Mixing Matching and Sabermetrics: Combining Advanced Analytics and the Generalized Matching Law in NFL Football Play-Calling

Jacob Bradley
Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Experimental Analysis of Behavior Commons, and the Industrial and Organizational Psychology Commons

Recommended Citation

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact wmu-scholarworks@wmich.edu.
MIXING MATCHING AND SABERMETRICS: COMBINING ADVANCED
ANALYTICS AND THE GENERALIZED MATCHING LAW
IN NFL FOOTBALL PLAY-CALLING

by

Jacob Bradley

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Psychology
Western Michigan University
April 2018

Dissertation Committee:

Douglas Johnson, Ph.D., Chair
Heather McGee, Ph.D.
Thomas Critchfield, Ph.D.
Derek Reed, Ph.D.
The fields of advanced analytics in sports and quantitative analysis of behavior as it applies to sports have developed independently over the last several decades. Both fields share the common goal of using a quantitative approach to describe and predict behavior within sports beyond the common traditional verbal accounts. To date, the two fields have not directly intersected. The current study provides an overview of advanced analytics and quantitative analysis of behavior in sports, demonstrates how the two fields can be combined to better account for the behavioral processes involved in decision-making in sports, and identifies several possible ways the two fields can be combined in future research. The Generalized Matching Equation (GME) from quantitative analysis of behavior has been successfully used to account for NFL play-calling behavior in previous studies when yards-gained on a play was the measure of reinforcement. The current study compares GME models using yards-gained as the measure of reinforcement with GME models using Success Rate—a seminal advanced analytics metric in football assessing the value of a play—as the measure of reinforcement. The GME models using Success Rate as the measure of reinforcement were largely found to account for more variance in play-calling behavior and provide better generalized matching outcomes, thus
demonstrating the potential value of combining advanced analytics in sports with the quantitative analysis of behavior to improve the description and prediction of choice behavior in elite sports.
ACKNOWLEDGMENTS

I would like to begin by thanking my advisor, Dr. Douglas Johnson for his guidance throughout this project. His continued direction and support were critical to completing this study. I would also like to thank Dr. Heather McGee, Dr. Thomas Critchfield, and Dr. Derek Reed. Their expertise, insight and suggestions, especially in the areas of behavior analysis, quantitative analysis of behavior, comparison of quantitative models, and applications of quantitative models of choice to sport behavior, made it possible for me to pursue and complete this project.

Most importantly, I would like to thank my wife, Jamie, for her support throughout the research and writing process and her encouragement to pursue my goals in advancing my education. Lastly, I would like to thank Western Michigan University’s Graduate College for their consideration.

Jacob L. Bradley
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................... ii  
LIST OF TABLES ................................................................................................................... viii  
LIST OF FIGURES ............................................................................................................... ix  
INTRODUCTION .................................................................................................................. 1  
  Advanced Analytics in Elite Sports .......................................................... 1  
  Advanced Analytics in Football .............................................................. 8  
    Success Rate .................................................................................. 8  
    Expected Points Added .............................................................. 10  
    Win Probability Added ............................................................. 11  
  Quantitative Analysis of Behavior in Sports .............................................. 12  
    Matching Law ........................................................................ 13  
    Applications to Basketball .......................................................... 15  
    Applications to Hockey ................................................................. 21  
    Applications to Baseball ................................................................. 22  
    Applications to Football ................................................................. 24  
  Limitations and Goals ......................................................................... 37  
  Preview of the Proposed Research .............................................................. 42  
GENERAL METHODS...................................................................................................... 43  
  Data Collection and Classification ....................................................... 43  
  GME Analyses ............................................................................... 44
Table of Contents—Continued

ANALYSIS 1: SEASON-AGGREGATE LEAGUE OUTCOMES ................. 45
   Analysis 1A: Yards-Gained as Reinforcement ......................... 46
      Analysis ................................................................. 46
      Results ............................................................... 46
   Analysis 1B: Success Rate as Reinforcement ......................... 47
      Analysis ................................................................. 47
      Results ............................................................... 47
   Analysis 1 Comparison ..................................................... 48
      Analysis ................................................................. 48
      Results ............................................................... 49
   Analysis 1 Discussion ....................................................... 50

ANALYSIS 2: GAME-BY-GAME LEAGUE OUTCOMES..................... 56
   Analysis 2A: Yards-Gained as Reinforcement ......................... 57
      Analysis ................................................................. 57
      Results ............................................................... 57
   Analysis 2B: Success Rate as Reinforcement ......................... 57
      Analysis ................................................................. 57
      Results ............................................................... 57
   Analysis 2 Comparison ..................................................... 58
      Analysis ................................................................. 58
      Results ............................................................... 59
Table of Contents—Continued

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis 2 Discussion</td>
<td>60</td>
</tr>
<tr>
<td>ANALYSIS 3: GAME-BY-GAME INDIVIDUAL TEAM OUTCOMES</td>
<td>62</td>
</tr>
<tr>
<td>Analysis 3A: Yards-Gained as Reinforcement</td>
<td>64</td>
</tr>
<tr>
<td>Analysis</td>
<td>64</td>
</tr>
<tr>
<td>Results</td>
<td>64</td>
</tr>
<tr>
<td>Analysis 3B: Success Rate as Reinforcement</td>
<td>64</td>
</tr>
<tr>
<td>Analysis</td>
<td>64</td>
</tr>
<tr>
<td>Results</td>
<td>66</td>
</tr>
<tr>
<td>Analysis 3 Comparison</td>
<td>66</td>
</tr>
<tr>
<td>Analysis</td>
<td>66</td>
</tr>
<tr>
<td>Results</td>
<td>67</td>
</tr>
<tr>
<td>Analysis 3 Discussion</td>
<td>70</td>
</tr>
<tr>
<td>ANALYSIS 4: CORRELATIONAL ANALYSIS OF MATCHING WITH WINNING PERCENTAGE</td>
<td>73</td>
</tr>
<tr>
<td>Analysis 4A: Yards-Gained as Reinforcement</td>
<td>73</td>
</tr>
<tr>
<td>Analysis</td>
<td>73</td>
</tr>
<tr>
<td>Results</td>
<td>73</td>
</tr>
<tr>
<td>Analysis 4B: Success Rate as Reinforcement</td>
<td>74</td>
</tr>
<tr>
<td>Analysis</td>
<td>74</td>
</tr>
<tr>
<td>Results</td>
<td>74</td>
</tr>
<tr>
<td>Analysis 4 Discussion</td>
<td>75</td>
</tr>
</tbody>
</table>
Table of Contents—Continued

GENERAL DISCUSSION .......................................................................................... 76

Impact .................................................................................................................. 76

Combining Quantitative Analysis of Behavior and Sabermetrics .................................. 76

Success Rate as a Measure of Reinforcement ......................................................... 77

Potential Applications .......................................................................................... 79

Benefits of Translational Research in Quantitative Analysis of Behavior ......................... 79

Limitations ............................................................................................................. 80

Future Research .................................................................................................... 83

Longitudinal Studies ............................................................................................... 84

Other Advanced Analytic Measures ......................................................................... 84

Non-Binary Response Options in Play-Calling ......................................................... 85

Simulation ............................................................................................................... 86

Other Choice Behaviors in Sports ............................................................................ 87

Conclusion ............................................................................................................... 88

REFERENCES ......................................................................................................... 90

APPENDICES

A. Plot of Residuals for Analysis 1A and Analysis 1B ................................................ 98
B. Plot of Residuals for Analysis 2A and Analysis 2B ................................................. 100
C. GME Graphs for Analysis 3A and Analysis 3B Arranged by Team ....................... 102
D. Plots of Residuals for Analysis 3A and Analysis 3B Arranged by Team ............. 114
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. Plot of $R^2$ Values and Means for Analysis 3A and Analysis 3B</td>
<td>126</td>
</tr>
<tr>
<td>F. R2 Values by Team</td>
<td>128</td>
</tr>
<tr>
<td>G. Required Protocol Clearance Letter</td>
<td>130</td>
</tr>
</tbody>
</table>
LIST OF TABLES

1. Summary of Analysis 3 Results: Matching Parameters of Comparisons for Each NFL Team in 2016 Regular Season ............................................................... 65

2. Analysis 4 Results: Correlation of Matching Parameters with Winning Percentage .................................................................................................................. 75
LIST OF FIGURES

1. Possible GME Regression Lines ................................................................. 14
2. GME Graphs for Analysis 1A and Analysis 1B ........................................ 47
3. GME Graphs for Analysis 2A and Analysis 2B ........................................ 58
INTRODUCTION

The fields of advanced analytics in sports and quantitative analysis of behavior have been developing independently over the course of several decades. Advanced analytics in sports has had as its focus the evaluation of player performance and game strategy through innovative metrics and novel applications of statistics. Generally, it involves the application of statistical analyses to performance data in sports in order to predict future performance (Berri & Bradbury, 2010). The field of quantitative analysis of behavior is concerned with understanding, predicting and controlling behavioral phenomena through behavioral principles and quantitative models (Nevin, 2008).

Advances are generally made in the laboratory research setting, but some models have been successfully applied to better understand behavioral processes in less controlled settings, including in elite sport competition (Borrerro & Vollmer, 2002; Poling, Weeden, Redner, Foster, 2011; Reed, Critchfield, Martens, 2006; Vollmer, 2000). These two fields have differed in both their courses of development and approaches, but share the common goal of using equations to describe and predict behavior within sports in ways that are more objective than traditional verbal accounts (Reed, 2011). To date, the two fields have not directly intersected. The current study will provide an overview of advanced analytics and quantitative analysis of behavior in sports and will attempt to demonstrate one initial way the two fields can be combined to better account for the behavioral processes involved in decision-making in elite sports competition.

Advanced Analytics in Elite Sports

The use of advanced analytics in professional sports can be traced back to one of its most notable pioneers, Earnshaw Cook. Cook was a Princeton University trained
metallurgist who went on to consult on the Manhattan Project before becoming an engineering professor at Johns Hopkins University. He was known as an opinionated intellectual and avid baseball fan, and he was one of the first to study baseball using advanced statistics (Deford, 1972). For decades a debate had raged in baseball fandom as to whether Ty Cobb or Babe Ruth was the better batter. Fans were divided over the debate and each side was entrenched in their view. Cook, a staunch Ty Cobb supporter, was no different. Cobb was known for hitting a high batting average and using speed to round and steal bases. Ruth, nicknamed the *Colossus of Clout*, was known as a big homerun hitter who used his batting power instead of speed to score (Shwarz, 2004). It was clear that a homerun, Ruth’s specialty, was more valuable than a base hit, Cobb’s specialty. What was not clear was how much more valuable a homerun was over a base hit. Base hits occur much more frequently than home runs, but the value they provide is the opportunity to advance around the bases to score, while a homerun hit by itself is a score. There had been no way of directly comparing a batter with a higher batting average to a batter with a higher home run percentage.

Cook decided to use statistical analysis to compare these baseball stats\(^1\) and solve many other baseball conundrums, such as when a relief pitcher should be used and how to arrange the most effective batting order. One advanced metric that Cook created, called the *Scoring Index*, was built by comparing the relative probabilities of every offensive act—such as hitting a single or a double, stealing a base, hitting a home run, etc.—and the ways those acts can be arranged to score runs in one inning of a baseball

---

\(^1\) *Stats* in sports refers to the standard way regular actions and outcomes within a game are documented and reported within a given sport. Stats are typically recorded as a count, rate, or percentage, though other forms are also used. This term is not to be confused with statistical analysis of data.
The final equation for the Scoring Index stated that the probability of scoring a run was proportional to the probability of a batter getting to first base multiplied by the probability of advancing that player to subsequent bases. Cook’s Scoring Index provided a metric to directly compare the offensive performances of stylistically different players like Ty Cobb and Babe Ruth. Cobb had a Scoring Index of .2161, while Ruth was close behind with a scoring index of .2109. Thus, Cook was able to demonstrate that, according to his Scoring Index, Cobb was indeed the better player (Schwarz, 2004). Many of his findings went starkly against common baseball wisdom. Cook’s work culminated in a 1964 book called Percentage Baseball. Cook was interviewed about his work for an article in Sports Illustrated magazine, which focused on Cook’s contrarian findings and the statistical evidence he used to back up his claims. The piece was so provocative that it was the lead feature story in the March 23, 1964, issue and was titled “BASEBALL IS PLAYED ALL WRONG.” In the article the author, Frank Deford, made the bold claim that, “Right now Earnshaw Cook knows more about baseball than anyone else in the world” (Deford, 1964, p.17).

Cook led an iconoclastic entrance of advanced analytics into professional baseball in the mid-1960s that decades later would go on to change the sport’s standards for strategy. Though Cook was one of the first to receive national media attention for his work on the statistical analysis of baseball, his approach was not well received by baseball experts or adopted by Major League Baseball (MLB) teams at the time. In addition to his contrarian approach and the difficulty in fully comprehending statistics without advanced training, Cook’s original analyses were laboriously done by hand with a calculator and slide rule which limited dissemination. The advent of digital computers
in the 1980s—which made it possible for many people to initiate calculations in volumes never dreamed of before—allowed for the exponential expansion of the use of statistical analyses in baseball (Schwarz, 2004).

Bill James, a U.S. Army veteran and aspiring sports writer, entered the baseball advanced analytics scene in the mid-1970s. James, like Cook, believed that baseball was played and understood wrong. James began publishing an annual book on advanced analytics in baseball called *Baseball Abstract* in 1977. These books presented various categories of statistical information that were not available anywhere else. The first several annual editions sold few copies, but by the mid-1980s James had garnered media attention for his work and his books were relatively widely distributed and read (Schwarz, 2004). James coined the term *sabermetrics*—named after the Society for American Baseball Research (SABR), of which James was a member—as the name of the field of advanced analytics in baseball (Birnbaum, n.d.).

Bill James became known as the father of sabermetrics and one of the first *sabermetricians*—a portmanteau of *sabermetrics* and *statistician*—and was the main figure who brought advanced analytics to popular American sports. James’ work was so transformative to the game of baseball and elite sports strategy in general that he was named one of *TIME* magazine’s *100 Most Influential People* in 2006 (Henry, 2006). James invented numerous analytical innovations in baseball, such as *Win Shares* and *Runs Created* (Braunstein, 2010; James & Henzler, 2002).

*Runs Created* is a metric that allows for an individual player’s contribution to a team’s number of runs scored to be measured, as well as an expected number of runs scored for a team to be calculated. Though the formula has been through numerous
rounds of revisions by many sabermetricians and has grown somewhat complex, the original formula created to calculate Runs Created was simply hits plus walks, multiplied by total bases, and dividing that product by plate appearances. The original Runs Created metric, when applied to a whole team, has a strong correlation with the actual number of runs scored for that team. This correlation shows that there is some measure of validity to the Runs Created metric (James, 2001). James’ Runs Created metric is one example of how simple and useful advanced analytics can be in elite sport competitions.

The value of the Runs Created metric, along with many other sabermetrics statistics, lies in two main areas: game strategy and player selection. It is obvious that in professional sports, including baseball, teams want to have the best players. Personnel decisions have historically been made with a strong emphasis on traditional baseball stats, and not sabermetrics. The best players were typically identified using traditional stats such as batting average and home run percentage. Players with great traditional stats are typically quite expensive for teams to sign because their skills are in high demand. For example, in 2017 the top starting pitcher for the Los Angeles Dodgers, Clayton Kershaw, was paid an annual salary of $35,571,428 per year, while the lowest paid relief pitcher on the team, Brock Stewart, was paid $537,500 per year (Los Angeles Dodgers 2017 salaries payroll, 2017). Though both are pitchers for the same team, Kershaw is paid more than 66 times as much as Stewart.

Teams want to sign players that contribute runs and wins to their team. Historically, traditional baseball stats were used as a way to assess a player’s value to a team. However, advanced analytics have shown that even if a player’s traditional baseball stats are not very impressive, a player could significantly contribute runs and wins to a
team. For example, a batter could have a high Runs Created metric while having an unimpressive home run percentage. The Runs Created metric can show that a player is significantly contributing to creating runs for the team—which is important to winning games—without having impressive traditional baseball stats. Given that traditional baseball stats have historically influenced the demand for and therefore the salary of players in the MLB, a player with higher sabermetric stats but unimpressive traditional stats could be a cost-effective, valuable player for a team to sign. This sabermetric-based approach to running an MLB team was first implemented on a club-wide scale by the Oakland Athletics in the late 1990’s. The story of how this underdog team from Oakland became a seemingly overnight success is told in the bestselling book, *Moneyball: The Art of Winning an Unfair Game* (Lewis, 2003). The success of the Oakland Athletics, driven by the extensive use of sabermetrics in player selection and game strategy, has led to most MLB teams hiring full-time sabermetricians and even creating entire staff departments dedicated to using sabermetrics to gain a competitive advantage (Baumer, 2015).

Following baseball’s sabermetrics trend, National Basketball Association (NBA) teams also began to adopt the use of advanced analytics. According to a 2015 report by *ESPN* titled, “The Great Analytics Rankings,” a majority of NBA teams use sabermetrics, with some teams using sabermetrics as the foundation around which their organization is built (e.g., Dallas Mavericks). Sabermetrics is also gaining ground in the National Football League (NFL) with roughly a third of teams employing specialized personnel to conduct advanced analytics and another third of teams just beginning to use advanced analytics with existing internal personnel and external consulting firms (Baumer, 2015).
There are differences in the adoption of sabermetrics across these three major sports for at least three reasons. First, the use of advanced analytics in professional sports started in baseball and the sport has therefore had the longest period of time to develop, refine and implement this approach to fit their sport. Second, raw baseball data is highly compatible with advanced analytics. Baseball is considered to be an atomistic sport in that player performance is highly separable from team outcomes and the performances of other players (Gerrard, 2007). The way the game of baseball is structured makes it is relatively easy to collect and analyze data—especially offensive data—based on the performance of individual players. When a batter is at the plate, the raw offensive data collected involve only one player—the batter. When the pitcher pitches the ball to the batter, there are only two players directly involved—the pitcher and the batter. Compared to invasion team sports like basketball and football, baseball is a relatively individual, atomistic sport. Gerrard provides a concise explanation of the differences between atomistic sports, like baseball, and team invasion sports, especially as it relates to the analysis of individual player performance:

…analysis in invasion team sports such as the various codes of football, field and ice hockey, and basketball is much more problematic. Invasion team sports are much more complex and hence the separability of individual player contributions is considerably more difficult. Invasion team sports involve a group of players cooperating to move an object (e.g., a ball or puck) to a particular location defended by opponents (e.g., across a line or between goalposts). (Gerrard, 2007, p. 222)

Gerrard’s (2007) explanation of atomistic and invasion sports can also be summarized in a simple metaphor, “Baseball is a thread; football is a fabric” (Carroll,
Invasion team sports are by their nature more complex in the following ways: (1) individual players engage in a wide breadth of actions such as tackling, passing, blocking, running, catching, shooting, etc.; (2) players act independently in their own actions, cooperatively within their squad, and in response to the actions of opposing offensive and defensive players; (3) some actions involve the joint effort of two players, such as when defensive players work together to tackle an offensive ball carrier in football; (4) in some invasion team sports like hockey and soccer play flows continuously with seamless and almost instant transitions from offense to defense; in others, like football, there is segmentation, meaning that play stops and squads of players are changed when possession changes; (5) different invasion team sports require various levels of player specialization, where a segmented sport like football has a high degree of player specialization and a continuous flow sport like soccer requires nearly all players to be able to play both offense and defense at a competent level (Gerrard, 2007).

