_i^h) \end{aligned}$$

ε_i^h , perception shock from external factors, is in η_{is}^h constructed from $\ln V_{is}$ and in the error term e_i . Because high school effects η_{is}^h are negatively correlated with the error term e_i , which affects the estimation of α . For attenuation bias, $\hat{\alpha}$ is attenuated towards 0. This bias will make coaches at top programs seemingly better and coaches at lower ranked programs seemingly worse.

I allow for a correlation between true quality θ and college attended such that better players are more likely to get recruited to better schools. Without altering this implication, I assume that ε_i^h is independent of which college the player attends. This implies that external shocks, or luck, doesn't affect the school a player attends. As a result, players who attend the top programs are no more likely to get a high ε_i^h than players who attend a mediocre program.

Let $\bar{\eta}_s^s$ denote the average η_i^s at college s . Similarly, define $\bar{\eta}_s^h$ as the average η_i^s amongst players at school s . The naive OLS estimates before correcting for attenuation bias of $\hat{\alpha}$ and \hat{c}_s are given by:

$$\hat{\alpha} = \frac{\frac{1}{n} \sum_i (\eta_i^s - \bar{\eta}_s^s)(\eta_i^h - \bar{\eta}_s^h)}{\frac{1}{n} \sum_i (\eta_i^h - \bar{\eta}_s^h)^2}$$

$$\hat{c}_s = \bar{\eta}_s^s - \hat{\alpha} \bar{\eta}_s^h$$

It is clear that when $\hat{\alpha}$ is mis-estimated it will affect the estimation of the coaching fixed effects \hat{c} because it is a component in that calculation. Since attenuation bias causes $\hat{\alpha}$ to be smaller than it should be, schools with higher quality high school recruits (i.e. with $\bar{\eta} > 0$) will have upwardly biased estimates of coaching effect, and schools with lower quality high school recruits (i.e. with $\bar{\eta} < 0$) will have downwardly biased estimates. So the effect of coaching will be dramatically over-stated.

Next, calculate the correction factor necessary to correct the biased estimates collected in the naive OLS estimation. In order to do this solve for $\hat{\alpha}$ in terms of α .

First, solve for an equation in terms of η and α to substitute into $\hat{\alpha}$. Let n_s denote the number of observations in school s , and define $var(\bar{\eta}_s^h) = \sum_s \frac{n_s}{n} (\bar{\eta}_s^h - \bar{\eta}^h)^2$. ($\bar{\eta}^h = 0$ by construction.) By the Law of Total Variance, $var(\eta_i^h) = var(\eta_i^h - \bar{\eta}_s^h) + var(\bar{\eta}_s^h)$.

$$\bar{\eta}_s^s = \frac{1}{n_s} \sum_{i \in s} (\alpha \eta_i^h + c_s + e_i)$$

$$= \alpha \bar{\eta}_s^h + c_s + \bar{e}_s$$

and so:

$$\eta_i^s - \bar{\eta}_s^s = (\alpha \eta_i^h + c_s + e_i) - (\alpha \bar{\eta}_s^h + c_s + \bar{e}_s)$$

$$= \alpha (\eta_i^h - \bar{\eta}_s^h) + (e_i - \bar{e}_s)$$

[Note that unlike the standard attenuation bias model, there is no α term in e_i .]

Returning to the expression for $\hat{\alpha}$ substitute the above. This gives:

$$\begin{aligned}
\hat{\alpha} &= \frac{\frac{1}{n} \sum_i (\eta_i^s - \bar{\eta}_s^s)(\eta_i^h - \bar{\eta}_s^h)}{\frac{1}{n} \sum_i (\eta_i^h - \bar{\eta}_s^h)^2} \\
&= \frac{\frac{1}{n} \sum_i [\alpha[\eta_i^h - \bar{\eta}_s^h] + (e_i - \bar{e}_s)] (\eta_i^h - \bar{\eta}_s^h)}{\frac{1}{n} \sum_i (\eta_i^h - \bar{\eta}_s^h)^2} \\
&= \alpha + \frac{\frac{1}{n} \sum_i (e_i - \bar{e}_s)(\eta_i^h - \bar{\eta}_s^h)}{\frac{1}{n} \sum_i (\eta_i^h - \bar{\eta}_s^h)^2} \\
&\rightarrow \alpha + \frac{cov(e_i, \eta_i^h - \bar{\eta}_s^h)}{var(\eta_i^h - \bar{\eta}_s^h)} \\
&= \alpha + \frac{cov(\varepsilon_i^s - \varepsilon_i^h, \theta_i - \bar{\theta}_s + \varepsilon_i^h)}{var(\theta_i - \bar{\theta}_s + \varepsilon_i^h)} \\
&= \alpha - \frac{\sigma_h^2}{\underbrace{\sigma_\theta^2 - var(\bar{\eta}_s^h)}_{=var(\theta_{is} - \bar{\theta}_s)} + \sigma_h^2}
\end{aligned}$$

Define the constant attenuation factor:

$$\lambda = \frac{\sigma_h^2}{\sigma_\theta^2 - var(\bar{\eta}_s^h) + \sigma_h^2}$$

such that $\hat{\alpha} = \alpha - \lambda$.

Note that \hat{c}_s is biased because $\hat{\alpha}$ is not estimated correctly. The estimated \hat{c}_s 's will be more spread out such that if $\bar{\eta}_s$ is positive \hat{c}_s is more positive. For the fixed effects bias return to the original equation and substitute:

$$\begin{aligned}
\hat{c}_s &= \bar{\eta}_s^s - \hat{\alpha} \bar{\eta}_s^h \\
&= (\alpha \bar{\eta}_s^h + c_s + \bar{e}_s) - \hat{\alpha} \bar{\eta}_s^h \\
&= c_s + (\alpha - \hat{\alpha}) \bar{\eta}_s^h + \bar{e}_s \\
&\rightarrow c_s + \lambda \bar{\eta}_s^h
\end{aligned}$$

A mis-estimated $\hat{\alpha}$ and \hat{c} will also affect the variances and error term calculated by the naive OLS. In order to correct those measures solve for $var(\hat{c}_s)$, \hat{e}_i , and $\hat{\sigma}_e^2$ using the attenuation factor λ .

Since $\bar{e}_s \rightarrow 0$ as $n \rightarrow \infty$. (The average measurement error is zero for each school.)

$$var(\hat{c}_s) \rightarrow var(c_s) + 2\lambda cov(c_s, \bar{\eta}_s^h) + \lambda^2 var(\bar{\eta}_s^h)$$

This expression includes the term $cov(c_s, \bar{\eta}_s^h)$ which is not directly observed. But $cov(\hat{c}_s, \bar{\eta}_s^h)$ can be computed.

$$\begin{aligned} cov(\hat{c}_s, \bar{\eta}_s^h) &\rightarrow cov(c_s + \lambda \bar{\eta}_s^h, \bar{\eta}_s^h) \\ &= cov(c_s, \bar{\eta}_s^h) + \lambda var(\bar{\eta}_s^h) \end{aligned}$$

Hence $cov(c_s, \bar{\eta}_s^h) \rightarrow cov(\hat{c}_s, \bar{\eta}_s^h) - \lambda var(\bar{\eta}_s^h)$. Then:

$$\begin{aligned} var(\hat{c}_s) &\rightarrow var(c_s) + 2\lambda cov(c_s, \bar{\eta}_s^h) + \lambda^2 var(\bar{\eta}_s^h) \\ &= var(c_s) + 2\lambda [cov(\hat{c}_s, \bar{\eta}_s^h) - \lambda var(\bar{\eta}_s^h)] + \lambda^2 var(\bar{\eta}_s^h) \\ &= var(c_s) + 2\lambda cov(\hat{c}_s, \bar{\eta}_s^h) - \lambda^2 var(\bar{\eta}_s^h) \end{aligned}$$

Thus, the residual is:

$$\begin{aligned} \hat{e}_i &= \eta_i^s - (\hat{\alpha} \eta_i^h + \hat{c}_s) \\ &= (\alpha \eta_i^h + c_s + e_i) - (\hat{\alpha} \eta_i^h + \hat{c}_s) \\ &= e_i + (\alpha - \hat{\alpha}) \eta_i^h + (c_s - \hat{c}_s) \\ &= e_i + (\alpha - \hat{\alpha}) \eta_i^h - (\alpha - \hat{\alpha}) \bar{\eta}_s^h - \bar{e}_s \\ &= (e_i - \bar{e}_s) + (\alpha - \hat{\alpha}) (\eta_i^h - \bar{\eta}_s^h) \\ &\rightarrow e_i + \lambda (\eta_i^h - \bar{\eta}_s^h) \end{aligned}$$

Hence, the estimated variance of the equation error is:

$$\hat{\sigma}_e^2 \rightarrow \sigma_e^2 + \lambda^2 var(\bar{\eta}_s^h)$$

and recall that $\sigma_e^2 = var(\varepsilon_i^s - \varepsilon_i^h) = \sigma_s^2 + \sigma_h^2$.

So far, the definition of the attenuation factor has been used to define corrected estimates for α , c , e , and the variances. Next, it is necessary to calculate λ using the estimates from

the naive OLS. Bringing everything together gives the following:

$$\begin{aligned}
(1) \quad & \sigma_\theta^2 + \sigma_s^2 + \text{var}(c) = 1 \\
(2) \quad & \sigma_\theta^2 + \sigma_h^2 = \alpha^2 \\
(3) \quad & \sigma_s^2 + \sigma_h^2 = \sigma_e^2 \\
(4) \quad & \lambda = \frac{\sigma_h^2}{\sigma_\theta^2 - \text{var}(\bar{\eta}_s^h) + \sigma_h^2} \\
(5) \quad & \alpha - \hat{\alpha} \rightarrow \lambda \\
(6) \quad & \hat{\sigma}_e^2 \rightarrow \sigma_e^2 + \lambda^2 \text{var}(\bar{\eta}_s^h) \\
(7) \quad & \text{var}(\hat{c}_s) \rightarrow \text{var}(c_s) + 2\lambda \text{cov}(\hat{c}_s, \bar{\eta}_s^h) - \lambda^2 \text{var}(\bar{\eta}_s^h)
\end{aligned}$$

The first 3 equations come directly from the model. Equations (4) and (5) define the attenuation factor. Equations (6) and (7) provide the asymptotic variance of the OLS residual and the OLS estimates of the fixed effects.

This gives have 7 equations that involve 7 unknown variables: σ_θ^2 , σ_h^2 , σ_s^2 , $\text{var}(c)$, α , λ , and σ_e^2 . All the remaining terms in the 7 equations are either fixed numbers or terms that can be computed/estimated, including: $\hat{\alpha}$, $\hat{\sigma}_e^2$, $\text{var}(\hat{c}_s)$, $\text{cov}(\hat{c}_s, \bar{\eta}_s)$, $\text{var}(\bar{\eta}_s)$. Given the 7 equations, it is possible to back out values for each of the 7 unknown terms. In this case, manipulating the equations for σ_θ^2 and σ_h^2 will result in an equation that can be solved for α . Because λ is necessary to correct all of the other measures from the naive OLS, equation (5) can be used to back out λ at the end.

I start by substituting equation (2) into equation (4) to get σ_h^2 in terms of λ . This gives:

$$\begin{aligned}
\sigma_h^2 &= \lambda [\sigma_\theta^2 - \text{var}(\bar{\eta}_s^h) + \sigma_h^2] \\
&= \lambda [(\alpha^2 - \sigma_h^2) - \text{var}(\bar{\eta}_s^h) + \sigma_h^2] \\
&= \lambda [\alpha^2 - \text{var}(\bar{\eta}_s^h)]
\end{aligned}$$

and so (by (2) again):

$$\begin{aligned}
\sigma_\theta^2 &= \alpha^2 - \lambda [\alpha^2 - \text{var}(\bar{\eta}_s^h)] \\
&= (1 - \lambda)\alpha^2 + \lambda \text{var}(\bar{\eta}_s^h)
\end{aligned}$$

Hence, the expressions for σ_h^2 and σ_θ^2 is in terms of λ and observable variables.

Next, I take equation (7). At line 2, I substitute for $\text{var}(c_s)$ using (1). Then at line 3, I

substitute σ_s^2 using (3). Then at line 4, I substitute for σ_e^2 using (6). This gives:

$$\begin{aligned}
var(\hat{c}_s) &= var(c_s) + 2\lambda cov(\hat{c}_s, \bar{\eta}_s^h) - \lambda^2 var(\bar{\eta}_s^h) \\
var(\hat{c}_s) &= [1 - \sigma_\theta^2 - \sigma_s^2] + 2\lambda cov(\hat{c}_s, \bar{\eta}_s^h) - \lambda^2 var(\bar{\eta}_s^h) \\
var(\hat{c}_s) &= [1 - \sigma_\theta^2] - (\sigma_e^2 - \sigma_h^2) + 2\lambda cov(\hat{c}_s, \bar{\eta}_s^h) - \lambda^2 var(\bar{\eta}_s^h) \\
var(\hat{c}_s) &= 1 - \sigma_\theta^2 + \sigma_h^2 - (\hat{\sigma}_e^2 - \lambda^2 var(\bar{\eta}_s^h)) + 2\lambda cov(\hat{c}_s, \bar{\eta}_s^h) - \lambda^2 var(\bar{\eta}_s^h) \\
var(\hat{c}_s) &= 1 - \sigma_\theta^2 + \sigma_h^2 - \hat{\sigma}_e^2 + 2\lambda cov(\hat{c}_s, \bar{\eta}_s^h)
\end{aligned}$$

Now, I can use the previous expressions for σ_h^2 and σ_θ^2 . This gives:

$$\begin{aligned}
var(\hat{c}_s) &= 1 - [(1 - \lambda)\alpha^2 + \lambda var(\bar{\eta}_s^h)] + \lambda[\alpha^2 - var(\bar{\eta}_s^h)] - \hat{\sigma}_e^2 + 2\lambda cov(\hat{c}_s, \bar{\eta}_s^h) \\
var(\hat{c}_s) &= 1 + (2\lambda - 1)\alpha^2 - 2\lambda var(\bar{\eta}_s^h) - \hat{\sigma}_e^2 + 2\lambda cov(\hat{c}_s, \bar{\eta}_s^h) \\
var(\hat{c}_s) &= 1 - \hat{\sigma}_e^2 - \alpha^2 + 2\lambda(\alpha^2 + cov(\hat{c}_s, \bar{\eta}_s^h) - var(\bar{\eta}_s^h))
\end{aligned}$$

Finally, using (5) to substitute for λ . This gives:

$$\begin{aligned}
var(\hat{c}_s) &= 1 - \hat{\sigma}_e^2 - \alpha^2 + 2[\alpha - \hat{\alpha}](\alpha^2 + cov(\hat{c}_s, \bar{\eta}_s^h) - var(\bar{\eta}_s^h)) \\
0 &= 2\alpha^3 - (2\hat{\alpha} + 1)\alpha^2 + 2(cov(\hat{c}_s, \bar{\eta}_s^h) - var(\bar{\eta}_s^h))\alpha + [1 - \hat{\sigma}_e^2 - var(\hat{c}_s) - 2\hat{\alpha}cov(\hat{c}_s, \bar{\eta}_s^h) + 2\hat{\alpha}var(\bar{\eta}_s^h)]
\end{aligned}$$

Once α is calculated, λ can be calculated using equation (5) and all the other variables can be recovered one by one.

2.5 Assuming $\sigma_h^2 = \sigma_s^2$

In the following subsection, I conduct a simulation study to check that this method recovers the true model parameters when these are known. The method seemed to work well when the true parameters satisfied $\sigma_h^2 = \sigma_s^2$, however produced biased results otherwise. The method also occasionally produced complex results, which obviously cannot be. I am not entirely clear about why this problem arises, but my guess is that equation (1) is very insensitive to the data, and so the randomness present in a medium-sized data set can cause the model estimates to systematically deviate from their true values.