The third reason there have been differences in the adoption of advanced analytics across these three major sports is that each MLB team plays 162 games each regular season, while NBA teams play 82 and NFL teams play only 16. There are more than 10 times as many regular season baseball games as football games. Baseball’s volume of games in a season produces a larger data set than is possible for football, which also makes the sport more conducive to advanced analytics (Quinn, 2012).

**Advanced Analytics in Football**

**Success Rate.** Despite the difficulties inherent to football play-by-play data, advanced analytics has made several inroads into the sport. One of the first published
works explaining advanced analytics in football was Carroll, Palmer & Thorn’s *The Hidden Game of Football* (1988). The book introduced a procedure to calculate an advanced metric which provided a new way to evaluate the effectiveness of a football play beyond the standard measure of yards-gained. Each play was labeled as either a *win* or a *fail*, based on whether the play was successful or unsuccessful at gaining a certain percentage of the yards needed for a first down\(^2\). The result is a measure of the rate of successful plays. The *Hidden Game of Football* described the process used to calculate this new metric, but never provided a specific name for it (see Carroll et al., 1988, p. 69). More recent publications detailing Carroll et al.’s procedures have called the metric *Success Rate* (see Berri & Burke, 2012, p. 147). Carroll et al. used a combination of three variables to determine the success of a given play: 1) yards-gained, which is the stand-alone traditional measure of play success; 2) down; and 3) distance needed to earn a first down. On first down, a play is to be considered a success if it results in a gain of at least 40% of the yards needed for a first down. For example, if the down-and-distance\(^3\) is 1\(^{st}\)-and-10, a play would be successful if it resulted in a gain of four or more yards for the offense. On second down, a play is successful if it gains 60% of the remaining yards to earn a first down. For example, on 2\(^{nd}\)-and-5, a play would be successful if it gained at least three yards. On third down, a play is successful only if it gains 100% or more of the

---

\(^2\) In football, *down* is the term used for a play. The offense has four downs, or plays, to gain 10 yards. If the offense gains 10 yards within four downs, then the offense retains possession of the ball and earns a new set of four downs starting again on first down. If the offense fails to gain 10 yards within four downs, then the ball changes possession and the team that was on defense becomes the new team on offense. The downs are numbered *first down*, *second down*, *third down*, etc., to identify their position within the series of four downs.

\(^3\) *Down-and-distance* states the play number in the sequence of four downs and the distance remaining to earn a new set of downs. For example, on 2\(^{nd}\)-and-8 it is second down with 8 yards remaining for the offense to earn a new first down.
yards remaining and results in a new set of downs. For example, if it is 3rd-and-2, then a play would be a success only if it gained 2 or more yards. Once each play in a game or a season is rated using the above rules, Success Rate can be calculated on the team or individual level by dividing the number of successful plays by the sum of successful and unsuccessful plays (Carroll et al. 1988; Berri & Burke, 2012).

**Expected Points Added.** Another advanced analytics method for determining the value of a football play is the expected points (EP) a given football situation adds to the offensive team. Expected Points Added (EPA) takes a different approach than Success Rate to determine the value of a play in that it measures and compares the value of situations before and after a play, instead of measuring the advancement of the ball during the play and qualifying the yards-gained based on their relationship to the current down-and-distance (Carroll et al. 1988) The first step to calculating EPA is to define a set of possible football situations, which is generally done by listing each possible combination of down-and-distance at each possible yard-line on the field. This creates literally thousands of possible combinations which each represent a unique football situation. Once the situations are defined, the past five to ten years of NFL seasons are reviewed to determine what the next points scored were for each of the possible football situations (Berri & Burke, 2012). Take, for example, the football situation of interest being 2nd-and-4, ball on the opponent’s 35 yard line. The expected points for that football situation would be determined by reviewing five to ten years of data to find each instance of 2nd-and-4, ball on the opponent’s 35 yard line, then determining the next points scored after each instance of that football situation, and then averaging all of those next points scored. The next points scored in any given instance of a defined football situation may
be by the offensive team (positive values) or by the team on defense (negative values). The averaged value of the next points scored from all of the instances of a defined football situation is an EP value. Every combination of down-and-distance and field position will have its own EP value.

EPA is calculated by subtracting the EP value from the football situation at the beginning of a play by the EP value from the football situation at the end of a play (Berri & Burke, 2012). For example, at the start of the play it is 2\textsuperscript{nd}-and-4 with the ball on the opponent’s 35 yard line, which has a hypothetical EP value of 3.1. A play is then run for a gain of four yards, which results in 1\textsuperscript{st}-and-10 ball on the opponent’s 31 yard line. This new football situation has a hypothetical EP value of 4.0. When 3.1 EP before the play is subtracted from 4.0 EP after the play, the result is the net gain of .9 EPA. This means that the play is expected to add .9 points to the score of the offensive team in the current game. If the play had only resulted in the gain of 1 yard, and the EP value of 3\textsuperscript{rd}-and-3 ball on the opponent’s 34 yard line was 2.0, then even though the offense gained a yard on the play, they would have an EPA value of -1.1 for that play. This basic model for calculating EPA can be modified by adding another variable to further specify football situations, such as time remaining in the game or a measure of the strength of the opposing defense.

**Win Probability Added.** Win Probability (WP) has a similar structure to EP in that a recent history of NFL seasons are analyzed by defined football situations to produce a value (Berri & Burke, 2012; Carroll et al., 1988). However, the variable of interest in WP is whether or not the offensive team in a specific football situation ended up winning the game, instead of recording the next points scored. A standard WP model will include
down-and-distance, field position, time remaining in the game and score, though other variables can be added (Lock & Nettleton, 2014). Once a WP model is constructed, WPA is calculated in the same manner as EPA by subtracting the pre-play WP from the post-play WP (Berri & Burke, 2012). Though at first glance it may seem that WP models are relatively easy to construct, the reality is that they are rather difficult to construct properly. Some EP models can essentially be classification functions that identify and classify instances of given football situations and average the next points scored across those instances, though there are multiple regression-based EP models (Berri & Burke, 2012; Carroll et al., 1988). Given the sheer volume of permutations of football situations created with the variables in a standard WP model, a classification function approach is not feasible for constructing WP models. Rather, WP models are examples of complex sabermetric analyses that are built on multiple regression analysis and may involve complex machine learning (Berri & Burke, 2012; Lock & Nettleton, 2014).

Quantitative Analysis of Behavior in Sports

Beyond the use of advanced analytic techniques to compare various sporting outcomes, there has also been a growing utilization of quantitative techniques for analyzing choice behavior more directly. Quantitative analysis of behavior involves the application of mathematical models to describe and predict functional relationships between environment and behavior. There are numerous research areas subsumed under the umbrella of quantitative analysis of behavior, one of which are the analyses of choice behavior. Choice behavior is analyzed with several different quantitative models, some of which are discounting, behavioral momentum, behavioral economics, and the matching law. Of these analyses, only the matching law has been directly applied to sports, though
it is reasonable to think that other analyses could be applied to sports with some success (Reed, 2011).

**Matching Law.** The matching law was developed in laboratory settings but has been applied to more socially significant areas ranging from complex problem behavior in clinical settings to MLB team-fan social media interactions (Borrero & Vollmer, 2002; Reed, 2016). Herrnstein (1961) introduced the matching law which states that organisms tend to allocate their responding between two simultaneously available response options proportionately to the reinforcement available for each response option. Herrnstein’s original matching law can be stated mathematically as follows:

\[
\frac{B_1}{B_1 + B_2} = \frac{R_1}{R_1 + R_2} \quad (1)
\]

In this equation, \(B_1\) and \(B_2\) represent two simultaneously available behaviors, while \(R_1\) and \(R_2\) represent the simultaneously available schedules of reinforcement corresponding to \(B_1\) and \(B_2\) respectively. The equation states that the proportion of the occurrence of one behavior \(B_1\) to the total amount of behavior that occurred \((B_1 + B_2)\) is equal to the proportion of reinforcement \(R_1\) received for that behavior to the total amount of reinforcement that was received \((R_1 + R_2)\). Herrnstein’s matching law equation has undergone expansions and revisions over the years. Baum (1974, 1979) introduced the generalized matching equation (GME) which both allows and accounts for deviation from strict matching. The GME can be stated as follows:

\[
\log \left( \frac{B_1}{B_2} \right) = a \log \left( \frac{R_1}{R_2} \right) + \log b \quad (2)
\]

Similar to Equation (1), \(B_1\) and \(B_2\) represent two behaviors while \(R_1\) and \(R_2\) represent the reinforcement corresponding to \(B_1\) and \(B_2\), respectively. The \(a\) parameter represents the slope of the function and is a measure of sensitivity to differentials in
reinforcement, while \( \log b \) corresponds to the y-intercept and is a measure of preference for one response option over the other—called bias—that is not accounted for by reinforcement ratios. The logarithmic transformation of the response and reinforcement ratios produces a linear function—as opposed to a curvilinear function—which is more conducive to evaluation (Reed & Kaplan, 2011).

Figure 1. Possible GME Regression Lines.

Strict matching is demonstrated by regression line (A) in Figure 1 with a slope \((a)\) of 1.0 and bias \((\log b)\) of 0.0. Slopes greater than 1.0 (B) represent overmatching, which occurs when the behavior ratio changes relatively more than the reinforcement ratio and is an uncommon outcome of a GME analysis. Slopes between 0.0 and 1.0 (C) demonstrate undermatching, which occurs when the behavior changes relatively less than the reinforcement ratio and is a common outcome of a GME analysis. Bias toward \(B_1\) (D) and bias toward \(B_2\) (E) are identified by having y-intercepts \((\log b)\) in the GME analysis) greater than or less than 0.0, respectively.

Another variation of the matching equation is the Concatenated Matching Equation (CME), which allows for the adjustment of the equation to factor in different amounts of reinforcement available for \(R_1\) and \(R_2\) (Davison & Hogsden, 1984). The CME
can be stated mathematically as:

\[
\frac{B_1}{B_1 + B_2} = \frac{R_1(A)}{R_1(A) + R_2}
\]  

(3)

where the equation is similar to Equation (1), with \(A\) functioning as a multiplier that represents the amount by which the reinforcer corresponding to one behavior is multiplied to represent the difference in the amount of reinforcement from the other behavior.

**Applications to Basketball.** Vollmer & Bourret (2000) used the CME stated in Equation (3) to analyze shot selection by college basketball players on two different teams. The two behaviors in the Vollmer & Bourret study were three-point and two-point shots attempted. The measure of reinforcement were the baskets made from successful three-point and two-point shots, with \(A\) set at 1.5 to account for the reinforcing difference between three-point and two-point shots. The CME, in addition to the plotting of log response ratios against log reinforcement ratios—which essentially converted the CME to the GME—was able to account for 90% or more of the variance between the selection of two-point and three-point shots across players for the two teams that were analyzed.

Romanowich, Bourret & Vollmer (2007) conducted a partial replication and extension of Vollmer & Bourret (2000) using 57 NBA players that met certain inclusion criteria over 9 NBA seasons. In 1994, the NBA adjusted the distance required for a three-point shot from an arc measuring 23 ft. 9 in. from the top of the arc and 22 ft. from the baseline, to an arc measuring 22 ft. both at the top of the arc and from the baseline. The three-point line was then returned to its original distance of 23 ft. 9 in. from the top of the arc and 22 ft. from the baseline in 1997. This change allowed for a quasi-experimental analysis to be conducted. The study used three 3-year periods for analysis which naturally
formed an A-B-A reversal design: 1991–1994, 1994–1997, 1997–2000. Both A conditions had the three-point arc set at 23 ft. 9 in. from the top of the arc and 22 ft. from the baseline. The B condition had the three point arc set at 22 ft. from both the top of the arc and the baseline. Romanowich et al. hypothesized that the change from A to B would have resulted in an increase in the reinforcer probability and reinforcement rate for three-point shots, given the shorter distance from the three-point arc to the hoop. Likewise, the authors hypothesized that the change from the B condition back to the A condition would have resulted in a decrease in the reinforcement probability and reinforcement rate, given the return to the longer distance from the three-point arc to the hoop. The study found that matching was closely approximated in all three conditions. A small increase in bias for three-point shots occurred when conditions changed from A to B. However, the bias toward three-point shots was maintained when conditions changed from B back to A. Undermatching was found across all three conditions. A statistically significant difference was found between the shot completion percentages in the first A condition and the B condition, but not between the B condition and the return to A condition.

This line of research was expanded by another 2007 study in which the GME was applied to the 11 teams that were in college basketball’s Big Ten Conference at the time (Hitt, Alferink, Critchfield, & Wagman, 2007). The results from the Hitt et al. (2007) study, in conjunction with Romanowich et al. (2007) and Vollmer & Bourret (2000) demonstrate that the GME is capable of accounting for a majority of the variance in shot allocation in elite basketball players and teams. Additionally, there was a rather consistent bias for choosing three-point shots and slight undermatching was found. These three studies provide a proof-of-concept and rationale for continued research using the
GME to describe and account for choice behavior in sports. However, the samples used in these studies have not been fully representative of the scope of basketball players and teams in both the NCAA and NBA. Vollmer & Bourret studied two collegiate teams from the same university, Romanowich et al. studied a total of 57 NBA players, and Hitt et al. studied 11 NCAA teams that played in the same conference.

Alferink, Critchfield, Hitt, & Higgins (2009) addressed this limitation in the research when they conducted a study of shot selection using the GME for all NCAA men’s Division I teams for the 2005-2006 regular season. Data were obtained from 320 of the 332 teams and players from those teams were included in the analysis if they attempted a minimum of 14 two-point and 14 three-point shots during the season. A total of 80% of the variance in shot selection for 295 of the 320 teams was accounted for by the GME. For 21 of the remaining 25 teams, the GME accounted for more than 50% of the shot selection variance. As predicted, shot selection varied with shots made and three-point shots were taken more than expected. 92.6% of teams exhibited undermatching, while 94.6% of teams demonstrated a bias for three-point shots. The GME often accounts for >95% of variance in tightly controlled laboratory studies (Baum, 1979). However, the matching law’s ability to account for such a substantial amount of variance in elite basketball shot-selection is impressive given the translational nature of the studies and lack of experimental control in these post hoc applications of the GME.

Alferink et al. (2009) conducted a second study aimed at assessing the explanatory flexibility of the matching law by answering three questions: 1) Did the GME show differences in shot selection between players on successful and unsuccessful teams? 2) Did the GME show differences between players on teams that participate in
NCAA Division I, II, and III? 3) Did the GME show differences between regular players and substitute players? The unit of analysis used in this second study—performance of individual players—differs from many previous analyses which have used team performance as the unit of analysis. Narrowing the unit of analysis from the team to the individual player reduces the amount of variability that is inherent with teams being built of players who perform at different levels. Similar inclusion criteria were used in this second study as were used in the first study. An individual player’s season aggregate data made one data point in the GME analysis.

In answering the first research question in their second study regarding differences between players on successful and unsuccessful teams, Alferink et al. (2009) used Rating Percentage Index (RPI) rankings to determine the top 30 and bottom 30 teams out of the total 330 teams in NCAA Division I basketball for both the 2004-2005 and 2005-2006 seasons. During the 2004-2005 season, the GME accounted for 94.1% of the variance between two-point and three-point shots for the top 30 teams and 91.1% for the bottom 30 teams. A larger difference was found in the 2005-2006 season, where 94.6% of the variance was accounted for among the top 30 teams, while only 83.6% of the variance was accounted for among the bottom 30 teams. For the top 30 and bottom 30 teams in both seasons, both undermatching and a bias for three-point shots were demonstrated. Top 30 teams in both seasons had more variance accounted for and a greater sensitivity to shot-making contingencies than the bottom 30 teams. This suggests that better teams consist of players whose behavior is more sensitive to the relevant contingencies regarding shot-making.
The second research question regarding differences between players on teams that participate in NCAA Division I, II, and III basketball was analyzed slightly differently than the first research question. In this analysis, data from players on the 25 most highly rated teams in each of the three divisions, as determined by a coaches’ poll, were used. Coaches’ polls were used instead of RPI rankings because RPI rankings were only available for Division I teams and not Division II or III teams. Also, coaches’ polls only rank the top 25 teams and do not provide rankings on the rest of the teams in each division, so 25 teams were analyzed instead of 30. The GME accounted for 94.0% of the variance for Division I players, 93.5% for Division II players, and 81.2% for Division III players for the 2005-2006 season. Similar to the outcomes of the analysis for the first research question, players in all three divisions demonstrated undermatching and a bias for three-point shots. The results of this analysis suggest that there is a relationship between shot-selection matching and competitive success.

The third research question addressed the differences between regular players and substitutes in terms of matching. Regular players were defined as those who played an average of 32 minutes or more per game, out of the 40 total minutes in a game (i.e., three fourths of the game). Substitute players were those who played an average of one third or less of each game they participated in. The GME accounted for 95.5% of the variance of in shot selection for regulars and only 72.6% of the variance for substitutes. As in the analyses for the first two research questions, undermatching and a bias for three-point shots were found for both regulars and substitutes. Regulars also demonstrated greater sensitivity than substitutes. These findings suggest regulars are—as is generally assumed—more skilled than substitutes. Also, regulars likely benefit from more extended
contact with in-game shot-taking contingencies than substitutes. This would be consistent with laboratory studies that have demonstrated that matching outcomes can differ based on the duration of exposure to contingencies (Todorov, de Olivera Castro, Hanna, Bittencourt de Sa, & Barreto, 1983).

Alferink et al. (2009) also conducted a third study in which they narrowed the unit of analysis even further, from aggregated individual data to the performance of specific individuals. The matching law, as it was developed, accounts for the choice of individual organisms (Herrnstein, 1961; Baum, 1974). Previous applications of the matching law in basketball have used aggregated player data or team data. Though these applications have shown the GME to be relevant and useful at the aggregate level, it was originally intended to be used at the individual level. This third study used various criteria to select players from a pool of the 50 greatest players in NBA history at the time as determined by NBA.com. Nine of the 50 greatest players listed met all of researchers’ criteria including Charles Barkley, Larry Bird, Clyde Drexler, Magic Johnson, Michael Jordan, Karl Malone, Scottie Pippin, John Stockton, and Isiah Thomas. Each data point for each player represented one season of play. The GME accounted for at least 89% of the variance in shot selection for each individual player and all players demonstrated undermatching. The variance in shot selection that was accounted for by the GME exceeded 99% for three players: Magic Johnson (99.1%), Scottie Pippin (99.2%) and Clyde Drexler (99.6%). Unlike previous analyses of aggregate player performance data, which repeatedly demonstrated a bias for three-point shots, some players, such as Magic Johnson and Isiah Thomas, exhibited no bias and others, like Michael Jordan and Karl Malone, showed a bias for two-point shots. Though data for only nine players—
considered to be some of the greatest players in NBA history—were analyzed, this study has successfully extended a traditional GME analysis of individual choice behavior from the laboratory setting to an elite sport setting.