Though my preference was to estimate the fully flexible model, I instead proceed by forcing the assumption that $\sigma_h = \sigma_s$. This implies that the idiosyncratic shock is as noisy at the

college level as at the high school level. With this assumption, the above system of equations simplifies as follows:

In the main analysis, (1) was used to identify σ_s^2 because it was not known. This also required $var(c)$, which is in (7). With the assumption that $\sigma_s^2 = \sigma_h^2$, equations (1) and (7) can be omitted. Additionally, (3) was used to isolate σ_h^2 (by using the empirical connection between σ_e^2 and $\hat{\sigma}_e^2$) which required knowing σ_s^2 . (3) becomes $2\sigma_h^2 = \sigma_e^2$.

By making the assumption that $\sigma_h^2 = \sigma_s^2$ the manipulation of equation to solve for λ becomes much easier. I started with (6), the asymptotic variance of the OLS residual, and substitute σ_h^2 for σ_e^2 from the simplified (3). Then I substitute for σ_h^2 as solved previously using (2) and (4) in terms of α and λ . Finally, I used (5) to substitute α for $\hat{\alpha}$, which is observable, and λ , which will be calculated.

$$\begin{aligned}\hat{\sigma}_e^2 &= \sigma_e^2 + \lambda^2 var(\hat{\eta}_s^h) \\ &= 2\sigma_h^2 + \lambda^2 var(\hat{\eta}_s^h) \\ &= 2\lambda [\alpha^2 - var(\bar{\eta}_s^h)] + \lambda^2 var(\hat{\eta}_s^h) \\ &= 2\lambda [(\lambda + \hat{\alpha})^2 - var(\bar{\eta}_s^h)] + \lambda^2 var(\hat{\eta}_s^h)\end{aligned}$$

This is a cubic equation purely in terms of one unknown λ and the observables $\hat{\alpha}$ and $var(\bar{\eta}_s^h)$. Importantly, it doesn't involve a calculation of $var(c)$ - which may result in imaginary roots.

λ is the solution to:

$$2\lambda^3 + \lambda^2(4\hat{\alpha} + var(\bar{\eta}_s^h)) + 2\lambda(\hat{\alpha}^2 - var(\bar{\eta}_s^h)) - \hat{\sigma}_e^2 = 0$$

Once λ is calculated, α can be calculated using equation (5) and all the other variables can be recovered one by one.

3 Analysis

With this structural model developed, I then tested to test if the empirics recovered the true parameters. I ran a Monte Carlo simulation to randomly simulate coaching effects. I defined the true coaching fixed effects c once, and then simulated the data 100 times to collect the estimated \hat{c} . In the simulation I created normal distributions for θ_{is} , η_{is}^h , and η_{is}^s . I generated

player data that assigned schools and rank. With the player data I then assigned them a Z score from η_{is}^h , and η_{is}^s . At this stage I then ran the OLS regression to get \hat{c} . In order to correct for attenuation bias I solved the equation developed in the model and ran adjusted the results. I ran the simulation at a small, medium, and large number for each of the parameters: N , S , σ_{θ}^2 , σ_h^2 , σ_s^2 , $var(c)$.

Table 1 shows the results of this simulation for $N = 1000$, $S = 10$. This table shows the values are seemingly recovered, and importantly the direction of the effects is captured. Figure 1 is a box and whiskers plot of the estimated \hat{c} from the simulation showing the median recovered value with the true c . Figure 2 is a scatterplot that shows the true c values plotted against the estimated \hat{c} . The data is clustered along the 10 sample schools and generally the values fall along the red line indicating a perfect fit. (See the appendix for more simulation results.)

School	True log(c)	Mean log(\hat{c})	Std. Dev.
1	-0.0463	0.0116	0.0902
2	-0.0836	-0.0539	0.1143
3	-0.1103	-0.0419	0.2055
4	-0.1416	-0.0888	0.1020
5	0.0103	0.0736	0.0954
6	0.0082	0.0528	0.0859
7	-0.0714	-0.0164	0.1830
8	0.0420	0.0965	0.0991
9	-0.1063	-0.0508	0.2095
10	0.0299	0.0780	0.1243
$\alpha=-0.0154,$		$\beta=1.1148$	

Table 1: Summary Statistics of Monte Carlo Simulation (N=1000)

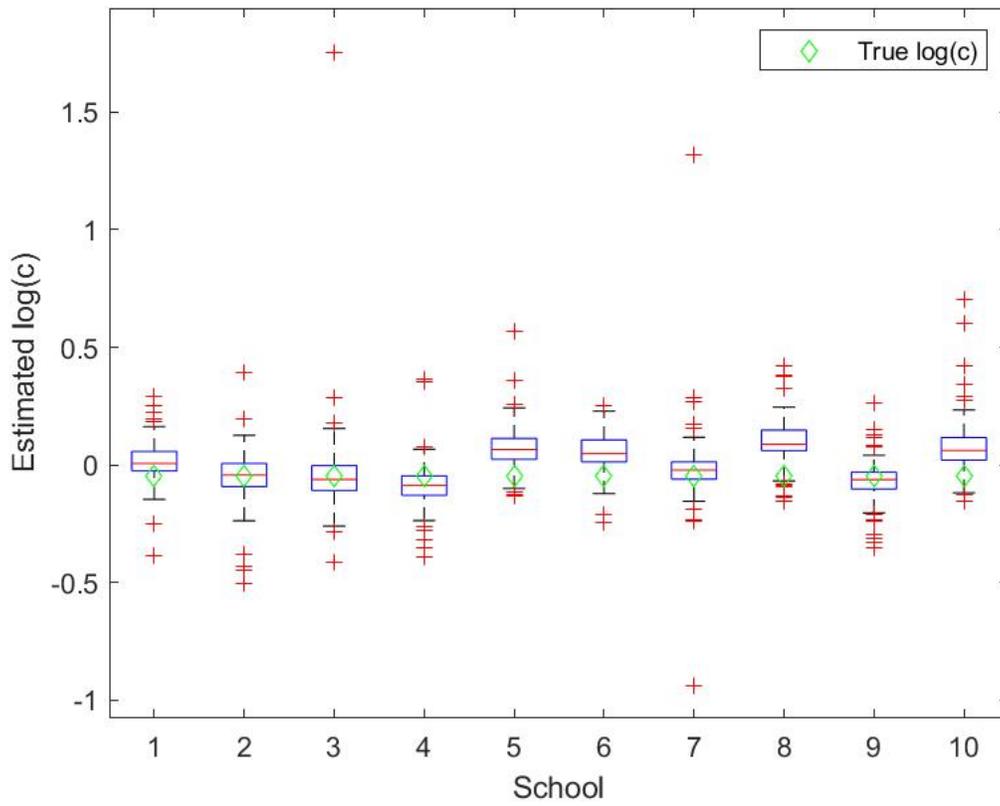


Figure 1: Graph of Estimated $\ln \hat{c}$ Distribution per School

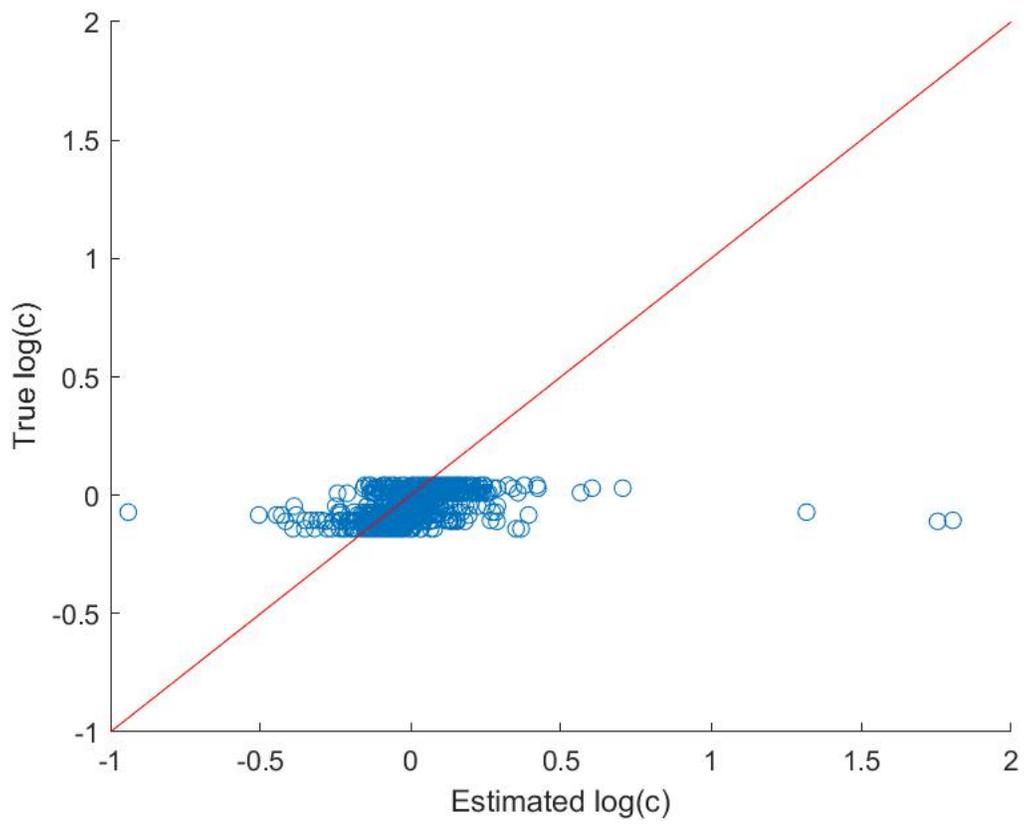


Figure 2: Graph of True $\ln c$ on Estimate $\ln \hat{c}$

Overall, this simulation supports the use of the model as a method of converting ordinal ranks to a continuous measure of quality in determining coaching effects.

4 Data

In order to address the question of the effect of college coaches on player development a new dataset was constructed. The final data set is a pooled cross-sectional dataset of recruited college football players by draft year out of college. The dataset has a player level unit of observation for the years 2008-2020.

ESPN.com has a dataset ranking the top 100 high school recruits from 2008-2022 which was used up to 2016. This ranking was used to create the variable *RK* indicating a player’s high school recruit ranking. Sports-reference.com has a dataset on every college player. This was used to create the variable *Combined2Pick* which is a player’s ranking in the professional draft. These measures operate on comparable scales of rankings except for the fact that the professional draft includes around 250 players each year.

Player observations were merged such that the final data set only has players who were both recruited in high school and drafted professionally. The resulting dataset has 180 players from 67 schools (Table 2). This means that each school has on average approximately 3 player observations, with 29 schools only having 1 player observation. Ohio State has the most player observations (12), followed by Mississippi State (10) and Alabama (9) (Figure 3).

Variable	N	Mean	Std. Dev.	Min	Max
High School Recruit Rank	180	41.45	28.3537	1	100
College Draft Rank	180	121.3278	75.3610	1	256

Source: ESPN.com, Sports-reference.com

Table 2: Summary Statistics

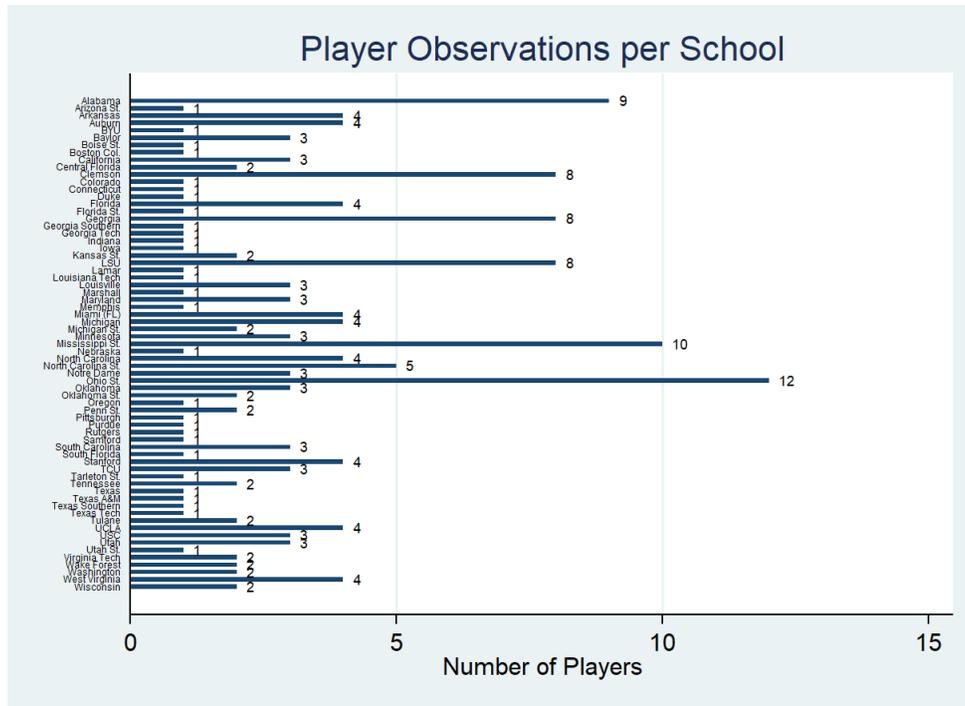


Figure 3: Graph of Player Observations Per School

Lastly, η_{is}^h and η_{is}^s were constructed by assigning a Z score to each player based on their high school and college rank. Figure 4 is a scatterplot of the constructed η_{is}^h and η_{is}^s modeling player quality. The line of best fit for these variables shows a small positive slope but they have a very low correlation of 0.072. It should be noted that this data has not been corrected for attenuation bias.

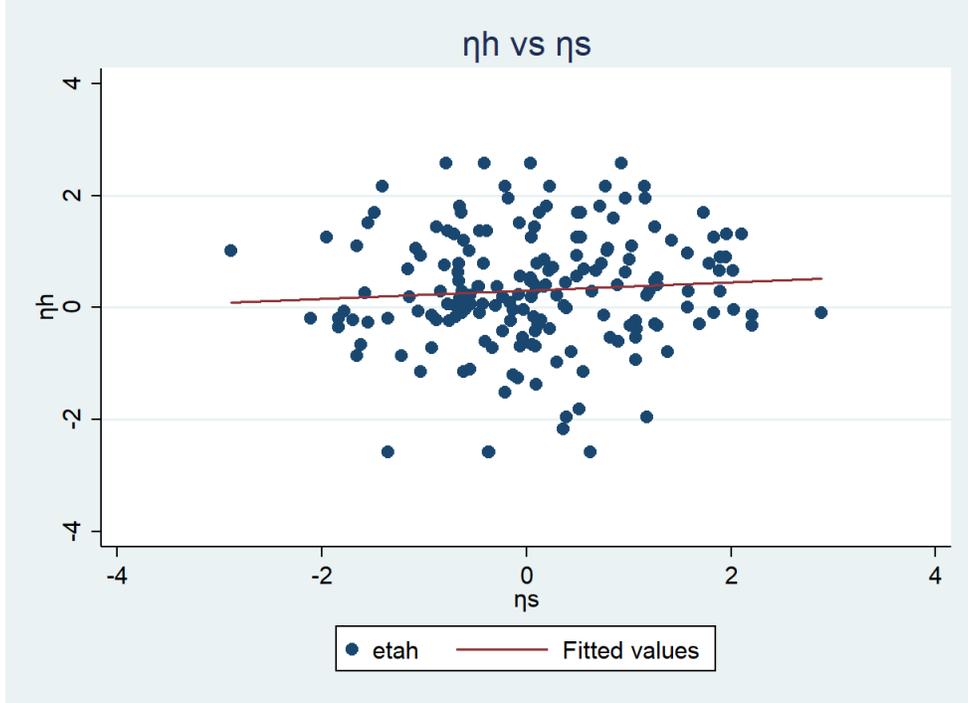


Figure 4: Graph of η_{is}^h and η_{is}^s

5 Methodology

This study aimed to see the effect of college coaches on player development. I used professional draft ranking of football players while controlling for their high school recruit position, as a proxy for player development. The method to ordinal rank to a continuous measure of quality and correct for attenuation bias is detailed in the model.