These findings demonstrate that choice behavior in elite level competitive sports is able to be analyzed using the GME. These studies are some of the first in a significant line of translational research applying analyses from the experimental analysis of behavior directly to the socially significant arena of elite sports. Though it has been noted that there is a bi-directional research influence between the applied and experimental branches of behavior analysis, this line of translational research follows the more standard pipeline model of taking findings from basic research and applying them to non-laboratory settings (Virues-Ortega, Huraod-Parrado, Cox, & Pear, 2014).

Applications to Hockey. Seniuk, Williams, Reed, & Wright (2015) expanded the line of shot selection research in elite sports using the GME by moving from basketball to hockey. Shooting in hockey obviously differs from shooting in basketball topographically, but also in at least two other major ways: 1) all shots in hockey, if made, are worth one point; 2) hockey shots are taken and made with a much lower frequency than basketball shots. Given the infrequency of shots made in hockey, the authors decided to use shots that hit the net (i.e., the goalie makes a save and a point was not scored) as the measure of reinforcement. Data from players in the National Hockey League (NHL) during the 2011-2012 season were analyzed using a modified version of the GME at the league, team, and individual levels. Four different shooting topographies served as the four response options in the analysis including wrist shots, snap shots, slap shots, and backhand shots. The standard GME equation is designed to analyze choice
behavior when there are two response options (Baum, 1974). The GME equations used in this study were designed to incorporate more than two response options and were similar to those used in the matching law analysis of the Rock/Paper/Scissors game by Kangas et al. (2009).

\[
\log \left( \frac{B_1}{B_2 + B_3 + B_4} \right) = a \log \left( \frac{R_1}{R_2 + R_3 + R_4} \right) + \log b \tag{4}
\]

The calculations were run four times such that each of the four behaviors and corresponding reinforcement metrics served as \( B_1 \) and \( R_1 \) while the other three served as \( B_2, B_3, B_4 \) and \( R_2, R_3, R_4 \). A composite was taken and the equation was fit to the data using a least squares linear regression. The analyses of league, team, and individual data showed the GME to be a useful descriptor of shot selection choice. Where the results of this study deviate from previous findings is that there were no significant relationships found between matching and success. There are numerous factors that could account for this apparent lack of relationship including hits on the net not being a key measure of successful performance in hockey.

**Applications to Baseball.** The field of advanced analytics in elite sports first began in baseball and in 2011 the first matching law analysis study of baseball behavior was published (Poling, Weeden, Redner, & Foster, 2011). Poling et al. set out to investigate switch hitting behavior in professional baseball. Switch-hitters are batters that are able to bat both right-handed and left-handed. Data from three former MLB switch-hitters were analyzed in this study including Mickey Mantle, Pete Rose, and Eddie Murray. All three players selected played professional baseball for at least 15 years. It was noted by the authors that switch-hitters are rare in the MLB. The choice behavior
analyzed was right-handed at bats versus left-handed at bats with three consequences of those at bats being total bases earned (TB), runs batted in (RBI), and home runs (HR).

When analyzing switch-hitting and TB data, all three batters demonstrated undermatching with Mantle and Murray showing a bias for batting left-handed and Rose showing a bias for batting right-handed. However, when confidence intervals were calculated for the intercepts, only Mantle’s data showed a reliable degree of left handed bias related to TB. For the analyses of switch-hitting with RBI data, all three batters demonstrated statistically significant undermatching and a bias for left-handed at bats. When HR is used as the measurement of reinforcement, all three batters demonstrated undermatching with a bias for left-handed at bats.

Poling et al. (2011) suggested that had the data been analyzed within the context of the specific game situations under which it occurred—including things such as the score, inning, and runners on base, etc.—that stronger relations may have been found. Though not specifically addressed by the authors, perhaps sabermetric measures which combine game situation specific information and several measures of batting data to form composite scores would provide a more accurate measure of reinforcement than TB, RBI and HR data alone.

The authors also suggest that switch-hitting is likely a rule governed behavior and not controlled by the consequences measured in the study. There is a behavioral rule statement in baseball—not to be confused with a rule of the game of baseball enforced by an umpire, such as three strikes and the batter is out—that Poling et al. (2011) state as, “bat right-handed against left-handed pitchers and bat left-handed against right-handed pitchers” (p. 287). The rule is largely supported by data which show that batters have
more success when facing a pitcher who is throwing with the opposite hand from which the batter is batting (Clotfelter, 2008). Given that the majority of pitchers in the MLB are right handed and following the rule stated above, it makes sense that Mantle, Murray and Rose would demonstrate a bias for batting left-handed. When the situational data were further analyzed, Mantle and Murray batted right-handed against right-handed pitchers only one time each during their MLB careers. Rose batted left-handed against left-handed pitchers on a total of three occasions. These five at bats were the only instances recorded where the handedness of the batter and pitcher were the same across the careers of the players. Five at bats out of the 33,355 at bats across the three players’ careers, the study shows, constitute merely 0.01% of their total batting opportunities. These data points are relatively miniscule and could not be reasonably used to determine whether batters actually benefited from switch hitting. The authors conclude that it should not come as a surprise that the GME did not describe the data well. In addition to the behavioral rule statement regarding switch-hitting, batting in baseball more closely resembles a discrete-trial procedure instead of a free-operant procedure. According to Hall-Johnson & Polling (1984), response allocation is not as sensitive to reinforcement parameters in discrete-trial procedures. While batting in baseball resembles a discrete trial procedure, shooting in basketball and hockey appear to more closely—albeit not perfectly—resemble free operant procedures. Poling et al.’s rationale for the GME not effectively describing switch-hitting does fit with the established findings regarding shooting in basketball and hockey, which have more similarities to free operant procedures.

Applications to Football. Reed, Critchfield, & Martens (2006) took the GME to the gridiron to analyze play-calling data from the 2004 NFL season. The generalized
matching law was used to compare the relative ratio of rushing plays to passing plays as a function of the relative ratio of yards-gained on rushing plays to yards-gained on passing plays. Though football is clearly a team sport, the decision to call a rushing play or a passing play is typically consistently made by either a team’s offensive coordinator or head coach (McCorduck, 1998). The level of analysis used by Reed, et al. was team performance, which was a proxy for the play-call choice behavior of each team’s individual play-caller. It is important to note at least three possible exceptions to each team having one individual play-caller. First, there may be times during a season when the dedicated play-caller for one team will switch from one person to another. For example, an offensive coordinator may start the season calling plays for the team and then the head coach decides to relieve the offensive coordinator of his play-calling duties and begin calling the plays himself. Second, after the quarterback receives the play-call from the team’s offensive coordinator or head coach, he may decide to call an audible and change the play-call based on what he sees the opposing defense doing before the start of the play. Typically, these various audible calls are rehearsed during practice so the quarterback has a pre-established list of audible options and is taught when to use each option. Third, some plays are designed as run/pass option plays. On these option style plays, the quarterback has the option to hand the ball off to a running back for a rushing play, run the ball himself for another type of rushing play, or pass the ball to a receiver for a passing play. In this case, the offensive coordinator or head coach does not call either a rushing or passing play, but rather a play that is designed to be either rushing or passing based on what the quarterback sees as the best option once the play has started and begins to develop.
In Reed et al. (2006) the plays are categorized as either rushing or passing based on what actually happened during each play. It is not possible to know what the original play-call was for each play during the season given that this information is highly confidential for each team and is not publicly accessible. Additionally, when a quarterback is tackled behind the line of scrimmage\(^4\), called a *sack*, it is considered to be a failed rushing play even though a sack is typically the result of a failed passing play. This general play misclassification was due to the way the data were coded and it was not possible to know if the design of the play and intention of the quarterback was to pass the ball or rush the ball. Finally, when a receiver catches a pass and then proceeds to fumble the ball, this study still used the yards-gained data as recorded. Reed et al. conducted three different GME analyses using the 2004 NFL play-calling data.

In the first study, Reed et al. (2006) treated each team’s season-aggregated stats as one data point, which was analogous to Vollmer & Bourret’s (2000) original application of the matching law to shot selection in basketball. In this analysis, the GME accounted for 75.7% of the variance in play-calling with demonstrated undermatching and a bias for calling rushing plays. The authors then repeated this analysis for each NFL season from 1972 to 2003 and found that the 2004 results fit within the historical range. However, it was noted that the variance accounted for by the GME decreased about 0.4% per year across the 32 years. That is to say, the measure of yards-gained used as reinforcement increasingly became a less reliable predictor of play-calling choice behavior. The analyses also showed that the sensitivity to yardage gained as a reinforcer decreased over

\(^4\) The *line of scrimmage* is an imaginary line that runs across the width of the field as determined by the placement of the ball and is the point from which each play begins. The offense and defense line up on opposite sides of the line of scrimmage and are not allowed to cross it until the play begins by the offense snapping the ball.
time with a reduction in the matching slope of about .005 per year. Also, the bias for rushing was shown to weaken over time at a rate of .004 log units per year. The authors suggest several possible causes of these trends including NFL rule changes that favor passing, the increased complexity of passing plays used by NFL offenses, and an increased rate of coach and player replacement caused by increasing pressure from the fan base and changes in NFL free agency rules.

The GME analysis was then repeated for six other professional football leagues and three NCAA Division I college teams. The variance accounted for across the leagues ranged from 57% to 95% and like the NFL, undermatching was present and most leagues showed a bias for rushing while a few showed either no bias or a bias for passing. Difference in bias between the leagues could likely be accounted for by the different rules in the leagues that may greater incentivize passing or rushing. Another interesting finding was that as the relative risk of a fumble increased the bias shifted from rushing to passing, as fumbles are typically less common on passing plays than rushing plays. This suggests that bias is at least in part influenced by relative turnover risk.

In football, the offense has four downs—or plays—to get to 10 yards and earn another first down. If the offense fails to gain 10 yards on the four downs, then a turnover occurs and the team that was on defense now becomes the team on offense. On first down the offense is thought to have the largest amount of viable play-call options because the play does not need to result in earning a first down since there are two more downs to earn the first down. Teams typically punt the ball on fourth down in order to give the ball to their opponent away from their own end zone. On third down, play-calling options are typically more restricted by the game situations, such as needing to gain many yards to
earn a first down. This 3\textsuperscript{rd}-and-long situation almost necessitates a passing play to have a reasonable chance at earning another first down. Given these situational differences, the authors hypothesized that there would be more sensitivity to relationship between play-call choice and yards-gained as a reinforcer on first down than on third down. The GME was applied to all NFL teams’ 2004 data combined on first, second, and third downs, respectively. The analysis produced three interesting findings. First, there was a greater level of sensitivity to the relationship between play-calling and yards-gained on first down than on second and third downs. Second, a bias for rushing was found on first and second down while a bias for passing was found on third down. Third, the variance accounted for by the GME was almost equal for first and second downs, but low for third down.

The second study in Reed et al. (2006) focused on analyzing variation in play-calling game-by-game, where each game is analogous to an individual experimental session, as opposed to season aggregate data used in the first study. The GME was applied to the 2004 NFL preseason, regular season, and postseason in three separate analyses. NFL teams play four games in the preseason for the purpose of evaluating talent to determine which players each team would like to keep on their final regular season roster. Wins and losses do not matter in preseason games and teams may even call certain plays for the purpose of evaluating talent instead of focusing on trying to win the game. As expected, there was a rather weak matching relationship between play-call choice and yards-gained during the preseason and only 48.4\% of the variance in play-calling was accounted for. The matching relationship was stronger during the regular
season and postseason and the variance accounted for by the GME increased to 62.2% and 81.6%, respectively.

The increase in slope and variance accounted for during the regular season and postseason could be attributed to at least a couple different factors. Given the evaluative focus of the preseason, it was not surprising to see the increase from preseason to the competitive regular season. The increase in variance accounted for from the regular season to the post season could be due to the fact that the cost of losing a game also increases. Losing a game in the regular season impacts a team’s regular season record, which is then used to determine which teams advance to the playoffs at the end of the season; losing a game in the postseason results in the team’s season ending immediately. However, the increase in variance accounted for from the regular season to the postseason could also be due to the fact that only the most successful teams in the regular season advance to the postseason and thus there is no input from unsuccessful teams. The authors also suggested that the change in the matching relationship over the three conditions could be the result of accumulated contingency exposure, which has been demonstrated in laboratory studies (Todorov et al., 1983). However, further analysis found that there were no systematic changes in sensitivity, bias or variance accounted for across successive four-game blocks of the regular season.

The third study shifted from game-by-game analysis for the NFL as a whole to game-by-game analyses for individual teams across the 2004 regular season. This narrowing of focus to the behavior of individual play-callers more closely resembles the level of analysis used in laboratory settings. The GME was found to account for 40% of the variance or more for 28 of the 32 teams. The authors note that these results are
modest when compared to laboratory studies, but that they are more in line with previous studies outside of the laboratory where choice behavior was analyzed in more complex settings (e.g., Martens, Halperin, Rummel, & Kilpatrick, 1990). However, it is worth noting that this is also substantially lower than the variance accounted for by GME analyses of other elite sport behavior, such as Vollmer & Bourett (2000) where 90% of the variance in shot selection was accounted for on two collegiate basketball teams.

In this third study, all teams were found to demonstrate undermatching with a majority of teams showing a bias for rushing plays. The authors found that there was a significant correlation between relative turnover rate and bias in play-calling; that is, as the risk of a fumble increased the bias for rushing decreased or, in some cases, shifted to a passing bias. Additional analyses also found that as variance in rushing yards increased, the bias towards rushing decreased or shifted to a passing bias. The authors discuss the possibility that inter-team variations in matching may be related to important football results, which suggests that the variations could be systematic.

Previous laboratory studies, along with Vollmer and Bourret’s (2000) shot selection study, have shown that choice patterns progressively approach matching law predictions as experience accumulates across sessions or games. Reed et al.’s (2006) analysis did not find a similar pattern in the relationship between professional football play selection and yardage gained. The authors suggest one reason for this could be that NFL play-callers are highly experienced with working in an elite football environment and this experience has made them particularly sensitive to reinforcement variables. Another explanation could be that the NFL play-callers, unlike the subjects in most matching law studies in laboratory settings, are humans and studies have shown that even
experimentally naïve human subjects begin to show steady-state performance in less than an hour of being exposed to the relevant contingencies (Critchfield, Paletz, MacAleese, & Newland, 2003). Also of note, NFL play-callers work with a team of coaches and analysts for several days to study the strategies and tendencies of their upcoming opponents. Specific game plans are then developed and rehearsed throughout the week in preparation for the next game. The game plans may function as behavioral rules that affect play-calling choice patterns beyond contingency shaping. (Martin, 2011; Skinner, 1969).

Given the findings in Reed et al. (2006) and Vollmer & Bourett (2000) before it, Reed et al. suspected that matching was an important component of sport behavior that may have a relationship with team successes. It was found that the top 10 most successful offensive teams during the 2004 NFL season had steeper slopes, larger rushing biases and data that better fit the GME than did the 10 least successful offensive teams. A correlational analysis was run between team winning percentage and each of the fitted parameters of the GME. There were statistically significant correlations between variance accounted for and winning percentage ($r = .409, p = .0225$) and between intercept and winning percentage ($r = -.579, p = .0007$), but not between slope and winning percentage. The lack of significant relationship between slope and winning percentage could be influenced by range restriction on the matching slope variable given that all slope estimates were between 0 and 1, while it is conceptually possible to have negative slopes and slopes greater than 1. The intercept, or bias for rushing, was found to be a stronger predictor of winning percentage than 25 of 30 offensive statistics used by football bettors (Reed et al., 2006).
Stilling & Critchfield (2010) expanded the application of the GME to play-call selection in NFL football to assess variation across game situations. Analyses were conducted at the season aggregate level for the 2006 NFL season and at the play-by-play level for six randomly selected games for each of the 32 NFL teams (192 total games) from the 2006 season. Each play was categorized by the following situational factors: down, yards needed for a first down, time remaining, score and field position. Fourth down plays were excluded from this analysis, so the downs were classified as first, second, or third. Plays were divided into three categories by yards needed for a first down: 1-4 yards, 5-10 yards, and >10 yards. Time remaining was divided into three categories: the combined final 2:00 minutes of the second quarter (the final two minutes before halftime) and the final 2:00 minutes of the fourth quarter (the final two minutes of the game), the remaining 13:00 minutes from the second quarter combined with the remaining 13:00 minutes of the fourth quarter, and the entire first quarter combined with the entire third quarter. These categories allowed for an analysis of time remaining in the half being >15:00 minutes (first and third quarters), 2:01-15:00 minutes (second and fourth quarters, minus the final two minutes of each), and ≤2:00 (the final two minutes of the second and fourth quarters). Score for the offensive team in comparison to the defensive team was classified as winning, tied, or losing. Finally, field position was recorded in three zones: 1-8 yards from the opponent’s goal long (near opponent’s goal line), 83-99 yards from the opponent’s goal line (near one’s own goal line), and 9-82 yards from the goal line (the rest of the field not included in the first two zones).

The GME was fitted to the data which consisted of one data point for each NFL team for the 1999-2008 regular season. Sensitivity measures for each season ranged from
.50 to .73 with bias measures ranging from -.14 to -.06 and variance accounted for ranging from 54.9% to 80.5%. The 2006 season had sensitivity of .57 with a bias of -.10 and 75.2% of variance accounted for. The authors conducted GME analyses for the situational variables described above using pooled play data from six randomly selected games for each team from the 2006 regular season. The GME accounted for >40% of the variance in play-calling choice across all game situations analyzed, and >50% for the majority of analyses. It is notable that the majority of analyses for the specific game situations accounted for a smaller amount of variance in play-calling choice than when all the plays were grouped together. A bias for calling rushing plays was found on both first and second downs while a bias for passing plays was found on third down. A bias for passing was found when more than 10 yards were needed for a first down and a bias for rushing when fewer than 10 yards were needed. A rushing bias was found throughout the game, except for when there was less than 2:00 remaining in a half, in which case there was a bias for passing. No significant bias was found when the offensive team was losing and a bias for rushing was found when the offensive team was winning. Finally, there was a strong bias for rushing when teams were 8 or fewer yards away from the opponent’s goal line and a more modest bias towards rushing plays at other locations on the field (Stilling & Critchfield, 2010).

Further analyses were conducted by Reed, Skoch, Kaplan, & Brozyna (2011) to identify possible explanations for bias in offensive play-calling as a function of the aggregate performance of the opposing defense in college football. Unlike previous studies, Reed et al. did not use within-offense analysis; rather, the study analyzed the offensive data of the opponents of the top 5th percentile and bottom 5th percentile of
defenses in college football. The GME used in this study was identical to the one used by Reed et al. (2006). Analyses were conducted on the play-calling distribution of the offenses playing against the top-ranked pass defenses, top-ranked run defenses, bottom-ranked pass defenses, and bottom-ranked run defenses. $R^2$ values for the four analyses ranged from .353 to .579, which represents a decrease in the amount of variance accounted for by the GME compared to Reed et al. (2006) and Stilling & Critchfield (2010). Similar to previous studies, a bias for rushing plays was found. However, the relative degree of the bias for rushing plays was significantly different between opponents of top-ranked pass defenses and opponents of bottom-ranked pass defenses in that opposing offenses of top-ranked pass defenses had a relative bias towards rushing plays.