The main independent variable of interest are the coach fixed effects which indicates the amount that coaches effect player development. I hypothesize that college football coaches are incentivized to focus on player development from recruitment to draft because developing players can improve a winning percentage beyond strategic game-time decisions and high draft picks can bring a program recognition. I expect a positive coefficient on the coach fixed effects indicating that coaches improve player’s rankings in the draft controlling for where they were recruited, and thus develop players.

6 Results

This study utilized a fixed effects regression of constructed player quality η_{is}^h and η_{is}^s for each college coach. Finally, the results collected from the regression were adjusted for the attenuation bias. The attenuation factor $\lambda = 0.835$. This significantly alters the results and their significance from Table 6 to Table 6 and emphasizes the importance of the correction for attenuation bias to accurately measure coaching effects.

Table 6 shows the final results which indicate that coaches have very little effect on player development. The coaching fixed effects coefficients are mostly negative indicating that compared to being coached at Alabama, colleges decrease player quality. The only schools with positive fixed effects are BYU, Baylor, Tarleton State, and Utah State. This is likely because Alabama is a top ranked school with a strong coaching program making other schools look worse in comparison. Furthermore, few of the fixed effects have significant coefficients likely due to the small number of observations. Only BYU, Boston College, North Carolina State, TCU, and Tarleton State are significant at the 10% level. It is notable that in Table 6 under the naive OLS many schools seem to have significant effects on player development but the significance goes away after correction in Table 6. Furthermore, the attenuation bias correction even changes the direction of the coaching effect for six of these schools.

Table 6 also shows player's talent in high school η_{is}^h increases talent in college by 0.968, holding school/coach equal. Importantly, without the correction for attenuation bias, the effect of high school talent on college talent is negative and insignificant (Table 6). Thus, the correction significantly alters the results of this regression. Overall, these results do not support my hypothesis that the coaching fixed effects would be positive. The error term which is composed of the unobservable perception shocks ε_{is}^s and ε_{is}^h indicates that factors other than coaches increase player development by 0.821, holding school equal, and is significant at the 5% level (Table 6).

VARIABLES	η_{is}^c	Std. Error	N	VARIABLES	η_{is}^c	Std. Error	N
η_{is}^h	-0.0168	0.0994		North Carolina State	-1.066*	0.613	5
Arizona State	-0.846	1.095	1	Notre Dame	-0.335	0.695	3
Arkansas	-1.191*	0.621	4	Ohio State	-0.0313	0.464	12
Auburn	-0.335	0.621	4	Oklahoma	-0.381	0.700	3
BYU	0.192	1.098	1	Oklahoma State	0.0146	0.812	2
Baylor	0.371	0.693	3	Oregon	-0.734	1.100	1
Boise State	-1.677	1.090	1	Penn. State	-1.483*	0.826	2
Boston College	-0.493	1.136	1	Pittsburgh	-0.0162	1.090	1
California	-1.154*	0.691	3	Purdue	-1.467	1.090	1
Central Florida	0.105	0.812	2	Rutgers	-0.448	1.094	1
Clemson	-0.443	0.505	8	Samford	-1.956*	1.093	1
Colorado	-1.442	1.092	1	South Carolina	-1.421**	0.702	3
Connecticut	-1.418	1.095	1	South Florida	-0.431	1.091	1
Duke	-0.595	1.099	1	Stanford	0.0523	0.623	4
Florida	-0.576	0.624	4	TCU	-1.081	0.717	3
Florida State	-0.634	1.090	1	Tarleton State	-1.233	1.148	1
Georgia	-0.945*	0.502	8	Tennessee	-0.880	0.809	2
Georgie Southern	-1.451	1.096	1	Texas	-0.977	1.097	1
Georgia Tech	-0.864	1.101	1	Texas A&M	-0.769	1.091	1
Indiana	-1.058	1.100	1	Texas Southern	-0.891	1.093	1
Iowa	-0.713	1.098	1	Texas Tech	-0.979	1.094	1
Kansas State	-0.760	0.809	2	Tulane	-1.127	0.826	2
LSU	0.0157	0.505	8	UCLA	-0.856	0.623	4
Lamar	-0.548	1.090	1	USC	-1.175*	0.690	3
Louisiana Tech	-1.561	1.100	1	Utah	-0.820	0.704	3
Louisville	0.433	0.690	3	Utah State	0.470	1.091	1
Marshall	-2.457**	1.089	1	Virginia Tech	-1.041	0.808	2
Maryland	-0.911	0.692	3	Wake Forest	-1.298	0.821	2
Memphis	-3.690***	1.090	1	Washington	-0.122	0.809	2
Miami	-1.463**	0.626	4	West Virginia	-0.340	0.626	4
Michigan	-1.099*	0.622	4	Wisconsin	-0.237	0.813	2
Michigan State	-0.707	0.810	2	u_{is}	0.821**	0.361	
Minnesota	-2.161***	0.705	3	Observations	180		
Mississippi State	-1.129**	0.494	10	R-squared	0.375		
Nebraska	-1.471	1.093	1				
North Carolina	-1.441**	0.649	4				

*** p<0.01, ** p<0.05, * p<0.1

Table 3: College Coach Differentials (Before Correction for Attenuation Bias)

VARIABLES	η_{is}^c	Std. Error	N	VARIABLES	η_{is}^c	Std. Error	N
η_{is}^h	0.968***	0.244		North Carolina State	-0.204***	0.289	5
Arizona State	-0.815	0.516	1	Notre Dame	-0.453	0.328	3
Arkansas	-1.911	0.293	4	Ohio State	-0.193	0.219	12
Auburn	-1.064	0.293	4	Oklahoma	-0.237	0.330	3
BYU	0.458*	0.518	1	Oklahoma State	-0.187	0.383	2
Baylor	0.086	0.327	3	Oregon	-2.886	0.519	1
Boise State	-2.880	0.514	1	Penn. State	-0.922	0.390	2
Boston College	1.320***	0.536	1	Pittsburgh	-0.864	0.514	1
California	-1.652	0.326	3	Purdue	-2.126	0.514	1
Central Florida	-0.139	0.383	2	Rutgers	-0.479	0.516	1
Clemson	-0.860	0.238	8	Samford	-2.114	0.515	1
Colorado	-1.686	0.515	1	South Carolina	-1.170	0.331	3
Connecticut	-1.387	0.516	1	South Florida	-0.810	0.514	1
Duke	-0.284	0.518	1	Stanford	-0.482	0.294	4
Florida	-1.994	0.294	4	TCU	-0.306**	0.338	3
Florida State	-1.352	0.514	1	Tarleton State	0.919***	0.541	1
Georgia	-1.927	0.237	8	Tennessee	-1.437	0.381	2
Georgie Southern	-1.378	0.517	1	Texas	-0.776	0.517	1
Georgia Tech	-0.414	0.519	1	Texas A&M	-1.171	0.514	1
Indiana	-0.702	0.519	1	Texas Southern	-1.092	0.515	1
Iowa	-0.447	0.518	1	Texas Tech	-1.052	0.516	1
Kansas State	-1.363	0.381	2	Tulane	-0.589	0.390	2
LSU	-0.442	0.238	8	UCLA	-1.301	0.294	4
Lamar	-1.152	0.514	1	USC	-1.759	0.325	3
Louisiana Tech	-3.713	0.519	1	Utah	-0.520	0.332	3
Louisville	-0.135	0.325	3	Utah State	0.020	0.514	1
Marshall	-3.379	0.514	1	Virginia Tech	-1.732	0.381	2
Maryland	-1.295	0.326	3	Wake Forest	-0.962	0.387	2
Memphis	-4.538	0.514	1	Washington	-0.659	0.381	2
Miami	-1.674	0.295	4	West Virginia	-0.555	0.295	4
Michigan	-1.685	0.293	4	Wisconsin	-0.385	0.383	2
Michigan State	-1.087	0.382	2	u_{is}	-0.815*	0.516	
Minnesota	-1.811	0.332	3	Observations	180		
Mississippi State	-0.884	0.233	10	R-squared	0.375		
Nebraska	-1.608	0.515	1				
North Carolina	-0.764	0.306	4				

*** p<0.01, ** p<0.05, * p<0.1

Table 4: True College Coach Differentials

There are some potential threats to the validity of these results. First, the data set of players who were both ranked as a recruit in high school and drafted in college was a very small number spread across a larger number of schools. This resulted in nearly half of the schools having only one play observation used to determine a coaching effect which is an imperfect representation. There are also limitations to the relative measure of player development including injured players, players who drop out, players who do not make the draft based on skill, as well as players who do not change ranking but instead developed in exact alignment with their peers. Lastly, there are limitations to assigning coaches to a player for coaching changes or transfer players.

7 Discussion & Conclusion

This study involved both the construction of a model, as well as an analysis of college football data. The model converting recruit and draft rank into a continuous measure of quality as well as a correction for attenuation bias will have wider implications about how to evaluate value added when performance is mis-measured. This model is useful for studying any features that use ordinal measures or exhibit attenuation bias. This study will contribute to sports economics literature in the areas of wage determination and incentives. When more broadly applied to labor economics it can be viewed as a relevant comparison to salary determination of managers and indicate where emphasis is or is not being placed by the salary.

Overall, this study looks at value added in the context of education and mis-measured performance due to a variety of issues, perhaps including systemic racism. This paper will hopefully show that estimates of value-added are systematically biased and add to the research on this topic.

The next step in this research on coaching effects would be to utilize the coaching effects of player development to determine their impact on college coaches' salaries. The second stage regression could predict the college coach's salary with the independent variable of interest being the coach fixed effects from this study. The literature suggests a positive effect of player development on coaches' salary indicating that coaches who develop players are paid more. However, it is likely that with the coaching fixed effects being so small, the effect on salary will be very low.

This research could also be extended to study development of football players with professional coaches as well. This study also raises the question as to whether those athletes who develop a lot in college will continue to see the same levels of success or development in the NFL. This is an important question in the labor market if employees who develop well under one manager are promoted or moved to a different division. To do this, one could see if the relative measure of player development in college is like that experienced in the NFL. However, this extended analysis would again present challenges in measures of player development during their time in the NFL as well as the natural plateauing effects of athletic development.

The results of this study may have relevance on future NCAA policy when considering regulating the compensation of college coaches or determining how much college players

should be paid. Additionally, it may influence colleges to incorporate measures of player development when creating college coaches' salaries. This could have an extended impact in the larger labor market with adjustments to manager pay to account for developing employees.

References

- Berkowitz, S , J Varney, E Johnson, and S Dougherty (n.d.). College football head coach salaries.
- Caro, Cary A (2012). College football success: The relationship between recruiting and winning. *International Journal of Sports Science & Coaching* 7(1), 139–152.
- Colbert, Gary J and E Woodrow Eckard (2015). Do colleges get what they pay for? evidence on football coach pay and team performance. *Journal of Sports Economics* 16(4), 335–352.
- Giacobbi Jr, Peter R , Joe Whitney, Emily Roper, and Ted Butryn (2002). College coaches' views about the development of successful athletes: A descriptive exploratory investigation. *Journal of Sport Behavior* 25(2).
- Gomez-Mejia, Luis R and David B Balkin (1992). Determinants of faculty pay: An agency theory perspective. *Academy of Management journal* 35(5), 921–955.
- Leeds, Michael A and Ngoc Tram Nguyen Pham (2020). Productivity, rents, and the salaries of group of five football coaches. *Journal of Sports Economics* 21(1), 3–19.
- Mixon, Philip , W Jennings Byrd, and Alan Wright (2013). Does pay lead to performance? using ncaa head football coaches as a surrogate for ceos. *Journal of Business and Behavioral Sciences* 25(2), 25.
- Naylor, Adam H (2007). The coach's dilemma: balancing playing to win and player development. *Journal of Education* 187(1), 31–48.
- Tuckman, Barbara H and Howard P Tuckman (1976). The structure of salaries at american universities. *The Journal of Higher Education* 47(1), 51–64.

Appendices

A Appendix

A Simulation Results

This section contains more results from the Monte Carlo simulation of the model detailed in the Analysis section when $N=200$, $S=10$.

School	True $\log(c)$	Mean $\log(\hat{c})$	Std. Dev.
1	-0.0463	0.0331	0.3441
2	-0.0836	-0.0542	0.1816
3	-0.1103	-0.1058	0.2437
4	-0.1416	-0.1030	0.1807
5	0.0103	0.0719	0.2127
6	0.0082	0.0567	0.2692
7	-0.0714	-0.0555	0.2664
8	0.0420	0.0785	0.2008
9	-0.1063	-0.0823	0.1993
10	0.0299	0.0685	0.2546

$\alpha=-0.2012,$ $\beta=1.0912$

Table 5: Summary Statistics of Monte Carlo Simulation ($N=1000$)