Critchfield, Meeks, & Stilling (2014) extended this line of matching research as it applies to football play-calling choice. The authors sought to determine if matching differed across conditionally defined game situations using the same data set as Stilling & Critchfield (2010) described above. First, play-calling choice was analyzed in the context of general down-and-distance situations. This analysis included the categories of 1st/2nd-and-long; 1st/2nd-and-short; 3rd-and-long; and 3rd-and-short. Short was defined as needing 1-4 yards to gain a first down and long was defined as needing 7+ yards to gain a first down. A second analysis of play-calling choice was conducted using game score and time remaining to play. With these variables, the second analysis had the categories of first half of the game and winning; second half of the game and winning; first half of the game and losing, and second half of the game and losing. The study found that the GME generally accounted for an average of 40% of the variance in play-calling. A bias for
calling passing plays was found for third down, while there was a bias for rushing plays on first/second down. Teams were generally biased towards running plays when winning and passing plays when losing, which the authors state as being in line with conventional football wisdom.

Matching analyses have also been conducted on football play-calling behavior in a risk tolerance and gain-loss framework. As previously noted, rushing plays are considered to be generally less risky than passing plays. Critchfield & Stilling (2015) stated that an analysis of risk aversion in play-calling cannot be conducted by simply stating whether rushing plays occur more often than passing plays, but rather the analysis can only be conducted by determining if rushing plays occur more often than expected. The GME was used to establish an objective criterion by which the frequency of rushing plays could be evaluated as occurring more often than expected. The authors hypothesized that teams would have a bias for calling rushing plays when winning and a bias for calling passing plays when losing or tied due to framing, which accounts for the variation in risk tolerance depending on the possibility of gains and losses. Play-calling data was collected from a sample of 6 out of 16 total regular season games each for 25 of 32 NFL teams in the 2006 NFL season. Only data from first down and third down were used in this analysis. Play-calling data were grouped into five scoring categories: winning by 8 or more points, winning by 1-7 points, tied, losing by 1-7 points, and losing by 8 or more points. The six games from each of the 25 teams were counted as one observation, for a total of 25 observations. Consistent with previous research, matching was evident and a bias for rushing plays was present across most game conditions, the only exception being a bias for passing when a team was losing by 8 or more points. Bias varied with
score in such a way that the bias for rushing was stronger when the score was more favorable for the offense, in accordance with gain-loss framing and risk tolerance. The variance in play-calling accounted for by the GME ranged from 46% - 75% across game score conditions.

The matching law has also been used to investigate other choice situations in football. Falligant, Boomhower, & Pence (2016) used the GME to analyze the play-call choice between kicking a point-after-touchdown (PAT) or attempting a two-point conversion after a touchdown is scored. On a PAT, the scoring team attempts to kick the football between the uprights of the goal post from the 2-yard line to earn one point. On a two-point conversion attempt, the scoring team attempts to execute a rushing or passing play from the 2-yard line which will result in scoring two points if the scoring team successfully rushes or passes the ball into the end zone. Should a team attempting a PAT or two-point conversion fail, they earn zero points. Teams chose PAT attempts more frequently than two-point conversion attempts. After deciding to kick a PAT, coaches have another choice of using a conventional-style kicker or a soccer-style kicker to attempt the PAT. The GME was used to analyze the decisions to attempt a PAT or a two-point conversion and, for PAT attempts, the decision to select a conventional-style kicker or soccer-style kicker to kick the ball. Points scored was used as the measure of reinforcement in both analyses. Data were gathered from the NCAA website on PAT and two-point attempts from 1958-2013 seasons and on conventional-style and soccer-style kickers from 1975-2004. In 1991, the width of the goal posts in college football was narrowed from 280 inches to 233 inches, which made PAT attempts more difficult. This rule change allowed for an analysis of selection behavior under two sets of rules.
The GME analyses account for a significant amount of the variance in kicking choices with \( R^2 \) values of .88 for PAT versus two-point plays before the 1991 rule change, to .98 after the rule change, and .95 for soccer-style versus conventional-style kickers. The seasons prior to the goalpost narrowing, there was a significantly greater tendency to attempt PAT plays relative to two-point plays which demonstrated as a bias for PAT plays. This bias virtually disappeared after the narrowing of the goalposts in 1991 from 0.33 before the rule change to 0.04 after the rule change and sensitivity significantly increased from 0.61 to 0.92, respectively.

**Limitations and Goals**

The above matching law analyses have largely demonstrated the utility of the GME in describing and accounting for decision making in NFL football. One of the primary limitations of these analyses, as with many translational studies, is that there has been no experimental manipulation of the relevant reinforcement ratios as is standard in laboratory research on matching. This lack of experimental manipulation means that causal relationships within the GME analyses are not able to be determined and that it is at least possible that matching is artefactual in football play-calling. It may be impossible, or nearly so, to gain full experimental control of the relevant variables in GME analyses of play-calling behavior in a natural football game setting. It would be infeasible to gain full control over the behaviors of the 22 players on the football field whose joint performances determine the outcomes of a given play. Even if such control was achieved, the act of manipulating the performance of players for the purpose of controlling the relative rates of reinforcement for calling passing plays and rushing plays by a coach
would change the fundamental nature of the football game, which would belie—at least in part—the very purpose of the analysis.

A second limitation in this line of research is that the amount of variance accounted for in football matching studies has been relatively low in comparison to laboratory studies and even studies in other sports, such as basketball (Baum, 1979; Vollmer & Bourret, 2000). This suggests that the GME analyses applied to football play-calling may not be properly accounting for a variable that is better accounted for in laboratory research and applications of the GME to shot selection in basketball.

Returning to the seminal Reed, et al. (2006) study, the authors identified limitations in their analysis, which largely apply to the line of research to date, stating that:

The present analysis overlooks much that is regarded as important to football success. For instance, a cornucopia of strategies exist for deciding what type of play is most suitable to call depending on the down, the plays that a team has executed correctly, the type of defensive strategy employed by an opponent, the game situation, the weather conditions, the abilities of key players….strategies may also shift as the magnitude of the reinforcer, or yardage gained, varies across game situations….However, despite ignoring most of these factors, the present analysis revealed a global association between play-calling and relative success of passing and rushing plays. (Reed et al., 2006, p. 293)

The authors list several factors that likely impact football play-calling decisions and an avid football fan may be able to come up with even more factors with only a little time. It is apparent that numerous factors influence play-calling in football, yet to date only yards-gained has been used as a measure of reinforcement in the relevant rushing
versus passing matching analyses. This leads to two limitations in the analyses; (1) other factors outside of yards-gained are not used as measures of reinforcement; (2) the reinforcing value of one yard gained changes throughout the course of the game, even play-by-play, based on mediating factors such as down-and-distance. Reed et al. (2006) addressed the issue of varying magnitude of yards-gained as a reinforcer. Unlike in a laboratory setting where, barring satiation, the reinforcing value of one pellet of food is largely the same over the course of an experimental session, the quality of reinforcement in the form of yards-gained in football can change significantly from play to play based on various situational factors. The quality of one yard-gained as a reinforcer varies throughout the course of the game. For instance, the down-and-distance an offensive play-caller faces may function as a type of conditioned motivating operation (Michael, 2004). Consider an offense in a 3rd-and-10 situation in which gains of 10 yards or more are established as reinforcing and the reinforcing value of gains of less than 10 yards is abolished. The play-caller’s behavior of calling passing plays that are designed to gain at least 10 yards is evoked and the behavior of calling rushing plays designed to reliably gain one yard is abated. Now consider an offense in a 3rd-and-1 situation in which a gain of one yard is established as reinforcing. Needing a gain of one yard to earn a first down, the play-caller’s behavior of calling relatively risky passing plays designed to gain 10 yards is abated, while the behavior of calling rushing plays designed to reliably gain one yard is evoked. According to McGee & Johnson (2015), “motivating operations, including various forms of conditioned motivating operations, are a critical component for an operant account of why rewards may vary in value” (p. 17).
Consider the difference in value of one yard gained in two different down-and-distance situations: one yard gained on 1st-and-10 is of little value, while one yard gained on 3rd-and-1 is highly valuable as it earns a new set of downs. The reinforcing value of these one yard gains is substantially different, but previous analyses either treat these one yard gains the same or try to analyze them separately in the context of first down and third down scenarios. One can imagine numerous situations where a gain of a small amount of yards is more valuable than a gain of a larger amount of yards based on different down-and-distance situations. Another example would be one yard gained on 4th-and-2 is likely of no reinforcing value, but one yard gained on 4th-and-1 is likely highly reinforcing. The variability in the quality of reinforcement of yards-gained based on different down and distance situations alone suggests that yards-gained is a structural measure of reinforcement and not a consistent or complete functional measure of reinforcement for play-calling behavior.

There are numerous other factors, such as time remaining in the game, which may also impact the reinforcing value of a yard gained. Take for example a gain of 10 yards on 1st-and-10: if this takes place in the first quarter of a game then 10 yards will be of one reinforcing quality, however if this same play happened with one minute left in the fourth quarter of the game then it would likely be of a significantly higher reinforcing quality. This difference is due to the fact that during the first quarter, the 10 yards is valuable in two ways: 1) it moves the offensive team 10 yards closer to the goal line, 2) it gives the offensive team a new set of downs (it is first down again, as opposed to second down). During the fourth quarter with one minute left to play, the 10 yards is valuable in the two ways mentioned above and possibly in the additional following ways: 3) if the offensive
team is winning, then the new set of downs with one minute to go gives them the
opportunity to “run out the clock” to win the game (i.e., the defensive team does not have
the time or timeouts remaining to get the ball back), 4) if the offensive team is losing by a
reasonable amount, then the new set of downs with one minute remaining in the game
gives the offensive team continued opportunities on what is likely their last chance to
score and win the game. Not only does down-and-distance impact the value of a yard
gained as a reinforcer, it is clear that time remaining in the game influences the value.
Additionally, other factors likely modify the value of a yard gained as a reinforcer, such
as the historical performance of the offense and defense, the score differential between
offense and defense, the number of timeouts each team has remaining, and the offense’s
current field position. Most of these factors and numerous others that potentially modify
the reinforcing value of one yard gained either have not been incorporated into previous
matching analyses, or have been analyzed in the context of specific segmented game
situations. For example, when the down was factored into the analyses (e.g., Reed et al.
2006; Stilling & Critchfield, 2010) the GME was fitted to the data for each down
independently. These and other analyses of segmented, situation-specific football data do
not accurately reflect the order and flow of play-call choice behavior within the game,
nor are they consistent with the majority of laboratory-based matching research where it
is not common practice to divide response data into disjointed segments for analysis.

The line of research on the matching law and football play-calling detailed above,
through the various analyses of segmented situation-specific data, acknowledges the
potential influence that situational factors have on the value of yards-gained as a measure
of reinforcement. To date, none have attempted to ascribe a modified value to yards-
gained as a measure of reinforcement based on these situational factors, or pursue a new and potentially more accurate measure of reinforcement.

Preview of the Proposed Research

Over the past three decades, the field of advanced analytics in football has been developing multiple approaches to measuring the value of the outcomes of each play in an NFL football game (Carroll et al., 1988; Lock & Nettleton, 2014). The progress and current state of matching analyses of football play-calling indicates a potential need for a more comprehensive functional measure of reinforcement beyond yards-gained alone. The current study used Success Rate (Carroll et al.) as the measure of reinforcement for play-calling behavior in the NFL and compared outcomes of GME analyses using both Success Rate and yards-gained as measures of reinforcement. By using both Success Rate—an advanced analytics measure of play success—and the GME, the current study combined the fields of football sabermetrics and the quantitative analysis of behavior.

Four groups of analyses were conducted in this study to investigate the differences between the use of yards-gained and Success Rate as measures of reinforcement in GME analyses of play-calling choice behavior in the NFL. First, the GME analyses were conducted at the league level with the season-aggregate data for each NFL team serving as one data-point. Then the data were arranged such that the GME analyses were conducted at the league level using each game in an entire NFL regular season as one data-point. Third, GME analyses were conducted and compared within and across teams to assess matching relations on an individual level. Finally, the fitted parameters from the GME analyses of the teams were correlated with winning percentage to determine if any significant relationships existed between matching and winning.
Up until this point, the fields of advanced analytics and the quantitative analysis of behavior in football and general elite sport competition have developed independently. Advanced analytics in sports has developed largely outside of academia and has been driven by professional applications, sports media, and fan interest with books, articles, websites and blogs being the main dissemination hubs for the field (Berri & Bradbury, 2010). On the other hand, the quantitative analysis of behavior, especially the matching law, was developed through laboratory research in academic settings with credentialed researchers publishing their findings in peer-reviewed journals (Baum, 1974, 1979; Herrnstein, 1961 McDowell, 2013; Poling, Edwards, & Weeden, 2011). This study will serve as an initial point of convergence for the two distinct, yet complementary lines of work.

GENERAL METHODS

Data Collection and Classification

NFL play-by-play data were purchased from Arm Chair Analysis (armchairanalysis.com), a premier source of accurate and comprehensive NFL data. Play-by-play data included, among other items, whether a play was a pass or rush, the down, the distance needed for a first down, location on the field, quarter, time remaining in the quarter, and Success Rate. Following the procedures explained in Carroll et al.’s The Hidden Game of Football (1988), plays were classified as successful (1) based on the following criteria: on first down, gaining 40% or more yards needed to earn a first down; on second down, gaining 60% or more of the yards needed to earn a first down; on third down only plays that earn a first down are successful. All plays that did not meet the criteria for success were classified as unsuccessful (0). The classification of plays as
passing or rushing was based on what actually took place on the field, which may or may not reflect the plays as they were called. It is possible that a play was called by the coach to be a rushing play, but was changed by the quarterback to be a passing play—or vice versa—which is called an *audible*. It is also possible that a play was called as a passing play, but the quarterback was not able to throw the ball to a receiver, so the quarterback ran the ball. In this case, the play was recorded as a rushing play even though it was the result of a failed passing play. Previous studies (e.g., Reed et al., 2006) classified sacks—plays on which the quarterback is tackled behind the line of scrimmage before the ball is thrown—as rushing plays instead of passing plays due to the nature of the data available at the time. Sacks have been appropriately classified as failed passing plays in the Arm Chair Analysis data set used in this study.

**GME Analyses**

Similar to Reed et al. (2006), the GME as previously stated in Equation 2 was used to analyze play-calling behavior as a ratio of reinforcement, measured as yards-gained on passing plays to yards-gained on rushing plays, as a predictor of the ratio of passing plays to rushing plays, such that:

$$\log \left( \frac{\text{Plays}_{\text{pass}}}{\text{Plays}_{\text{rush}}} \right) = a \log \left( \frac{\text{Yards}_{\text{pass}}}{\text{Yards}_{\text{rush}}} \right) + \log b$$

(5)

As an expansion of this line of research, a second GME analysis was also conducted on play calling behavior using Success Rate instead of yards-gained as the measure of reinforcement, such that:

$$\log \left( \frac{\text{Plays}_{\text{pass}}}{\text{Plays}_{\text{rush}}} \right) = a \log \left( \frac{\text{Successes}_{\text{pass}}}{\text{Successes}_{\text{rush}}} \right) + \log b$$

(6)

It is thought that play Success Rate, a conceptually basic and seminal sabermetric measure of the outcome of football plays, may provide a more accurate functional
measure of reinforcement beyond simply yards-gained. Additionally, the nature of play Success Rate classification data is more consistent with matching analyses both as they are typically conducted in basic research settings and as they have been applied to other elite sports, including basketball and hockey shot selection (McDowell, 2005; Romanowich et al., 2007; Seniuk et al., 2016; Vollmer & Bourett, 2000). When yards-gained is used as the measure of reinforcement, it naturally produces a high amount of variability in the magnitude of reinforcement that can change from one response to the next. This is not typical of the standard approach used in laboratory research, where reinforcer quality and magnitude are tightly controlled and held constant across responses while reinforcement schedules are varied (McDowell, 2005; Poling et al., 2011).

A nonlinear regression package within GraphPad Prism 7.0 was used to fit straight lines to the data. This allowed for the evaluation of hypotheses using ANCOVA, as described below. Asymmetrical 95% Confidence Bands were be plotted around the best fit lines.

**ANALYSIS 1: SEASON-AGGREGATE LEAGUE OUTCOMES**

Vollmer & Bourret’s (2000) seminal study of choice behavior in elite sports analyzed the behavior of individuals players grouped as members of a team, such that the analysis was focused at the team level but was composed of choice behavior data from individuals. Similarly, Reed et al. (2006) first applied the GME analysis to NFL play-calling by treating each NFL team’s season aggregate data as an individual data point, such that the analysis of choice behavior was at the league level but was composed of choice behavior data from the individual play-callers of each team. This level of analysis is applied throughout Analysis 1 in the present study.
Using the D’Agostino-Pearson omnibus normality test in GraphPad Prism 7.0, it was determined that the data used in Analysis 1 came from Gaussian distributions with none of the $p$-values being significant at $alpha$ level .05, including the distributions of

$$\log\left(\frac{\text{Plays}_{\text{pass}}}{\text{Plays}_{\text{rush}}}\right) (K^2 = 1.804, p = .4058), \log\left(\frac{\text{Yards}_{\text{pass}}}{\text{Yards}_{\text{rush}}}\right) (K^2 = 1.841, p = .3982), \text{and}$$

$$\log\left(\frac{\text{Success}_{\text{pass}}}{\text{Success}_{\text{rush}}}\right) (K^2 = 4.451, p = .1080).$$

**Analysis 1A: Yards-Gained as Reinforcement**

*Analysis.* In Study 1A each of the 32 teams’ 2016 season aggregate play-calling data with yards-gained as a measure of reinforcement were treated as one data point in the GME analysis, as stated in Equation 5 and similar to the methods used in Study 1 in Reed et al. (2006). Data were analyzed for outliers using ROUT analysis in GraphPad Prism 7.0 with $Q$ set at 5.0%. $Q$, a measure of False Discovery Rate, is the probability of the ROUT analysis falsely identifying an outlier and is properly set at the same level as $alpha$ (.05) for significance testing in this study (Motulsky & Brown, 2006). One data point—specifically, Buffalo’s $\log\left(\frac{\text{Yards}_{\text{pass}}}{\text{Yards}_{\text{rush}}}\right)$ ratio—was identified as an outlier at the $Q = 5.0\%$ level and removed, resulting in a dataset of 31 total data points.

*Results.* In Analysis 1A, the GME using yards-gained as the measure of reinforcement accounted for 57.9% of the variance in play calling ($R^2 = .579$). This is roughly in line with the $R^2$ values reported in previous play-calling matching law studies, but is markedly lower than the $R^2$ value of .757 reported in Study 1 of Reed et al. (2006). Root mean square error, or $RMSE$, was .0419. Additionally, the slope of the matching relation was positive ($a = .548$). Undermatching was present with a slope less than 1.0. A
bias for calling rushing plays was present as evidenced by a negative $y$-intercept ($\log b = -.023$).