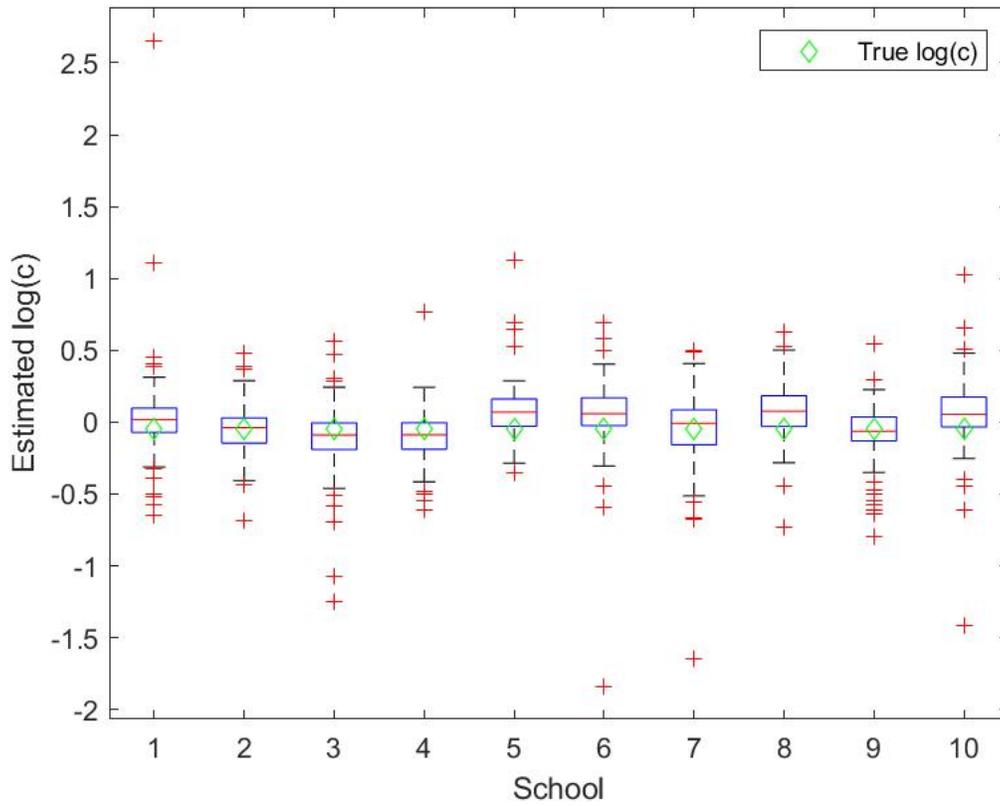


Figure 5: Graph of Estimated $\ln \hat{c}$ Distribution per School

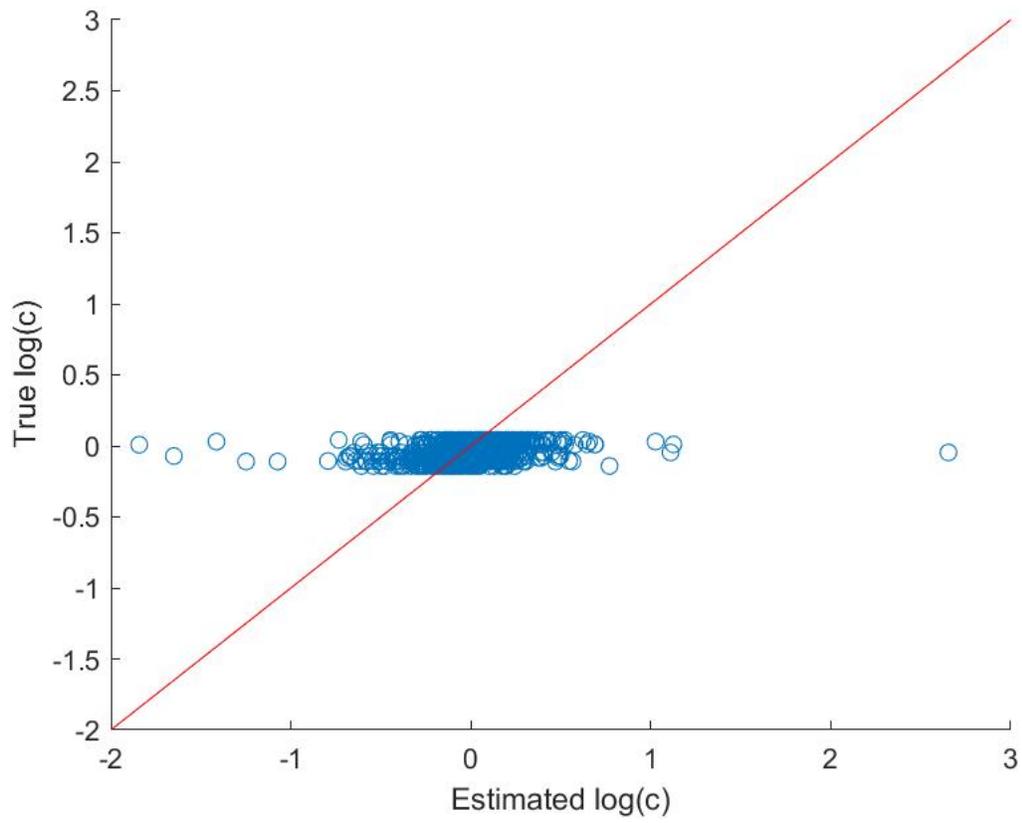


Figure 6: Graph of True $\ln c$ on Estimate $\ln \hat{c}$

B Analysis Code

This section contains the Matlab code used to run the analysis on the model as detailed in the Analysis section. The code is a Monte Carlo simulation on the function simulating player data and regression results with a correction for attenuation bias.

```
clear;
clc;

%[a, r, b1, b, t] = name(500, 10, 0.5, 0.006, 0.2, 0.1)

ns = [200, 500, 1000];
ss = [5, 10, 15];
sts = [0.3, 0.5, 1];
shs = [0.0055, 0.006, 0.0065]; %larger values to create attenuation bias
    on the same magnitude of scs
sss = [0.15, 0.2, 0.25];
scs = [0.05, 0.1, 0.15];

c = normrnd(0,scs(2),ss(2),1);    % True coach effect

mbtable = [];
mttable = [];
stdbtable = [];
table = [];

for x = ns %change to test each parameter
    bs = [];
    ts = [];
    alphas = [];
    betas = [];
    table = [];
    for i=1:100 %simulate data 100 times
        [a, r, b1, b, t, ua, uc] = name(x, 10, 0.5, 0.006, 0.2, 0.1, c);
        %call function
        while isnan(b) %run until no naan
```

```

        [a, r, b1, b, t, ua, uc] = name(x, 10, 0.5, 0.006, 0.2, 0.1, c);
    end
    bs(end+1, :) = uc; %record values
    ts(end+1, :) = t;
    alphas(end+1, :) = ua;
    betas(end+1, :) = b1(2);
end

%plots
%bs = bs(:, 2:end); %remove first value constant
ts = ts(:, 2:end);
boxplot(bs) %symmetric for normality
hold on
plot(ts(1:end), 'dg');
xlabel('School');
ylabel('Estimated log(c)');
legend('True log(c)', 'Location', 'northeast');
hold off
ax = gcf;
exportgraphics(ax, x + "nsbox.jpg")
bs = bs';
ts = ts';
scatter(bs(:),ts(:)) % Plot true c's v estimated c's.
    (property of ols that not many outliers at the top)
hline = reffline(1,0); %45 degree line for means
hline.Color = 'r';
xlabel('Estimated log(c)');
ylabel('True log(c)');
ax = gcf;
exportgraphics(ax, x + "nsscatter.jpg")

%summary statistics for table (only for ns)
bm = nanmean(bs,2); %mean
tm = nanmean(ts,2);
bstd = nanstd(bs, [], 2); %stdv
z_vector = (bm - tm) ./ bstd; %z-score

```

```

p_one = normcdf(z_vector); %convert to p-value
mbtable(:, end+1) = bm; %record values for table
mttable(:, end+1) = tm;
stdbtable(:, end+1) = bstd;
alphan = nanmean(alphas);
betam = nanmean(betas);

%make table
table(:, end+1) = tm;
table(:, end+1) = bm;
table(:, end+1) = bstd;
table(:, end+1) = p_one';
table(end+1, 1) = alphan;
table(end+1, 2) = betam;
writematrix(table, x + "nstable.xls")
end

function [alpha, R2, beta1, beta, true, un_alpha, un_c] =
    name(N, S, Sigmatheta, Sigmah, Sigmas, Sigmac, C)

%Convert the parameters to ensure sigmatheta^2+sigmas^2+sigmac^2=1
    (normalize std) (1.3);
sigmatheta = Sigmatheta/sqrt(Sigmatheta^2+Sigmas^2+Sigmac^2);
sigmah = Sigmah/sqrt(Sigmatheta^2+Sigmas^2+Sigmac^2);
sigmas = Sigmas/sqrt(Sigmatheta^2+Sigmas^2+Sigmac^2);
sigmac = Sigmac/sqrt(Sigmatheta^2+Sigmas^2+Sigmac^2);

%Simulate the talent and shock variables by creating a vector from a normal distribu
T = normrnd(0,sigmatheta,N,1); % Talent distribtion
Eh = normrnd(0,sigmah,N,1); % High school effect
Es = normrnd(0,sigmas,N,1); % College effect
%C = normrnd(0,sigmac,S,1); % Coach effect

% To generate the school data.
Q = S*rand(N,1); %student quality= vector of random numbers
    that is player size scaled to equate to schools

```

```

chi = sort([S*rand(S-1,1);S]);
%schools= matrix of random numbers that is schools-1 size scaled to
schools by # of schools and add max/total to the bottom
%allow for unequal number of players in each school with a random 1st
threshold and upper bound for if statement
J = ones(N,1); %vector of ones as large as players to assign school

for i=1:N
    for j=2:S %for every player go through every school
        if Q(i)<chi(j) && Q(i)>chi(j-1) %if they are less than the current
            school rank but greater than the previous
                J(i) = j;           % Student i goes to school j
            end
        end
    end
end

D=zeros(N,S); % Creating schools dummies
for i=1:N
    D(i,J(i))=1; %putting students in schools,
    Naan because some schools get no players
end

%Set up parameters of normal distribution
w=1/N*sum(D); % Weight of each school in regressions. (% of players at school)
meanC = w*C; %dot product: w1*C1+w2*C2+..., weighted average coach effects
scaled by # of players
varc = (C-meanC*ones(S,1))'*diag(w)*(C-meanC*ones(S,1)); %vector of
variance from mean

% Generate perceived quality and rank data (probabilities)
V = T+Eh; % High school perceived quality= talent+other effects
W=zeros(N,1); % College perceived quality
for i=1:N
    W(i) = T(i)+Es(i)+C(J(i)); %talent+other effects+coach effects
end