**Analysis 1B: Success Rate as Reinforcement**

*Analysis*. Analysis 1B was structured in the same way as Analysis 1A, except yards-gained were replaced with Success Rate as the measure of reinforcement in the GME, as stated in Equation 6. Data were analyzed for outliers using ROUT analysis with $Q$ set at 5.0%. No outliers were detected and the full 32 team data set was analyzed.

*Results*. In Analysis 1B, where Success Rate was used as the measure of reinforcement, the GME accounted for 86% of the variance in play calling ($R^2 = .86$). $RMSE = .026$. Additionally, the slope of the matching relation was positive and undermatching was present ($a = .699$). A bias for calling passing plays was also present with a positive $y$-intercept ($\log b = .064$). Results of the GME in Analysis 1A and Analysis 1B are graphically depicted in Figure 2. A plot of residual values of the GME analyses from Analysis 1A and Analysis 1B were arranged for visual analysis (see Appendix A). No systematic departures from homoscedasticity were identified.

*Figure 2. GME Graphs for Analysis 1A and Analysis 1B*
Analysis 1 Comparison

**Analysis.** The outcomes of the matching analyses conducted in Analysis 1A and Analysis 1B were compared to evaluate yards-gained and Success Rate as measures of reinforcement. It is important to compare the $R^2$ and $RMSE$ values to determine which model has the better goodness-of-fit. If the GME model from Analysis 1B, which used Success Rate as the measure of reinforcement, was a better fit of the data than Analysis 1A, which used yards-gained as a measure of reinforcement, then Analysis 1B would have the higher $R^2$ value and lower $RMSE$ value. Also, if Success Rate was a more accurate measure of reinforcement than yards-gained, then it may be expected that slope $a$—the measure of sensitivity to reinforcement—would be closer to 1.0 when Success Rate is used than when yards-gained is used. Due to response data and GME structure being the same in both Analysis 1A and Analysis 1B, the study that produces a slope closer to 1.0 may provide a more functionally accurate measure of reinforcement. In order to determine if there were significant differences between the slopes produced in Analysis 1A and Analysis 1B, an ANCOVA analysis similar to Reed et al. (2011) was conducted. The ANCOVA procedure in GraphPad Prism 7.0 was used to compare the slopes and intercepts from the best fit lines in Analysis 1A and Analysis 1B (see Motulsky & Christopoulos, 2004).

The slopes were compared to determine if the parameters were different between the two data sets. It was predicted that the ANCOVA analysis would find that the slopes between the data sets in Study 1A and Study 1B were significantly different at an *alpha* level of .05. If no statistically significant difference between the slopes were found, then an analysis of the y-intercepts ($\log b$ in the GME)—the measure of bias—would be
conducted. Statistical comparison of the biases is only possible if the slopes in Analysis 1A and Analysis 1B were found to be indistinguishable. Finally, the best fit slopes were analyzed to determine if each one differed from 1.0. Likewise, the biases were analyzed to determine if they differed from 0.0.

**Results.** The GME in Analysis 1A accounted for 57.9% of the variance in play-calling, while 86% of variance in play-calling was accounted for in Analysis 2 when Success Rate was used as the measure of reinforcement. When $RMSE$ values were compared, Analysis 1B with $RMSE = .026$ was lower than Analysis 1A with $RMSE = .042$.

When the slopes from Analysis 1A ($a = .5477$) and Analysis 1B ($a = .6986$) were compared, no significant differences were detected at an *alpha* level of .05 ($F = 2.458 [1, 60], p = .1222$). If the slopes from the two analyses were identical, there is a 12.22% chance that data points with slopes this different would be randomly selected. The absence of a significant difference between the slopes allowed for the further comparison of intercepts of the two GME models.

The intercepts between Analysis 1A ($\log b = -.023$) and Analysis 1B ($\log b = .064$) were determined to be significantly different from each other at an *alpha* level of .05 ($F = 83.92 [1, 60], p < .001$). If the intercepts from the two models were identical, then there is less than a .1% chance of randomly selecting data points with intercepts this different.

The slopes of the two GME models were then each compared to a slope of 1.0, which represents perfect matching. The slope produced in the GME from Analysis 1A using yards-gained as the measure of reinforcement was determined to be significantly
different from a slope of 1.0 at the *alpha*.05 level (*F* = 34.86 [1, 30], *p* < .001). The slope produced in the GME from Analysis 1B using Success Rate as the measure of reinforcement was also determined to be significantly different from a slope of 1.0 at the *a* = .05 level (*F* = 34.41 [1, 30], *p* < .001).

The bias terms in the GME analyses as represented by the *y*-intercepts were subsequently compared to a *y*-intercept of 0.0, which represents the absence of bias for calling either passing plays or rushing plays. The *y*-intercept from Analysis 1A in which yards-gained were used as the measure of reinforcement was determined to not be significantly different from an intercept of 0.0 at an *alpha* level of .05 (*F* = .306 [1, 30], *p* = .584). While the *y*-intercept in Analysis 1B where, Success Rate was the measure of reinforcement, was found to be significantly different from 0.0 (*F* = 45.68 [1, 30], *p* < .001).

**Analysis 1 Discussion**

When comparing regression models, such as two versions of the GME in the current study, it is important to keep in mind that there is no one number that can be taken alone as the final arbiter in selecting the better model (Dallery & Soto, 2013; Motulsky & Christopoulis, 2004). Rather, multiple pieces of information should be considered together in building the case for identifying the better model. In the current study, the GME used in Analysis 1A and Analysis 1B were structurally similar in that they both used \( \log \left( \frac{\text{Plays}_{\text{pass}}}{\text{Plays}_{\text{rush}}} \right) \) as the outcome variable and the only predictor variable was a measure of the value of an offensive football play. In the case of Analysis 1A, the predictor variable was \( \log \left( \frac{\text{Yards}_{\text{pass}}}{\text{Yards}_{\text{rush}}} \right) \), while in Analysis 1B, it was \( \log \left( \frac{\text{Successes}_{\text{pass}}}{\text{Successes}_{\text{rush}}} \right) \).
outcome variable \( y \) that can be accounted for by the predictor variable \( x \) and the regression model. The higher the \( R^2 \) value, the more variance accounted for and thus the better the regression model. In Analysis 1A, \( R^2 = .579 \) demonstrating that 57.9\% of the variance in calling passing plays versus rushing plays can be accounted for by the logarithmically transformed ratio of yards gained on passing plays to rushing plays, and the regression model. This value is broadly in line with previous GME analyses of football play calling (Reed et al. 2006, Stilling & Critchfield, 2010). In Analysis 1B accounted for 86\% of the variance (\( R^2 = .86 \)) which means using Success Rate as the measure of reinforcement allowed the GME to account for 28.1\% more variance than using yards-gained as the measure of reinforcement in Analysis 1A (\( R^2 = .579 \)). The amount of variance accounted for by using the sabermetric measure of Success Rate in Analysis 1B was nearly one-and-a-half times as much as that of yards-gained in Analysis 1A. While there are methods for statistically comparing \( R^2 \) values between competing models with a single data set and different numbers of predictor variables, the current analysis compares competing models with only one predictor variable each (Motulsky & Christopoulis, 2004). For this reason the \( R^2 \) values are reported and compared without additional inferential statistics. In doing so, variance accounted for is reported in a similar fashion as previous studies using matching law applications in football play calling choice (Critchfield, Meeks, & Stilling, 2014; Reed et al. 2006;Reed et al. 2011).

It is also helpful to compare the root mean square error, or \( RMSE \) values between the two models. \( RMSE \) is a measure of the distance of residuals from the regression line and is another measure of goodness of fit, in addition to \( R^2 \). Unlike when comparing \( R^2 \) values, the smaller \( RMSE \) value is more preferable (Huitema, 2011). It is also important
to note that $RMSE$ values are reported in the same units as $y$ in the regression model, which is $\log \left( \frac{\text{plays}_{\text{pass}}}{\text{plays}_{\text{rush}}} \right)$ in the current analysis. Analysis 1A had $RMSE = .042$, while in Analysis 1B $RMSE = 0.026$. The $RMSE$ value produced by the GME using Success Rate as the measure of reinforcement was lower than the corresponding value produced by the GME using yards-gained as the measure of reinforcement. This information, in combination with the difference in $R^2$ values, indicates that the GME model using Success Rate as the measure of reinforcement was the model with the better goodness of fit.

Due to the fact that the regression models being compared are the product of the GME, it is important to compare the models based on the slope and intercept values within the context of what is known about matching. Previous analyses have established the general application of the GME to NFL football play-calling behavior using yards-gained as the measure of reinforcement (Critchfield & Stilling, 2010; Reed et al. 2006). It may be the case that better measures of reinforcement would result in slopes and intercepts more closely resembling perfect matching with values approaching $a = 1.0$ and $\log b = 0.0$, respectively.

First, the slopes from Analysis 1A and Analysis 1B were compared to each other. Though the GME analysis using Success Rate as the measure of reinforcement produced a slope ($a = .699$) steeper than the GME analysis using yards-gained as the measure of reinforcement ($a = .548$), the two slopes were not found to be significantly different from each other. It was expected that the slope in Analysis 1B would be both steeper than and significantly different from the slope in Analysis 1A. Such results would be another indication that Success Rate may be a better measure of reinforcement than yards-gained.
While the results were not significantly different, \( p = .122 \) approaches significance and may be limited by the relatively small number of data points. Perhaps future studies across multiple seasons would be able to determine whether or not significant differences in the slopes from Analysis 1A and Analysis 1B appear over time with larger datasets.

Each slope was independently compared to a perfect matching slope of 1.0. Slopes from both Analysis 1A and Analysis 1B were found to be significantly different from 1.0 and demonstrate undermatching. The finding of undermatching is consistent with the previous studies in this line of research (Critchfield & Stilling 2010, Reed et al 2006, Reed et al. 2011). Undermatching demonstrates a lack of perfect sensitivity in play calling choice in the GME analyses. This could be a function of genuine undermatching, however, it could also be influenced by an imperfect measure of reinforcement. Though the slopes produced using yards-gained as the measure of reinforcement and Success Rate as reinforcement were not significantly different, it is possible that another sabermetric measure of the value of a play would produce a slope even steeper than the Success Rate GME analysis that would be significantly different from yards-gained GME analysis.

After the slopes from Analysis 1A and Analysis 1B were found to not significantly differ from each other, a significant difference was found between the intercepts in Analysis 1A (\( \log b = -.023 \)) Analysis 1B (\( \log b = .064 \)). When yards-gained was used as a measure of reinforcement, a slight, though not statistically significant, bias for rushing was present. This finding is consistent with previous studies and historical analyses conducted in previous publications (Reed et al., 2006). A 32 year historical analysis of matching relationships in football play calling conducted by Reed et al.
ranging from 1972 – 2004 determined that negative log \( b \) values, which demonstrate a bias for rushing plays, were slowing growing towards zero at a rate of .004 log units per year. Extrapolating that trend 12 more years to the 2016 season, it would be predicted that the GME analyses in 2016 would produce an intercept of \( \log b = -.081 \) \([b = -.129 \text{ in } 2004 + .048 \text{ log units} = -.081] \) (Reed et al.). This is broadly similar to the intercept (\( \log b = -.023 \)) found in the current study, though the actual 2016 intercept is .058 log units closer to 0.0, suggesting a more aggressive erosion of rushing bias over the past 12 years than predicted. Similar analyses were conducted by Reed et al. for the slope and variance accounted for. The matching slope was found to grow shallower at a rate of -.005 per year, which would predict a slope of \( a = .645 \) in 2016 \([a = .725 \text{ in } 2004 - .08 = .645] \). Similar to the erosion of rushing bias discussed above, the 2016 slope of .548 is broadly similar to the predicted slope, though the actual slope is shallower than predicted. Finally, the variance accounted for was found to be decreasing at a rate of .4% per year in the Reed et al. study, which would predict that in 2016 \( R^2 = .677 \). The actual 2016 \( R^2 = .579 \), which follows the general trend of a reduction in variance accounted for but represents a noticeably lower amount of variance accounted for than predicted by the trend identified by Reed et al. Though all of the GME outcomes in Analysis 1A using 2016 data were in line with the direction of the trends identified by Reed et al., they all exceeded the predicted values in the correct direction. This is an area for future research to explore and perhaps identify when the year-over-year erosion rate of rushing bias, slope, and variance accounted for changed. It is possible that an NFL rule change coincided with a significant change in any one or all three of these parameters. The impact of league rule changes on
matching relations would not be unprecedented, as such findings have already been documented in basketball by Romanowich, Bourrett, and Vollmer (2007).

When Success Rate was used as the measure of reinforcement, a bias for passing was apparent ($\log b = .064$) and was found to be significantly different from 0.0. A football fan may expect a bias for passing in the NFL in 2016, as the league has been described by pundits as being a “pass-heavy league” (Bailey, 2016). The alignment between the outcomes of this GME analysis and common perception of play-calling in the NFL perhaps provides the use of Success Rate as the measure of reinforcement with higher level of face-validity than the yard-gained as the reinforcement metric.

The findings from Analysis 1 make a solid case for preferring the use of Success Rate as a measure of reinforcement over yards-gained in GME analyses of football play-calling behavior. Using Success Rate as the reinforcer allowed the GME to account for substantially more variance in play-calling with lower standard deviation of the residuals than when yards-gained were used. The Success Rate matching slope was closer to 1.0, though not significantly different from the yards-gained matching slope. This piece of information perhaps leans toward supporting the use of Success Rate over yards-gained, but is ultimately inconclusive. Further analyses are needed to determine if there are potential significant differences between matching slopes at different levels of analysis using the GME. A significant bias for passing was evident when Success Rate was the measure of reinforcement, which may add face-validity to the model, given the current emphasis on passing plays in the NFL. However, the slight bias for rushing when yards-gained were used as the measure of reinforcement may suggest that yards-gained is the better measure of reinforcement, given that the intercept value more closely resembles
perfect matching. In summation, a majority of the evidence in Analysis 1 suggests that Success Rate was the better measure of reinforcement.

ANALYSIS 2: GAME-BY-GAME LEAGUE OUTCOMES

Analysis 1 analyzed season-aggregate league outcomes which provided value in understanding the league’s matching outcomes at the season level, but obfuscated the variability that occurs across games. A game-by-game analysis corresponds more closely to an analysis across experimental sessions in laboratory research than does an analysis by team at the season aggregate level. Analysis 1 also failed to find a significant difference between the slopes of the GME analyses using yards-gained and Success Rate as the measures of reinforcement. One possible reason for the lack of significant differences could have been the small number of data points used in the analysis, such that differences may have been present but were not detectable given the small sample size. By analyzing the play calling data across all individual games during the same regular season, instead of aggregating the data by team for the entire season, 512 total data points (32 teams x 16 games) were able to be analyzed, which was 16 times more than Analysis 1. While this analysis was conducted at a different level from Analysis 1, and thus it would not be possible to compare the results to each other directly, the findings would still be able to shed light on the differences between the use of yards-gained and Success Rate as measures of reinforcement. This level of analysis is similar to Study 2 in Reed et al. (2006).

Using the D’Agostino-Pearson omnibus normality test, it was determined that the data came from Gaussian distribution for $\log\left(\frac{\text{Plays}_{pass}}{\text{Plays}_{rush}}\right)$ ($K^2 = 3.932, p = .140$), $\log\left(\frac{\text{Yards}_{pass}}{\text{Yards}_{rush}}\right)$ ($K^2 = 1.711, p = .425$), and $\log\left(\frac{\text{Successes}_{pass}}{\text{Successes}_{rush}}\right)$ ($K^2 = 3.645, p = .162$). Data
were analyzed for outliers using ROUT analysis with \( Q \) set at 5.0%. Four data points—
specifically, Baltimore’s week 7 \( \log \left( \frac{\text{Yards}_{\text{pass}}}{\text{Yards}_{\text{rush}}} \right) \) and \( \log \left( \frac{\text{Success}_{\text{pass}}}{\text{Success}_{\text{rush}}} \right) \) ratios, Denver’s
week 14 \( \log \left( \frac{\text{Success}_{\text{pass}}}{\text{Success}_{\text{rush}}} \right) \) ratio, and San Francisco’s week 13 \( \log \left( \frac{\text{Yards}_{\text{pass}}}{\text{Yards}_{\text{rush}}} \right) \) ratio—
were identified as outliers in their respective distributions and were removed, resulting in
a dataset of 510 total data points for each GME analysis.

**Analysis 2A: Yards-Gained as Reinforcement**

*Analysis.* In Analysis 2A, each 2016 regular season game across the league was
treated as one data point in a GME analysis. Yards-gained was be used as the measure of
reinforcement, as in Equation 5, similar to Reed et al. (2006) Study 2.

*Results.* In Analysis 2A, the GME accounted for 49.5% of the variance in play
calling \( (R^2 = .495) \). \( RMSE \) was .131. Additionally, the slope of the matching relation was
positive \( (a = .489) \). Undermatching was present with a slope less than 1.0. A slight bias
for calling passing plays was present as evidenced by a positive \( y \)-intercept \( (\log b = .001) \).

**Analysis 2B: Success Rate as Reinforcement**

*Analysis.* Analysis 2B was modeled after Analysis 2A with the exception that
Success Rate was replaced with yards-gained as the measure of reinforcement, as in
Equation 6.

*Results.* In Analysis 2A, the GME accounted for 67.5% of the variance in play-
calling \( (R^2 = .675) \), with \( RMSE = .105 \). Additionally, the slope of the matching relation
was positive \( (a = .627) \). Undermatching was present with a slope less than 1.0. A slight
bias for calling passing plays was present as evidenced by a positive \( y \)-intercept \( (\log b = .071) \). Results of the GME analyses in Analysis 2A and Analysis 2B are graphically
depicted in Figure 3. Residuals were plotted for the GME analyses in both Analysis 2A
and Analysis 2B for visual analysis (see Appendix B). No systematic departures from homoscedasticity were readily identified.

**Figure 3.** GME Graphs for Analysis 2A and Analysis 2B

**Analysis 2 Comparison**

**Analysis.** The ANCOVA procedure detailed in Analysis 1 Comparison above was replicated by replacing Analysis 1A and Analysis 1B data with Analysis 2A and Analysis 2B data, respectively. As in Analysis 1 Comparison, it was predicted that there would be a significant difference between the slopes produced in Analysis 2A and Analysis 2B, with the slope in Analysis 2B more closely resembling perfect matching (i.e., $a = 1.0$). This finding would suggest that Success Rate may be a more accurate functional measure of reinforcement than yards-gained. If no statistically significant difference between the slopes was found, then an analysis of the bias terms could be conducted. Statistical comparison of the biases would only be appropriate if no significant difference was found between the slopes in Analysis 2A and Analysis 2B. Additionally, the best fit slopes were then analyzed to determine if each one differed from 1.0. Similarly, the biases were analyzed to determine if they differed from 0.0.
**Results.** When the slopes from Analysis 2A and Analysis 2B were compared, significant differences was detected at the *alpha* level of .05 (*F* = 27.7 [1, 1016], *p* < .001). If the overall slopes were identical, there is less than a .1% chance of randomly choosing data points with slopes this different. Because the slopes were found to be significantly different, it was not possible to test whether the intercepts differed significantly.

The slopes of the two GME models were then each compared to a slope of 1.0, which represents perfect matching. The slope produced in the GME analysis from Analysis 2A using yards-gained as the measure of reinforcement was determined to be significantly different from a slope of 1.0 at the *a* = .05 level (*F* = 544.1 [1, 508], *p* < .001). The slope produced in the GME from Analysis 1B using Success Rate as the measure of reinforcement was also determined to be significantly different from a slope of 1.0 at the *a* = .05 level (*F* = 375.2 [1, 508], *p* < .001).

Though it was not possible to directly compare bias terms in the GME analyses given the significant difference in slopes, the *y*-intercepts were able to be compared independently to a *y*-intercept of 0.0, which represents the absence of bias for calling either passing plays or rushing plays. The *y*-intercept from Analysis 1A in which yards-gained was used as the measure of reinforcement was determined to not be significantly different from an intercept of 0.0 at the *alpha* level of .05 (*F* = .022 [1, 508], *p* = .884). However, the *y*-intercept in Analysis 1B (*log b* = .071) where Success Rate was the measure of reinforcement was found to be significantly different from 0.0 at the *a* = .05 level (*F* = 152.1 [1, 508], *p* < .001). This finding is in line with the results from Analysis 1.
**Analysis 2 Discussion**

As in Analysis 1, the current Analysis 2A and Analysis 2B were structurally similar, sharing the same outcome variable, \( \log \left( \frac{\text{Plays_pass}}{\text{Plays_rush}} \right) \), and having only one predictor variable each, \( \log \left( \frac{\text{Plays_pass}}{\text{Plays_rush}} \right) \) and \( \log \left( \frac{\text{Plays_pass}}{\text{Plays_rush}} \right) \), respectively. Thus, the two models could be adequately compared using \( R^2 \) and a residual analysis, such as \( RMSE \). In analysis 2A, 49.5% of the variance in play-calling was accounted for using yards-gained and the GME model. In Analysis 2B, 67.5% of the variance in play-calling was accounted for, which means that using Success Rate as the measure of reinforcement allowed the GME to account for 18% more variance than using yards-gained as the measure of reinforcement.

In order to get a more complete assessment of the goodness-of-fit for the two models, the \( RMSE \) values were compared. Analysis 2A had \( RMSE = .131 \), while in Analysis 2B \( RMSE = 0.105 \). The \( RMSE \) value from Analysis 2B was lower than the corresponding value in Analysis 2A. This represents only a slight difference in \( RMSE \) values, though favoring Analysis 2B using Success Rate as reinforcement. However, these \( RMSE \) values, taken in combination with the difference in \( R^2 \) values, indicates that the GME model using Success Rate as the measure of reinforcement was the model with the better goodness-of-fit.

A noteworthy observation was that the variance accounted for decreased from Analysis 1A (\( R^2 = .579 \)) to Analysis 2A (\( R^2 = .495 \)) and from Analysis 1B (\( R^2 = .86 \)) to Analysis 2B (\( R^2 = .675 \)). This reduction in variance accounted for was also in line with the findings of Study 1 and Study 2 in Reed et al. (2006). The \( RMSE \) values increased from Analysis 1 to Analysis 2, however, in both analyses the \( RMSE \) values were smaller.
when Success Rate was the measure of reinforcement than when yards-gained was used as reinforcement.

Similar to Study 2 in Reed et al. (2006), a decrease in matching sensitivity, as indicated by the slope, was apparent from the GME analysis at the season-aggregate league level to the game-by-game league level. A significant difference was detected between the slopes produced in Analysis 2A ($a = .489$) and Analysis 2B ($a = .627$), demonstrating a difference in sensitivity to reinforcement based on the measure of reinforcement. Though both GME analyses showed significant undermatching, Analysis 2B had a significantly steeper slope than Analysis 1A, which suggests that play calling is more sensitive to Success Rate as reinforcement than yards-gained and thus Success Rate may be a better measure of reinforcement. These slope findings help provide more clarity in addition to the findings from Analysis 1, where the matching slope with Success Rate was steeper than the matching slope with yards-gained, but not significantly so. The difference between slopes in Analysis 1 appeared to approach significance, but was perhaps limited in its ability to identify a significant difference because of the small number of data points. When the 2016 play calling data were arranged as one data point per game instead of one data point per team, thus producing 16 times as many data points, Analysis 2 found a significant difference between the two matching slopes.

Another point of interest was the shift from slight rushing bias to slight passing bias when yards-gained was used as the measure of reinforcement, though neither the bias in Analysis 1A ($\log b = -.023$) nor Analysis 2A ($\log b = .001$) were significant. A similar shift occurred in Reed et al. 2006, however the bias for rushing only weakened instead of changing to a bias for passing.
Results from Analysis 2 largely support the use of Success rate over yards-gained as the measure of reinforcement in a GME analysis of NFL play-calling. Similar to the results from Analysis 1, the GME analysis using Success Rate accounted for more variance in play-calling behavior and had a slightly lower standard deviation of the residuals than the GME analysis using yards-gained. Both of these results indicate that Analysis 2B had a better goodness-of-fit than Analysis 2A. Significant differences between the matching slopes produced by Success Rate and yards-gained were detected in Analysis 2, with Success Rate producing the steeper slope closer to 1.0. Due to the significant differences found between the two matching slopes, it was not possible to statistically compare matching intercepts to each other. However, when the bias term in each GME analysis was compared to a non-biased \( y \)-intercept of 0.0, it was found that there was no significant difference between \( \log b \) in Analysis 2A and 0.0. A significant difference was detected between \( \log b \) in Analysis 2B and 0.0, with a significant bias for calling passing plays. The findings from comparing the \( y \)-intercept values to 0.0 are consistent with the results in Analysis 1. Given Success Rate’s higher variance accounted for, lower RMSE value, and the slope being significantly steeper and closer to 1.0, it is reasonable to conclude from Analysis 2 that Success Rate is the better measure of reinforcement than yards-gained in the GME analysis of play-calling in NFL football.

**ANALYSIS 3: GAME-BY-GAME INDIVIDUAL TEAM OUTCOMES**

Analyses 1 and Analysis 2 used play-calling data compiled from the league’s 32 teams in the form of season-aggregated data by team and game-by-game aggregated data across the league, respectively. These analyses were helpful in gaining an overall understanding of how matching relations function in the NFL and demonstrated some of
the differences made by using yards-gained and Success Rate as measures of reinforcement. However, matching is a behavioral phenomenon that has its roots in the analysis of individual choice behavior (Baum, 1974; Baum, 1979; Herrnstein, 1961).

Analysis 3 applied the GME to the play-calling behavior of the individual play-callers for each team. There are possible exceptions to the notion that each team had one individual play-caller, such as quarterback audibles or changes in the coach calling the plays as previously discussed, but this information is not publically available and such cases are thought to be the exceptions to the norm of each team having an individual who calls plays for the entire season.

Using the D’Agostino-Pearson omnibus normality test, it was determined that all of the data used in Analysis 3 within each team came from Gaussian distributions for log \( \frac{\text{Plays}_{\text{pass}}}{\text{Plays}_{\text{rush}}} \) ratios (\( K_2 \) ranging from .134 to 4.458, \( p \) ranging from .108 to .935), log \( \frac{\text{Yards}_{\text{pass}}}{\text{Yards}_{\text{rush}}} \) ratios (\( K_2 \) ranging from .052 to 5.8, \( p \) ranging from .055 to .974), and log \( \frac{\text{Successes}_{\text{pass}}}{\text{Successes}_{\text{rush}}} \) ratios (\( K_2 \) ranging from .120 to 5.677, \( p \) ranging from .059 to .942).

Data were analyzed for outliers using ROUT analysis by team with \( Q \) set at 5.0%, the same level as alpha in this study. Eight data points—specifically, Baltimore’s week 7 game log \( \frac{\text{Yards}_{\text{pass}}}{\text{Yards}_{\text{rush}}} \) and log \( \frac{\text{Successes}_{\text{pass}}}{\text{Successes}_{\text{rush}}} \) ratios, Carolina’s week 4 game log \( \frac{\text{Yards}_{\text{pass}}}{\text{Yards}_{\text{rush}}} \) ratio, Chicago’s week 13 game log \( \frac{\text{Plays}_{\text{pass}}}{\text{Plays}_{\text{rush}}} \) ratio, Denver’s week 14 game log \( \frac{\text{Successes}_{\text{pass}}}{\text{Successes}_{\text{rush}}} \) ratio, Detroit’s week 11 game log \( \frac{\text{Yards}_{\text{pass}}}{\text{Yards}_{\text{rush}}} \) ratio, Minnesota’s week 15 game log \( \frac{\text{Successes}_{\text{pass}}}{\text{Successes}_{\text{rush}}} \), and San Francisco’s week 13 game log \( \frac{\text{Yards}_{\text{pass}}}{\text{Yards}_{\text{rush}}} \) ratio.
ratio—were identified as outliers in their respective team distributions and removed, resulting in slightly smaller datasets for those teams.

**Analysis 3A: Yards-Gained as Reinforcement**

*Analysis*. An individual GME analysis was conducted for each of the 32 NFL teams using yards-gained as the measure of reinforcement, as expressed in Equation 5. Each regular season game for each team was treated as one data point in a GME analysis, totaling 16 data points for each team, with the exception of the teams that had outliers removed resulting in 15 data point, as specified above.

*Results*. Across the 32 teams, GME analyses using yards-gained as the measure of reinforcement had slopes ($a$) ranging from .204 to .732, with a mean of .488 and 30 of 32 slopes being significantly different from 1.0 at the *alpha* level of .05. Two teams had slopes that were not significantly different from 1.0. Intercept values (log $b$) ranged from -.113 to .151, with a mean of .0003 and only one team, Cleveland (log $b$ = .151; $p = <.001$), having an intercept significantly different from 0.0. Variance accounted for ($R^2$) ranged from .156 to .810, with a mean of .478, and $RMSE$ ranged from .070 to .164 with a mean of .119. GME outcomes including slopes ($a$), intercepts (log $b$), variance accounted for ($R^2$), and root mean square error ($RMSE$) are reported for each of the 32 teams in the Table 1.

**Analysis 3B: Success Rate as Reinforcement**

*Analysis*. Study 3B was structured similarly to Study 3A, except yards-gained was replaced with Success Rate as the measure of reinforcement in the GME analysis using Equation 6. GME outcomes are also reported for each of the 32 teams in the Table 1 as described above in Study 3A.
### Table 1. Summary of Analysis 3 Results: Matching Parameters and Comparisons for Each NFL Team in 2016 Regular Season

<table>
<thead>
<tr>
<th>Team</th>
<th>Analysis 3A: Yards-Gained as Reinforcement</th>
<th>Analysis 3B: Success Rate as Reinforcement</th>
<th>Analysis 3 Comparison: ANCOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a$</td>
<td>$\log b$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>Arizona</td>
<td>.512</td>
<td>-.054</td>
<td>.610</td>
</tr>
<tr>
<td>Atlanta</td>
<td>.537</td>
<td>-.008</td>
<td>.597</td>
</tr>
<tr>
<td>Baltimore A B</td>
<td>.363</td>
<td>.132</td>
<td>.185</td>
</tr>
<tr>
<td>Buffalo</td>
<td>.310</td>
<td>.010</td>
<td>.519</td>
</tr>
<tr>
<td>Carolina A</td>
<td>.461</td>
<td>-.030</td>
<td>.282</td>
</tr>
<tr>
<td>Chicago A B</td>
<td>.732</td>
<td>-.046</td>
<td>.752</td>
</tr>
<tr>
<td>Cincinnati</td>
<td>.476</td>
<td>-.041</td>
<td>.515</td>
</tr>
<tr>
<td>Cleveland</td>
<td>.345</td>
<td>.151</td>
<td>.665</td>
</tr>
<tr>
<td>Dallas</td>
<td>.522</td>
<td>-.075</td>
<td>.318</td>
</tr>
<tr>
<td>Denver B</td>
<td>.635</td>
<td>.067</td>
<td>.705</td>
</tr>
<tr>
<td>Detroit A</td>
<td>.246</td>
<td>.149</td>
<td>.156</td>
</tr>
<tr>
<td>Green Bay</td>
<td>.499</td>
<td>-.057</td>
<td>.397</td>
</tr>
<tr>
<td>Houston</td>
<td>.449</td>
<td>.038</td>
<td>.466</td>
</tr>
<tr>
<td>Indianapolis</td>
<td>.597</td>
<td>-.033</td>
<td>.246</td>
</tr>
<tr>
<td>Jacksonville</td>
<td>.536</td>
<td>.046</td>
<td>.539</td>
</tr>
<tr>
<td>Kansas City</td>
<td>.458</td>
<td>.007</td>
<td>.239</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>.440</td>
<td>.050</td>
<td>.409</td>
</tr>
<tr>
<td>Miami</td>
<td>.486</td>
<td>.048</td>
<td>.342</td>
</tr>
<tr>
<td>Minnesota B</td>
<td>.510</td>
<td>-.018</td>
<td>.263</td>
</tr>
<tr>
<td>New England</td>
<td>.491</td>
<td>.081</td>
<td>.536</td>
</tr>
<tr>
<td>New Orleans</td>
<td>.664</td>
<td>-.069</td>
<td>.692</td>
</tr>
<tr>
<td>New York (Giants)</td>
<td>.569</td>
<td>-.038</td>
<td>.554</td>
</tr>
<tr>
<td>New York (Jets)</td>
<td>.522</td>
<td>.002</td>
<td>.551</td>
</tr>
<tr>
<td>Oakland</td>
<td>.320</td>
<td>.066</td>
<td>.473</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>.613</td>
<td>-.009</td>
<td>.602</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>.546</td>
<td>.025</td>
<td>.577</td>
</tr>
<tr>
<td>San Diego</td>
<td>.443</td>
<td>-.018</td>
<td>.737</td>
</tr>
<tr>
<td>San Francisco A</td>
<td>.216</td>
<td>.052</td>
<td>.181</td>
</tr>
<tr>
<td>Seattle</td>
<td>.204</td>
<td>.113</td>
<td>.198</td>
</tr>
<tr>
<td>Tampa Bay</td>
<td>.673</td>
<td>.113</td>
<td>.475</td>
</tr>
<tr>
<td>Tennessee</td>
<td>.601</td>
<td>-.071</td>
<td>.654</td>
</tr>
<tr>
<td>Washington</td>
<td>.631</td>
<td>-.066</td>
<td>.810</td>
</tr>
</tbody>
</table>

In Table 1 above, $a$ is the matching slope which measures sensitivity to reinforcement, $\log b$ is the y-intercept which measures bias, $R^2$ measures variance accounted for, and $RMSE$ is root mean square error which measures residuals. A notes teams with 15 instead of 16 games in Analysis 3A due to the removal of outliers. B notes teams with 15 instead of 16 games in Analysis 3B due to the removal of outliers.
**Results.** Across the 32 teams’ GME analyses using yards-gained as the measure of reinforcement, slopes \((a)\) ranged from .394 to .893, with a mean .631 and 23 of 32 slopes being significantly different from 1.0 at the \(\alpha .05\) level. Intercept values \((\log b)\) ranged from .007 to .143, with a mean of .066 and 16 of 32 teams having intercepts significantly different from 0.0. \(R^2\) ranged from .271 to .914, with a mean of .638, and \(RMSE\) ranged from .062 to .150 with a mean of .097. The graphs of the GME analyses conducted in Analysis 3A and Analysis 3B were arranged by team for visual comparison (see Appendix C). Likewise, plots of residuals for the GME analyses were also arranged by team for visual comparison (see Appendix D).

**Analysis 3 Comparison**

**Analysis.** GME outcomes for Study 3A and Study 3B were first compared within each team. The ANCOVA procedure detailed in Study 1 Comparison above was replicated by replacing Study 1A and Study 1B data with Study 3A and Study 3B data, respectively, for each individual NFL team. The \(R^2\) and \(RMSE\) values from the two models were compared for each team to assess the relative goodness-of-fit for the models. As previously discussed, higher \(R^2\) values and lower \(RMSE\) values would be used to identify the GME model that best fits the data. Slopes from Study 3A and 3B for each individual NFL team were compared such that the procedure described in Study 1 Comparison was repeated 32 times—one for each team. As in Study 1 Comparison, it was predicted that there would be a significant difference between the slopes produced in Study 3A and Study 3B for each team, with the slopes in Study 3B more closely resembling perfect matching (i.e., \(a = 1.0\)).
This finding would suggest that Success Rate was a more accurate functional measure of reinforcement than yards-gained. If no statistically significant difference between the slopes were found, then an analysis of the bias terms could be conducted. Statistical comparison of the biases would only be appropriate if no significant difference was found between the slopes in Study 3A and Study 3B. Finally, the best fit slopes were then analyzed to determine if each one differed from 1.0 and the biases were analyzed to determine if they differed from 0.0.

Paired sample $t$-tests across the 32 teams were used to determine if there was a significant difference between the matching slopes in the GME analyses from Analysis 3A and Analysis 3B. It was predicted that there would be statistically significant differences found between the slopes for the GME analyses in Analysis 3A and Analysis 3B. More specifically, if Success Rate was a better functional measure of reinforcement than yards-gained, then the slopes in Analysis 3B would be closer to 1.0 than those found in Analysis 3A. It would not be possible to compare biases between Analysis 3A and Analysis 3B if the slopes were found to be significantly different. If there were no difference found between the two slopes, then comparing biases would be of some interest and a paired samples $t$-test would be conducted on the bias terms in Study 3A and Study 3B across the 32 NFL teams. In order to visually assess the differences in the variance accounted for between Study 3A and Study 3B across the 32 teams, $R^2$ values for 3A and 3B for each team were plotted by model (see Appendix E) and by team (see Appendix F).

**Results.** When the $R^2$ values were averaged across the 32 teams for Analysis 3B the mean was found to be .638, which indicates that on average the GME analysis using
Success Rate as the measure of reinforcement accounted for 16% more variance in play-calling than the GME analysis using yards-gained as reinforcement ($R^2 = .478$). Range minimums and maximums for Analysis 3B ($min = .271; max = .914$) were found to be higher than the corresponding values in Analysis 3A ($min = .156; max = .810$). The mean $RMSE$ value for the 32 teams from Analysis 3B was .097, which was approximately 20% lower than the corresponding mean $RMSE$ value of .119 in Analysis 3A. The range minimums and maximums of the $RMSE$ values in Analysis 3B ($min = .062; max = .150$) were also lower than those in Analysis 3A ($min = .070; max = .164$). A visual inspection of the plotted residuals for each GME analysis found instances of departure from homoscedasticity, such as San Diego’s Analysis 3A and Washington’s Analysis 3B. For the majority of the teams, the plotted residuals appeared either more homoscedastic and more closely grouped around 0 in Analysis 3B, or relatively similar between Analysis 3A and Analysis 3B generally suggesting a better goodness-of-fit for Analysis 3B.