```

```

[sortedV , rankV] = sort(V,'descend'); %Rank(i) gives the index in V
    that corresponds to ith location. gives the id number of the player
[sortedW , rankW] = sort(W,'descend');

nh = zeros(N,1);
ns = zeros(N,1);
for i=1:N
    nh(rankV(i))=i; % This gives the rank of each player i for every id number
    ns(rankW(i))=i;
end

etah = zeros(N,1);
etas = zeros(N,1);
for i=1:N
    etah(i) = norminv(1-(nh(i)-0.5)/N); % inverse CDF to get values;
        assumes selection at middle of interval, could modify to be
        random in range
    etas(i) = norminv(1-(ns(i)-0.5)/N);
end

%Define y=Xbeta + M
X = [etah D]; %high school effects for each player and what school they went to
M = eye(N)-X*(X'*X)^(-1)*X';
beta = (X'*X)^(-1)*X'*etas;

alpha = sqrt((sigmatheta^2+sigmah^2)/(sigmatheta^2+sigmas^2+varc));
    %True alpha from model
true = [alpha;C];

[beta true [0,w]'] %Comparison of estimates against the true parameters

% Regress estimated c's on true c's + constant
X1 = [ones(S,1), C];
beta1 = (X1'*X1)^(-1)*X1'*beta(2:S+1)
    % First number is constant, second is coefficient on true c.
R2 = 1- (beta(2:S+1)'*[eye(S)-X1*(X1'*X1)^(-1)*X1']*beta(2:S+1))/

```

```

(beta(2:S+1)'*[eye(S)-1/S*ones(S,1)*ones(1,S)]*beta(2:S+1))

un_alpha = alpha;
un_c = beta1(2:end);
if ~isnan(beta1)
    % Attenuation bias correction
    sigmahat2 = 1/(N-S-1)*etas'*M*etas; %RSS with var correction
    c_hat = beta(2:end);
    alpha_hat = beta1(1,1);
    etah_bar = (D'*D)^(-1)*D'*etah;
    covar = cov(c_hat, etah_bar);

    %solve equation
    syms lambda
    eqn = 2*lambda^3 + lambda^2*(4*alpha_hat+ var(etah_bar)) +
        2*lambda*(alpha_hat^2 - var(etah_bar)) - sigmahat2 ==0;
    sol = solve(eqn,lambda, 'MaxDegree', 3);
    sol = sol(sol==real(sol)); %remove imaginary solutions
    atten_factor = max(double(sol));

    %plug in and compute
    un_alpha = alpha_hat + atten_factor; %5
    un_c = c_hat - atten_factor*etah_bar;
    var_e = var(sigmahat2) - atten_factor^2*var(etah_bar); %6
    var_c = var(c_hat) - 2*atten_factor*covar(2,1) +
        atten_factor^2*var(etah_bar); %7
    var_h = var_e/2; %1.4.1
    var_s = var_h; %1.4.1
    var_theta = 1 - var_s - var(c_hat); %1
end
end

```

C Results Code

This section contains the code used to run the data in the model as detailed in the Results section. The Matlab code assigns each player a Z score and then is exported. A simple OLS

regression is run in the Stata code below and then is exported. Finally, the Matlab code contains a correction for attenuation bias for the final results.

Matlab:

```
clear;

%import data
data = readtable("Player Data.xlsx");

%Define parameters
nh = table2array(data(:, {'RK'})); %high school recruit rank
ns = table2array(data(:, {'Combined_2__Pick'})); %college draft rank

data.player_id = (1:height(data)).'; %create id number for each player

N = length(table2array(data(:, {'Player'}))); %Number of player observations
NH = 100; %Total players in hs
pro = readtable("Pro-Football Draft.xlsx");
Y = unique(table2array(pro(:, ('DraftYear'))));
NS = []; %Total players in college
for j=1:length(Y) %for every draft year
    counter = 0;
    for i=1:length(table2array(pro(:, ('DraftYear')))) %for all players drafted
        if pro.DraftYear(i) == Y(j)
            %if Draft year is equal to the item in the list add one
            counter = counter+1;
        end
    end
    NS(end+1, :) = counter; %record values
end

S_unique = unique(table2array(data(:, {'Combined_2__College_Univ'})));
%list of unique schools
S = length(S_unique); % Number of schools

for i=1:N
```

```

    data.school_id(i) = find(strcmp(cell2mat(data.Combined_2__College_Univ(i)),
        S_unique), 1); %assign each school an id number
end

%Assign z-score to each player (get rid of naan)
etah = zeros(N,1);
etas = zeros(N,1);
for i=1:N
    etah(i) = norminv(1-(nh(i)-0.5)/NH); % inverse CDF to get values; assumes
        selection at middle of interval
    for j=1:length(Y) %for every draft year
        if data.Combined_2__DraftYear(i) == Y(j) %find the index for the draft year
            etas(i) = norminv(1-(ns(i)-0.5)/NS(j));
        end
    end
end
end

%save file
data.etah = etah;
data.etas = etas;
writetable(data, "EtaData.xlsx")

%regress in stata

% Attenuation bias correction
sigmahat2 = 1.0336;
coef = readtable("StataCoef.txt");
c_hat = table2array(coef(2:end-3,2));
var_chat = table2array(coef(2:end-3,3));
alpha_hat = .1325349;
etah_bar = [];
for i=2:S %remove first school as reference for fixed effects
    l = [];
    for j=1:N
        if data.school_id(j)==i
            l(end+1, :) = etah(j);
        end
    end
end

```

```

        end
    end
    etah_bar(end+1, :) = mean(1);
end
covar = cov(c_hat, etah_bar);

%solve equation
syms lambda
eqn = 2*lambda^3 + lambda^2*(4*alpha_hat+ var(etah_bar)) +
    2*lambda*(alpha_hat^2 - var(etah_bar)) - sigmahat2 ==0;
sol = solve(eqn,lambda, 'MaxDegree', 3);
sol = sol(sol==real(sol)); %remove imaginary solutions
atten_factor = max(double(sol));

%plug in and compute
un_alpha = alpha_hat + atten_factor; %5
un_c = c_hat - atten_factor*etah_bar;
var_e = var(sigmahat2) - atten_factor^2*var(etah_bar); %6
var_c = var_chat.*(var_e/sigmahat2);
var_c = abs(var_c); %positive
var_h = var_e/2; %1.4.1
var_s = var_h; %1.4.1
var_theta = 1 - var_s - var(c_hat); %1

```

Stata:

```

clear
set more off
cd "C:\Users\allys\OneDrive\Senior Year\Thesis"

*import data with eta generated in matlab*
import excel EtaData, firstrow

*descriptive statistics *
sum RK Combined_2__Pick

```

```

histogram RK, percent
graph export HRK-dist.png, replace

histogram Combined_2__Pick, percent
graph export CRK-dist.png, replace

scatter RK Combined_2__Pick
graph export Rank-dist.png, replace

regress RK Combined_2__Pick, robust

scatter etah etas || lfit etah etas, ///
yttitle("{&eta}h") ///
xttitle("{&eta}s") ///
title("{&eta}h vs {&eta}s")
graph export Eta-dist.png, replace
corr etah etas

tab Combined_2__College_Univ
graph hbar (count), ///
over(Combined_2__College_Univ, label(labsize(tiny))) ///
yttitle("Number of Players") ///
title("Player Observations per School") ///
blabel(bar, si(vsmall))
graph export School-dist.png, replace

by Combined_2__College_Univ, sort: gen freq = _N
tab Combined_2__College_Univ if freq ==1

tab Player

*regress*
regress etas etah, robust

xtset school_id
xtline Combined_2__Pick Combined_2__DraftYear

```

```
graph export Draft-panel.png, replace

xi: regress etas etah i.school_id
matrix list e(b)
outreg2 using StataCoef.doc, side noparen replace
outreg2 using StataCoef.txt, side noparen noaster replace
```