Each of these descriptive comparisons between the ranges of goodness-of-fit values between Analysis 3A and Analysis 3B suggest that Success Rate is a better measure of reinforcement than yards-gained when analyzing NFL play-calling choice using the GME. Although on average both $R^2$ and $RMSE$ values were better when Success Rate was the measure of reinforcement, there were six teams (18.75% of league) for which $R^2$ was higher when yards-gained was the measure of reinforcement. The most extreme example was San Diego, where $R^2 = .773$ in Analysis 3A, which accounted for 16.9% more variance than Analysis 3B, where $R^2 = .604$. Five of those six teams also had $RMSE$ values that were lower when yards-gained was used as the measure of
reinforcement, with San Diego again having the most extreme difference with $RMSE$ where Analysis 3A $RMSE = .07$ was .023 lower than Analysis 3B where $RMSE = .093$.

Analysis 3A slopes ($a$) ranged in value across the 32 teams from .204 to .732, with a mean of .488. In Analysis 3B, the slope values ($a$) across the league ranged from a minimum of .394 to a maximum of .893, with a mean of .631 which were all higher and closer to a perfect matching slope of 1.0 than the corresponding slopes in Analysis 3A.

The intercept ($\log b$) values measuring bias for Analysis 3A were mostly grouped around 0.0, with a range of -.113 to .151 and a mean of .0003. In comparison, the range of intercept values ($\log b$) for Analysis 3B was .007 to .143 with a mean of .017. In Analysis 3A, 18 teams (56.25% of league) had negative intercepts which demonstrated a bias for calling rushing plays, while 14 teams had positive intercepts which showed a bias for passing plays. All 32 teams had positive intercept values in Analysis 3B, which demonstrated a bias for calling passing plays.

Slopes ($a$) from the GME analysis using yards-gained as the measure of reinforcement were compared to slopes from the GME analysis with Success Rate as the measure of reinforcement within each team using ANCOVA and it was found that for all 32 teams, there were no significant differences between slopes with $p$-values ranging from .11 to .974 across the league with none of the values meeting the significance criteria of $\alpha = .05$.

Since no significant differences between slopes were found within any of the 32 teams, the intercepts ($\log b$) of the two GME analyses were subsequently compared within each team. It was found that 18 (56.25%) of the 32 teams had intercepts that were
significantly different from each other, while the other 14 teams had intercepts that were not significantly different. For the 32 teams, $p$-values ranged from $>.001$ to $.81$.

A D’Agostino-Pearson omnibus normality test determined that the matching parameters across the 32 teams in Analysis 3A and Analysis 3B came from Gaussian distributions with none of the $p$-values being significant at alpha level .05, including distributions of slopes ($a$) from the 32 teams in Analysis 3A ($K^2 = 1.611, p = .447$) and Analysis 3B ($K^2 = 4.554, p = .103$), the distribution of intercepts (log $b$) in Analysis 3A ($K^2 = 2.83, p = .243$) and Analysis 3B ($K^2 = 1.378, p = .502$), and the distribution of $R^2$ values in Analysis 3A ($K^2 = 3.717, p = .156$) and Analysis 3B ($K^2 = 2.813, p = .245$).

A paired-samples $t$-test across teams was used to test for differences between the slopes produced for each team in Analysis 3A, where yards-gained was the measure of reinforcement, and Analysis 3B, where Success Rate was the measure of reinforcement. A significant difference was found between the slopes in Analysis 3A and Analysis 3B $t(31) = 6.392, p < .001$, 95% confidence interval (.097, .188). Given the significant differences found between the slopes, no further inferential statistics were conducted to compare the intercepts from Analysis 3A and Analysis 3B. GME analysis outcomes for Analysis 3A and Analysis 3B, along with ANCOVA results from Analysis 3 Comparison are reported in Table 1 with each team representing one row of data.

**Analysis 3 Discussion**

The majority of results from Analysis 3 indicate that Success Rate is a better measure of reinforcement than yards-gained when assessing play-calling choice behavior in a generalized matching paradigm. For 26 of the 32 teams, $R^2$ values were higher when using Success Rate as the measure of reinforcement and for 27 teams $RMSE$ scores were
lower with Success Rate. When $R^2$ values were compared across the league, Analysis 3B accounted for an average of 16% more variance in play-calling than Analysis 3A and the average $RMSE$ was lower by 20%. These goodness-of-fit measures show that the GME model using Success Rate as reinforcement in Analysis 3B fit the data better than the GME model using yards-gained in Analysis 3A.

It was predicted that there would be a significant within-subject difference between the sensitivity of reinforcement measure—slope—in Analysis 3A and Analysis 3B. However, no significant difference was found between slopes within any of the 32 teams. It is interesting to note that while no significant differences were found between the two matching slopes within any one team, the paired-samples $t$-test found a significant difference between the two slopes across the 32 teams, with the average slope in Analysis 3B being steeper and closer to the perfect matching slope of 1.0 than Analysis 3A. This may be due to the fact that teams only had 16 data points—or fewer, for the teams which had outliers removed—for each matching analysis to be compared. When comparisons were made across teams at the league level there were 32 data points.

Several teams’ $p$-values in the ANCOVA analysis approached statistical significance at the .05 level, but were not sufficiently low. Future studies could conduct similarly structured analyses over the course of multiple seasons or include playoff game data to determine if significant differences appear when more data points are used. Additionally, different analyses were used to detect differences between the slopes within each team (ANCOVA) and across the 32 teams (paired-samples $t$-test).

When the intercept values were compared within each team using ANCOVA, just over half of the teams were found to have matching slopes that differed significantly from
each other, with the log $b$ value for each team being larger and more biased towards passing when Success Rate was used as the measure of reinforcement. Analysis 3A produced only one intercept across the 32 teams that was significantly different from 0.0 and Analysis 3B found that 16 teams had intercepts significantly different from 0.0 and all biased towards passing.

There are two possible conclusions to draw from this information on intercept comparisons. First, it might be the case that yards-gained is the better measure of reinforcement, given that 31 of the 32 teams did not have significant biases in their play-calling and thus more closely resembled perfect matching. Second, it may be the case that Success Rate is the more accurate measure of reinforcement and a true bias for calling passing plays did exist within the NFL in the 2016 regular season which was better detected with Success Rate than yards-gained as the measure of reinforcement. The second explanation fits better with the common understanding of the NFL being primarily a passing league. The first explanation also does not seem to fit as well with a majority of the evidence previously discovered in this study (i.e., goodness-of-fit measures and several matching slope comparisons) as does the second explanation, which suggests Success Rate is the better measure of reinforcement. Though the current analysis is not able to determine conclusively which of the two conclusions is more correct—or if a possible third conclusion is correct—it appears that a stronger argument can be made for the use of Success Rate as the better measure of reinforcement.
ANALYSIS 4: CORRELATIONAL ANALYSIS OF MATCHING WITH WINNING PERCENTAGE

Analysis 1 through Analysis 3 in this study focused on comparing the parameters from two GME analyses to each other in various descriptive and inferential ways to determine if yards-gained or Success Rate was the better measure of reinforcement. Analysis 4 sought to answer the same question—which measure of reinforcement is better—by investigating the relationship between the matching parameters from the two analysis and team winning percentage, and comparing those relationships to each other. The relationships were assessed using a Pearson correlational analysis to compare the parameters in Table 1 with the winning percentage for each team. It is thought that if matching is a meaningful element in NFL play-calling behavior, then it may have a relationship with meaningful football outcomes. Winning percentage was selected for comparison as it is almost inarguably the most important football outcome in the regular season.

As shown in Table 2, all data in the following analyses were determined to come from Gaussian distributions with no p-values being significant at alpha level .05 using the D’Agostino-Pearson omnibus normality test. A ROUT analysis was conducted for each fitted parameter and winning percentage and no outliers were detected at Q = 5%.

Analysis 4A: Yards-Gained as Reinforcement

**Analysis.** A Pearson correlational analysis was conducted for each of the slope, bias and $R^2$ data of Study 3A with team winning percentages for the 2016 regular season.

**Results.** As shown in Table 2, no significant correlations were found between winning percentage and any of the fitted parameters from Analysis 3A. While no significant differences were found, the relationship between log $b$ and winning
percentage approached significance \((p = .092)\) with \(r = -.303\), which would have represented that a bias for rushing plays tended to be associated with winning more games. This finding, which was not significant, would have been consistent with the finding in Reed et al. (2006) where a significant negative correlation was found for a similar analysis \((r = -.579, p = .0007)\). The reduction in significance of the correlation between intercept and winning percentage from Reed et al. to the current study may be due in part to the erosion of rushing bias over the past 12 seasons, as previously discussed.

**Analysis 4B: Success Rate as Reinforcement**

*Analysis.* A Pearson correlational analysis was conducted for each of the slope, bias and \(R^2\) data of Study 3B with team winning percentages for the 2016 regular season.

*Results.* Of the three correlational analyses conducted, the only significant relationship found was a negative correlation between \(\log b\) and winning percentage \((r = -.535, p = .002)\). This finding is in line with Reed et al. (2006) in that a significant negative correlation existed between \(\log b\) and winning percentage. There were no significant correlations between any other fitted parameters of the GME analyses and winning percentage.

Given that only one significant correlation was found between any of the matching parameters in Analysis 4A or Analysis 4B and winning percentage, the corresponding correlation coefficients between Analysis 4A and Analysis 4B were not statistically compared to each other.
Analysis 4 Discussion

Six correlational analyses were conducted between winning percentage and the parameters from the GME analyses using yards-gained as the measure of reinforcement and Success Rate as the measure of reinforcement. Of the six correlational analyses, only one relationship was found to be significant and it was a negative correlation between the bias term (log \( b \)), which is the \( y \)-intercept value, and winning percentage. The negative correlation does not emphasize a rushing bias, as it did in the analysis in Reed et al. (2006), given that all 32 teams in the present analysis had matching intercepts greater than 0.0 (passing bias) when Success Rate was used as a measure of reinforcement. Rather, the negative correlation in this analysis shows that teams with intercepts closer to 0.0—demonstrating no bias in play-calling—won more games than teams with more significant passing biases, which highlights the potential importance of unbiased play-calling. The current analyses were not designed to determine causal relations, so the direction or relationship of causality between winning percentage and bias terms in Analysis 4B is not certain.

Table 2. Analysis 4 Results: Correlation of Matching Parameters with Winning Percentage

<table>
<thead>
<tr>
<th>Test</th>
<th>Winning %</th>
<th>Analysis 4A: Yards-Gained as Reinforcement</th>
<th>Analysis 4B: Success Rate as Reinforcement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( a )</td>
<td>( \log b )</td>
</tr>
<tr>
<td>D’Agostino-Pearson Omnibus Normality Test</td>
<td>( K^2 )</td>
<td>0.909</td>
<td>1.611</td>
</tr>
<tr>
<td></td>
<td>( p )</td>
<td>0.634</td>
<td>.447</td>
</tr>
<tr>
<td>Pearson Correlation with Winning %</td>
<td>( r )</td>
<td>-</td>
<td>.066</td>
</tr>
<tr>
<td></td>
<td>( p )</td>
<td>-</td>
<td>.721</td>
</tr>
</tbody>
</table>

In Table 2 above, \( a \) is the matching slope which measures sensitivity to reinforcement, \( \log b \) is the \( y \)-intercept which measures bias, \( R^2 \) measures variance accounted for.
GENERAL DISCUSSION

Impact

*Combining Quantitative Analysis of Behavior and Sabermetrics.* Until now, the matching law—and quantitative analysis of behavior—as applied to sports has developed independently of the field of sabermetrics. To the author’s knowledge, these two complementary fields have now been directly combined for the first time in this study. The current study used the seminal sabermetric measure of Success Rate as the measure of reinforcement in various GME analyses of NFL football play-calling behavior and found it to be a broadly more accurate measure of reinforcement than yards-gained.

Previous applications of the GME to football play-calling behavior used yards-gained as the measure of reinforcement. While this measure was used with some success over several studies, the amount of variation in play-calling it accounted for within the GME analysis was substantially lower than laboratory studies or even GME analyses in other sports (Baum, 1974; Critchfield & Stilling, 2010; Reed et al., 2006; Vollmer & Bourret, 2000). This was an indication that variables important to matching were not being appropriately accounted for in matching analyses of play-calling behavior. Closer examination of the variables used in the previous matching analyses revealed that yards-gained may have been a structural and not fully functional measure of reinforcement. The field of advanced analytics in football has developed multiple methods of measuring the value of a football play, including Success Rate, which incorporates more relevant variables than simply yards-gained alone.

Success Rate was selected for use in this study for several reasons. First, it factored in more relevant variables that affect the quality of reinforcement of a football
play—namely, down-and-distance and yards-gained—than just yards-gained alone. Second, Success Rate classifies plays as either successful or unsuccessful, which more closely resembles how reinforcers are measured in laboratory matching studies. Third, Success Rate is relatively easy to understand and calculate compared to other sabermetric measures in football, which makes it an approachable metric for those not familiar with sabermetrics. Fourth, Success Rate is a seminal measure of the value of a football play in advanced football analytics and is used as a conceptual foundation for more advanced analyses (Berri & Burke, 2012).

**Success Rate as a Measure of Reinforcement.** Across Analysis 1 through Analysis 4, the preponderance of the evidence suggests that Success Rate was a better measure of reinforcement than yards-gained. This was demonstrated through the goodness-of-fit measures in Analysis 1 – Analysis 3 with consistently higher $R^2$ values and lower RMSE values when Success Rate was used as the measure of reinforcement. Matching slopes in Analysis 2 and Analysis 3 were found to be significantly steeper and closer to 1.0, demonstrating a higher sensitivity to reinforcement, when reinforcement was measured with Success Rate instead of yards-gained. The intercept values were found to be significantly different in Analysis 1, with a significant bias for calling passing plays when Success Rate was the measure of reinforcement. Similar results were found for over half of the teams in Analysis 3. In Analysis 2, a significant passing bias was also present when Success Rate was the measure of reinforcement. These findings are in line with conventional wisdom of both football fans and commentators that the NFL is a pass-heavy league. In Analysis 4, none of the fitted parameters from the GME analysis using yards-gained as the measure of reinforcement had a significant correlation with winning.
percentage, while a significant negative correlation was found between winning percentage and the bias term when Success Rate was used as the measure of reinforcement. Since the biases across all of the teams were positive, the negative correlation with winning percentage demonstrated that teams with a bias closer to perfect matching \((\log b = 0.0)\) tended to have a higher winning percentage than teams with larger passing biases.

Though the majority of results favor the use of Success Rate as the measure of reinforcement, some results were inconclusive or could be argued to favor the use of yards-gained as the measure of reinforcement. No significant differences between matching slopes—the measure of sensitivity to reinforcement—were found at the season-aggregate league level in Analysis 1A. Such differences were, however, demonstrated across the league level in Analysis 2. No teams were found to have significant differences between their matching slopes in the within-team portion of Analysis 3, though significant differences were found between the slopes in the across-team portion of Analysis 3. For six teams in Analysis 3, \(R^2\) values were higher when yards-gained was used. For five of those teams, \(RMSE\) was lower when yards-gained was used. This means that for those teams, the model using yards-gained as a measure of reinforcement fit the data better than the model using Success Rate. In Analysis 4, only intercepts were found to significantly correlate with winning percentage when Success Rate was used as a measure of reinforcement. While this was one more significant correlation between a matching parameter and winning percentage than when yards-gained was used in the current study, Reed et al. (2006) found significant correlations with winning percentage for both variance accounted for and intercepts using yards-gained as reinforcement.
**Potential Applications.** The GME could potentially be used by a coaching staff at least three ways. First, play-callers can use GME analyses to influence their play calling strategy in attempt to maximize reinforcement. Sakai & Fukai (2008) suggested that using a strategy to achieve matching behavior can be beneficial to reward maximization when the specific conditions of maximization are not known to the performer, which is the case in football play-calling. It is generally understood, to the point of being regularly mentioned by football announcers on television, that coaches typically script the first several plays of a game, sometimes up to two full offensive series. Coaches could potentially use a GME analysis of the first two drives to set or adjust play calling strategy as the game progresses in an attempt to achieve perfect matching and thus potentially maximize reinforcement. Second, play-callers could be provided feedback from, and potentially be evaluated on their matching performance when calling plays. There is currently no broadly applied way of determining whether a coach called the “correct” ratio of passing plays to rushing plays. The GME could provide a way of identifying bias in the OC’s play calling, and providing feedback on if he called the best possible ratio of plays given the game situation he was in. Third, coaches could be selected for play-calling roles based in-part on their matching history when calling plays throughout their career. If attempting to match can serve as a viable approach to maximizing reinforcement, then matching performance would be of interest in selecting a coach to call plays.

**Benefits of Translational Research in Quantitative Analysis of Behavior.** Nevin (2008) argued that the value of quantitative models of behavior are only really relevant to the extent that they can be applied to the natural complex environment. In behavior
analysis, the translation of the matching law outside the operant lab has been regarded as one of the most promising advances of the field (Reed, et al., 2011). Sports provides the perfect arena for this translational research for several reasons. First, it provides a natural, but partially controlled environment. The environment is natural and uncontrolled in that there is no experimental manipulation. It is partially controlled in that sports have very specific structures and rules which are enforced. Second, sports are clearly socially significant in that tens of millions of people in the U.S. are interested, at some level, in professional sports. Third, highly detailed and comprehensive data sets, like the one employed in the current study, are available for most of the major professional and college sports either for free or reasonable purchase. These data provide both measures of performance and descriptions of the context in which the performance occurred, which means that behavior and potential reinforcers can be derived directly or indirectly from the existing available datasets. Fourth, many behavioral researchers are not familiar with the technical language and math involved in quantitative analysis of behavior. Many people likely find elite sports more interesting and may be more likely to understand quantitative models as they are applied to the performance of their favorite sports team or league, versus a traditional laboratory research setting. Continuing the line of translational research between quantitative analysis of behavior and sports may soon open the door for applied opportunities where quantitative models of behavior can be applied in sports settings to improve decision making for athletes and coaches.

Limitations

The field of matching law research in football play-calling has had the limitation of not accounting for or incorporating numerous variables that are thought to be
important in determining the value of a play. The current study evaluated Success Rate as compared to yards-gained as a measure of reinforcement to address this limitation. While Success Rate does incorporate both yards-gained and down-and-distance into determining the value of a play, it does not incorporate any other variables. Field position, time remaining in the game, score, strength of opposing defense, home-field advantage, weather, injuries, time-of-possession, number of time outs remaining, and many other variables thought to be important to play-calling decisions are not accounted for in the current study. Despite this limitation, the use of Success Rate alone as the measure of reinforcement in a GME analysis accounted for 86% of the variance in NFL play-calling behavior.

There were also limitations in the comparative analyses used in the current study. First, there was no statistical comparison of $R^2$ values, rather $R^2$ was compared descriptively and through visual analysis. The current study compared models which had only one predictor variable each. Procedures for comparing $R^2$ between two models which use the same set of data but a different number of predictor variables are available, but were not appropriate for the current study (Motulsky & Christopoulis, 2004). Second, the ANCOVA procedure implemented in this study to analyze difference in matching parameters could not statistically compare intercepts if slopes were found to be significantly different. This limitation could possibly be overcome in future research by implementing methods used to compare regression models in biological sciences research, which have been applied with some success to behavioral research as well (see Magoon & Critchfield, 2008).
Questions regarding causation were not able to be answered in the current study due to its observational nature and lack of experimental manipulation. It is possible that the matching phenomena in the present study are artifactual and do not demonstrate a real functional relationship between the measures used as reinforcement in the analyses and play-calling behavior. While an artifactual relationship is not suspected to be the case, the lack of experimental manipulation precludes the ability of the analyses to conclusively determine otherwise.

Success Rate was chosen as the advanced analytics measure of reinforcement for this study due in part to its simplicity so that it could be understood and used by those outside of the field of sabermetrics. Due to its simplicity and the fact that it has been in use at least since Carroll et al. (1988), Success Rate is hardly a cutting-edge metric in the field of football sabermetrics. Currently, EPA models are considered the standard in evaluating the value of a play with WPA models being more cutting-edge (Berri & Burke, 2012; Lock & Nettleton, 2014).

The current study only conducted correlational analyses between the matching parameters and winning percentage. Winning percentage is arguably the most important football outcome in the regular season, but it certainly is not the only important football outcome. There are numerous relevant metrics that should be analyzed for significant correlations with matching parameters such as offensive ranking, yards-gained on offense, points-scored on offense, time-of-possession, and perhaps countless others.

Reed et al. (2006) analyzed the pre-season, regular-season, and post-season games using the GME with yards-gained as the reinforcer and also conducted season-aggregated analyses across three decades of NFL seasons. The current study analyzed data from one
NFL regular season at several different levels. Data from other NFL seasons should be analyzed to determine trends over time. Additionally, analyses of post-season data would be of interest as the teams playing in the post season are more competitive and the extreme consequences of winning or losing games—advancing in the playoffs or immediate elimination from the playoffs, respectively—are commonly thought to impact play-calling decisions.

The current study only evaluated play-calling behavior in the NFL, while though it is the premier professional football league, it is not the only highly-competitive football league. It is expected that different GME outcomes would be found in different leagues, as was the case in Reed et al. (2006). Perhaps Success Rate would be found to be less effective than yards-gained as a measure of reinforcement in different leagues. If this were the case, then it may be that different variables influence the quality of reinforcement for play-calling behavior in different leagues.

Records detailing who called each play, the actual play called, and whether an audible was called were not available and likely may never be available to the public. Given this limitation in data availability, the analyses in the current study were forced to assume that there was only one play-caller for each team. This is certainly not true, given that quarterbacks call audible on the field after the coach calls a play.

**Future Research**

The results of the current study demonstrate the value of Success Rate in GME analyses of play-calling behavior in professional football. Future studies should expand the current analysis in several ways. Longitudinal studies should be conducted at the league level across several seasons to determine if the use of Success Rate and yards-
gained produces significant differences in GME analyses of play-calling behavior as were found in this study. Once this level of replication has been completed, focusing attention on the team and individual level become of great interest.

**Longitudinal Studies.** Team personnel and coaching staff largely remain the same during the NFL season with more substantial changes occurring during the offseason between NFL seasons. Investigating changes and trends over several years in matching behavior at the team level across the league could provide insight as to when, how, and, to a lesser extent, why changes happen within a team over time. One specific area of interest would be to determine how matching changes at the team level when key personnel, such as the starting quarterback, change between seasons. Assuming that the play-caller stays the same through the quarterback change, would the play-caller’s matching behavior change with a significant change in personnel, or would it remain unchanged perhaps due to a well-established play-calling repertoire with an idiosyncratic matching history? A similar experimental question can be investigated when a play-caller changes teams. If it is the case that play-callers maintain similar matching behavior in their play-calling over time and across several different environmental conditions—different quarterbacks, different teams, etc.—then the GME analysis could potentially be used as a tool to evaluate and select play-callers. This would serve as an immense opportunity for behavior analysis, and more specifically the quantitative analysis of behavior, and serve as an inroad to application in professional sports.

**Other Advanced Analytic Measures.** A limitation of the current study discussed above was that Success Rate is not a cutting-edge tool in advanced analytics of football and that it only incorporates yards-gained and down-in-distance in determining the value
of a play. Future studies should address this limitation by using more relevant and complex sabermetric measures, such as EPA or WPA, as the measure of reinforcement. Perhaps one would account for more variance than Success Rate or produce significantly different slopes and intercepts that more accurately represent matching behavior in NFL play-calling. Though this is a promising area of study and would further combine sabermetrics with quantitative analysis of behavior, there are two foreseeable obstacles that will need to be overcome in conducting this research. First, EPA and WPA are far more difficult to calculate than Success Rate. In order to determine EPA or WPA, an EP or WP model must first be built which requires several seasons of play-by-play data and advanced statistical analysis. Once a model is built, it must be tested and adjusted to ensure its accuracy. Given the dearth of academic research in football sabermetrics, it is difficult to use previously existing models that have been sufficiently tested and documented. Lock & Nettleton (2014) provided a method for constructing and testing a WP model, which can then be used to calculate WPA, but it is more a demonstration of machine learning using random forest\textsuperscript{5} method than a beginner’s guide to building WP models. The second obstacle to overcome is that EPA and WPA can produce negative numbers at the game and season level for both passing and rushing plays. A transformation of negative values to positive values would be required in order for the data to be subjected to the logarithmic transformation inherent in the GME.

\textit{Non-Binary Response Options in Play-Calling.} To date, the line of research on play-calling in football has focused on the dichotomy of calling passing-plays or rushing

\textsuperscript{5} Random forest is an ensemble learning approach involving repeated independent random vector sampling of data with the same distribution for use in decision trees which each produce a predicted value. The predicted values from these trees are combined together in a forest of prediction trees to generate combined predictions (Breiman, 2001; Lock & Nettleton, 2014).
plays. A casual observer of football will quickly note that there are many types of passing plays and rushing plays. One reason for analyzing play-calling options as a merely dichotomous choice has been the lack of more specific play-calling data. Highly detailed NFL play-by-play datasets, such as the one purchased from Arm Chair Analysis at armchairanalysis.com for this study, are now available which provide play-calling data such as whether a pass was thrown short, medium, or long range and whether a rushing play ran to the left, center, or right side of the field. The data provided in these datasets opens the door to thousands of potential studies using quantitative analysis of behavior, traditional sports stats, and even advanced analytics.

**Simulation.** The studies of matching behavior in football and sports in general have analyzed existing data collected from the natural sporting environment without any experimental manipulation. It would be implausible to gain experimental control of any relevant variables in a professional football game. For this reason, simulations of play-calling choice behavior could provide a more reasonable approach to gaining experimental control of the relevant variables. Video game based simulations would seem to provide a suitable platform as the play-caller could chose a play and then watch the play unfold, similar to the sequence of events that takes place in a live-game play-calling situation. However, video game simulations would still require gaining control of the relevant variables within the video game itself, which could prove difficult as football video games are designed to provide an enjoyable gaming experience, not to be easily manipulated as a research tool by end users. Analog simulations could be created on a computer using relatively simple programming where the play-caller is presented with relevant game situation information (e.g. field position, down-and-distance, score, time
remaining on the game clock, etc.) selects a passing play or rushing play button on a computer screen and the result of the play in yards-gained or Success Rate is displayed immediately afterwards. While this would not provide a visual simulation of in-game play-calling, it could provide a robust and more readily controllable simulation of the choices and outcomes of play-calling.

**Other Choice Behaviors in Sports.** Football play-calling is only one of many choice situations in competitive sports to which the GME can be applied. For example, the GME could be applied to other choice behaviors in football such as a quarterback’s decision to run the ball himself or pass the ball on read-option plays. An analysis of this specific choice behavior may be more difficult than passing/rushing play-calling, but it would be both possible. The breadth and depth of data in play-by-play records in NFL may make this analysis possible using existing datasets. Otherwise, it would be possible to record the necessary data by watching video of the relevant game situations. As in the current study, yards-gained and Success Rate could both be considered as potential measures of reinforcement.

On the defensive side of football, a defensive coordinator’s play-calling choices could be analyzed using the GME and various traditional football stats and sabermetric measures as reinforcement. One example might be whether the defensive coordinator calls for the defense to play zone coverage or man-to-man coverage against the opposing offense’s passing attack. The offense’s Success Rate measure used in the current study could be inverted to measure play success for the defense. Another example of choice in defensive play-calling might be whether or not a defensive coordinator calls a blitz-play. Sacks or hits on the quarterback may be counted as a traditional measure of
reinforcement, while many sabermetric measures of the opposing offense’s performance could be inverted to be measures of defensive performance.

The GME has already been shown to have value in analyzing choice behavior in other sports outside of football such as basketball and hockey (Vollmer & Bourret, 2000; Senuick et al., 2014). Future research could use a method similar to the one in the current study for combining the GME and sabermetric measures of reinforcement. Baseball seems to be the sport most open to this type of analysis, given its established use of sabermetrics. Poling et al. (2011) conducted a GME analysis of batting handedness in baseball which used several traditional baseball stats including TB, RBI, and HR data as measure of reinforcement for batting. A similar study of batting behavior could be conducted using a sabermetric measure of batting performance as the measure of reinforcement. Pitching behavior could also be evaluated with the type of pitch (fast ball, curve ball, knuckle ball, slider, etc.) being the response options and traditional or sabermetric pitching stats being the measure of reinforcement.

**Conclusion**

This current study is the first to combine the quantitative analysis of behavior with sabermetrics. Success Rate, a sabermetric measure of the value of a football play, was combined with the GME, a model of choice from quantitative analysis of behavior, to successfully account for play-choice selection in NFL football. The sabermetric measure used is simple to understand and calculate and is therefore relatively accessible to non-sabermetricians within the field of behavior analysis. The accessibility of Success Rate creates an on-ramp for behavior analysts using quantitative analysis of behavior to merge their work with the field of sabermetrics. Likewise, the matching law at the core of
the GME is conceptually straightforward and the GME is relatively easy to calculate and interpret for those familiar with regression analysis. In this way, the GME can serve as a useful and approachable analysis to introduce sabermetricians to the field of quantitative analysis of behavior, especially the analysis of choice behavior in elite sport competition. The current findings suggest that future studies in this line of research should use Success Rate as a measure of reinforcement of football play-calling behavior and that other advanced analytics measures should be studied in quantitative models of behavior. Behavior analysts serious about incorporating the quantitative analysis of behavior in sports would do well to familiarize themselves with sabermetrics and to continue building a bridge between the two fields through combined research, emphasizing the potential practical applications of such analyses to providing a competitive advantage to sports teams.


Deford, F. (1972, March 6). It ain’t necessarily so, and never was. *Sports Illustrated*, p. 59-60.


James, B. (1977). Baseball abstract: Featuring 18 categories of statistical information that you just can’t find anywhere else. (n.p.): Author.


Appendix A

Plot of Residuals for Analysis 1A and Analysis 1B
Appendix B

Plot of Residuals for Analysis 2A and Analysis 2B
Analysis 2A: Residuals

Analysis 2B: Residuals
Appendix C

GME Graphs for Analysis 3A and Analysis 3B Arranged by Team
Arizona

Analysis 3A: Yards-Gained as Reinforcement

\[
\begin{align*}
y &= 0.512x + 0.054 \\
R^2 &= 0.610
\end{align*}
\]

Analysis 3B: Success Rate as Reinforcement

\[
\begin{align*}
y &= 0.709x + 0.088 \\
R^2 &= 0.844
\end{align*}
\]

Atlanta

Analysis 3A: Yards-Gained as Reinforcement

\[
\begin{align*}
y &= 0.537x - 0.068 \\
R^2 &= 0.597
\end{align*}
\]

Analysis 3B: Success Rate as Reinforcement

\[
\begin{align*}
y &= 0.609x + 0.040 \\
R^2 &= 0.564
\end{align*}
\]

Baltimore

Analysis 3A: Yards-Gained as Reinforcement

\[
\begin{align*}
y &= 0.363x + 0.132 \\
R^2 &= 0.185
\end{align*}
\]

Analysis 3B: Success Rate as Reinforcement

\[
\begin{align*}
y &= 0.829x + 0.081 \\
R^2 &= 0.432
\end{align*}
\]
Buffalo

Analysis 3A: Yards-Gained as Reinforcement

\[ y = .310x + .010 \]
\[ R^2 = .519 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = .464x + .058 \]
\[ R^2 = .546 \]

Carolina

Analysis 3A: Yards-Gained as Reinforcement

\[ y = .461x - .030 \]
\[ R^2 = .282 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = .625x + .061 \]
\[ R^2 = .686 \]

Chicago

Analysis 3A: Yards-Gained as Reinforcement

\[ y = .732x - .046 \]
\[ R^2 = .752 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = .647x + .083 \]
\[ R^2 = .786 \]
Denver

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.635x - 0.067 \]
\[ R^2 = 0.705 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.790x + 0.034 \]
\[ R^2 = 0.704 \]

Detroit

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.246x + 0.149 \]
\[ R^2 = 0.156 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.421x + 0.128 \]
\[ R^2 = 0.453 \]

Green Bay

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.499x + 0.057 \]
\[ R^2 = 0.397 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.826x + 0.043 \]
\[ R^2 = 0.801 \]
Kansas City

Analysis 3A: Yards-Gained as Reinforcement
\[ y = 0.458x + 0.007 \]
\[ R^2 = 0.239 \]

Analysis 3B: Success Rate as Reinforcement
\[ y = 0.642x + 0.048 \]
\[ R^2 = 0.651 \]

Los Angeles

Analysis 3A: Yards-Gained as Reinforcement
\[ y = 0.440x + 0.050 \]
\[ R^2 = 0.409 \]

Analysis 3B: Success Rate as Reinforcement
\[ y = 0.456x + 0.143 \]
\[ R^2 = 0.438 \]

Miami

Analysis 3A: Yards-Gained as Reinforcement
\[ y = 0.486x - 0.048 \]
\[ R^2 = 0.342 \]

Analysis 3B: Success Rate as Reinforcement
\[ y = 0.618x + 0.008 \]
\[ R^2 = 0.605 \]
Minnesota

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.510x - 0.018 \]
\[ R^2 = 0.263 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.797x + 0.024 \]
\[ R^2 = 0.622 \]

New England

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.494x - 0.081 \]
\[ R^2 = 0.536 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.768x + 0.018 \]
\[ R^2 = 0.786 \]

New Orleans

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.664x - 0.069 \]
\[ R^2 = 0.692 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.890x + 0.023 \]
\[ R^2 = 0.709 \]
New York (Giants)

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.569x - 0.038 \]
\[ R^2 = 0.554 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.893x + 0.007 \]
\[ R^2 = 0.757 \]

New York (Jets)

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.522x + 0.002 \]
\[ R^2 = 0.551 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.627x + 0.098 \]
\[ R^2 = 0.721 \]

Oakland

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.320x + 0.066 \]
\[ R^2 = 0.437 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.413x + 0.104 \]
\[ R^2 = 0.597 \]
Philadelphia

Analysis 3A: Yards-Gained as Reinforcement
\[
y = 0.613x - 0.009 \\
R^2 = 0.602
\]

Analysis 3B: Success Rate as Reinforcement
\[
y = 0.741x + 0.098 \\
R^2 = 0.581
\]

Pittsburgh

Analysis 3A: Yards-Gained as Reinforcement
\[
y = 0.546x - 0.025 \\
R^2 = 0.577
\]

Analysis 3B: Success Rate as Reinforcement
\[
y = 0.705x + 0.065 \\
R^2 = 0.914
\]

San Diego

Analysis 3A: Yards-Gained as Reinforcement
\[
y = 0.443x - 0.018 \\
R^2 = 0.773
\]

Analysis 3B: Success Rate as Reinforcement
\[
y = 0.452x + 0.094 \\
R^2 = 0.604
\]
San Francisco

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.216x + 0.052 \]
\[ R^2 = 0.191 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.456x + 0.074 \]
\[ R^2 = 0.694 \]

Seattle

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.204x + 0.113 \]
\[ R^2 = 0.198 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.415x + 0.107 \]
\[ R^2 = 0.319 \]

Tampa Bay

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.673x + 0.113 \]
\[ R^2 = 0.475 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.773x + 0.011 \]
\[ R^2 = 0.601 \]
Tennessee

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.601x - 0.071 \]

\[ R^2 = 0.654 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.533x + 0.070 \]

\[ R^2 = 0.650 \]

Washington

Analysis 3A: Yards-Gained as Reinforcement

\[ y = 0.631x - 0.066 \]

\[ R^2 = 0.810 \]

Analysis 3B: Success Rate as Reinforcement

\[ y = 0.711x + 0.056 \]

\[ R^2 = 0.878 \]
Appendix D

Plots of Residuals for Analysis 3A and Analysis 3B Arranged by Team
Arizona

Analysis 3A: Residuals

Analysis 3B: Residuals

Atlanta

Analysis 3A: Residuals

Analysis 3B: Residuals

Baltimore

Analysis 3A: Residuals

Analysis 3B: Residuals
Cincinnati

Analysis 3A: Residuals

Analysis 3B: Residuals

Cleveland

Analysis 3A: Residuals

Analysis 3B: Residuals

Dallas

Analysis 3A: Residuals

Analysis 3B: Residuals
Kansas City

Los Angeles

Miami
Philadelphia

Pittsburgh

San Diego
Tennessee

Analysis 3A: Residuals

Analysis 3B: Residuals

Washington

Analysis 3A: Residuals

Analysis 3B: Residuals
Appendix E

Plot of $R^2$ Values and Means for Analysis 3A and Analysis 3B
Yards-gained was used as the measure of reinforcement in Analysis 3A. Success Rate was used as the measure of reinforcement in Analysis 3B. Bars were plotted to identify the means.
Appendix F

$R^2$ Values by Team
Appendix G

Required Protocol Clearance Letter
Date: December 20, 2017

To: Douglas Johnson, Principal Investigator
   Jacob Bradley, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: Approval not needed for HSIRB Project Number 17-12-21

This letter will serve as confirmation that your project titled “Mixing Matching and Sabermetrics: Combining Advanced Analytics and the Generalized Matching Law n NFL Football Play-Calling” has been reviewed by the Human Subjects Institutional Review Board (HSIRB). has been reviewed by the Western Michigan University Institutional Review Board (WMU IRB). Based on that review, the WMU IRB has determined that approval is not required for you to conduct this project because the scope of work does not meet the Federal definition of human subject.

45 CFR 46.102 (f) Human Subject

(f) Human subject means a living individual about whom an investigator (whether professional or student) conducting research obtains

(1) Data through intervention or interaction with the individual, or
(2) Identifiable private information.

Intervention includes both physical procedures by which data are gathered (for example, venipuncture) and manipulations of the subject or the subject's environment that are performed for research purposes. Interaction includes communication or interpersonal contact between investigator and subject. Private information includes information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information which has been provided for specific purposes by an individual and which the individual can reasonably expect will not be made public (for example, a medical record). Private information must be individually identifiable (i.e., the identity of the subject is or may readily be ascertained by the investigator or associated with the information) in order for obtaining the information to constitute research involving human subjects.

“About whom” – a human subject research project requires the data received from the living individual to be about the person.

Thank you for your concerns about protecting the rights and welfare of human subjects.

A copy of your protocol and a copy of this letter will be maintained in the HSIRB files.