Big Data and Machine Learning for Behavioral Analytics and Inference: Cases in Sports and Education

by

Emrah Yılmaz

A Dissertation Submitted to the
Graduate School of Business
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Business

KOÇ ÜNİVERSİTESİ

August 24, 2022

Big Data and Machine Learning for Behavioral Analytics and Inference: Cases in Sports and Education

Koç University

Graduate School of Business

This is to certify that I have examined this copy of a master's thesis by

Emrah Yılmaz

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Assoc. Prof. Özden Gür Ali (Advisor)

Prof. Seda Ertaç Güler

Asst. Prof. Muhammed A. Yıldırım

Prof. Theos Evgeniou

Asst. Prof. Mustafa Gökçe Baydoğan

Date:

This thesis contains no material which has been accepted for any award or any other degree or diploma in any university or other institution.

Chapter 1 is individual work under the supervision of my advisor - Assoc. Prof. Özden Gür Ali.

Chapter 2 is individual work under the supervision of my advisor - Assoc. Prof. Özden Gür Ali.

Chapter 3 is joint work with Prof. Seda Ertaç Güler, Daniel L Chen, Theodoros Evgeniou, Xin Miao and Ali Nadaf.

Signature: ______________________________ Emrah Yılmaz

ABSTRACT

Big Data and Machine Learning for Behavioral Analytics and Inference: Cases in Sports and Education
Emrah Yılmaz
Doctor of Philosophy in Business
August 24, 2022

This thesis focuses on the use of big data and machine learning methods in behavioral analytics and causal inference. The main motivation of the thesis is to illustrate how the researchers working with traditional econometric methods can benefit from big data and causal ML methods. In the absence of well-established literature, finding the right regression specification is a challenging task, especially when working with high dimensional data set. In this study, I have combined causal ML techniques with explainable AI methods and provided guidelines on how to measure heterogeneous treatment effects with the right regression specification (i.e. which main effects and interactions to be used, what control variables to be included). To empirically test these guidelines, I have curated a large data set in football including detailed variables about interim feedback, match-specific conditions, team features, and most importantly manager characteristics. Empirical evidence contributes to the sports analytics literature suggesting when and how risk-taking behavior of football managers pays off in light of interim and ex-ante information revealed to the manager (i.e. the decision maker). Moreover, this thesis contributes to the causal ML literature by evaluating the performances of two well-known causal ML techniques (a recently popular matching algorithm focusing on finding average treatment effects (FLAME) and Causal Forest that directly aims to estimate

heterogeneous treatment effects) are evaluated by using synthetic data generated with known heterogeneous treatment effects.

In addition to sports analytics, I have also worked with education data and demonstrated how grit, a non-cognitive skill, predicts academic achievement for students. I used a unique dataset from a digital learning platform to construct a behavioral measure of grit and showed that behavioral grit is a better predictor of student performance compared to survey grit that has been traditionally used by the researchers. I have also found that machine learning algorithms perform well in predicting academic resilience even without constructing any structural model or regression specification, thanks to the power of big data. I believe that my findings from cases in sports and education put forward the benefits of using Machine Learning and big data for researchers working with traditional and theory-based models for causal inference.

ÖZETÇE

Davranış Analitiği ve Nedensel Çıkarımlarda Büyük Veri ve Makine Öğrenimi: Spor ve Eğitimden Vakalar

Emrah Yılmaz
İşletme, Doktora
24 Ağustos 2022

Bu tez, davranış analitiği ve nedensel çıkarımda büyük veri ve makine öğrenimi yöntemlerinin kullanımına odaklanmaktadır. Tezin temel motivasyonu, geleneksel ekonometrik yöntemlerle çalışan araştırmacıların büyük veri ve nedensel Makine Öğrenmesi yöntemlerinden nasıl yararlanabileceğini göstermektir. Eğer bir konuda kapsamlı bir literatür yoksa, doğru regresyon spesifikasyonunu bulmak, özellikle yüksek boyutlu veri seti ile çalışırken zorlu bir iştir. Bu çalışmada nedensel Makine Öğrenimi tekniklerini açıklanabilir Yapay Zeka yöntemleriyle birleştirdim ve heterojen tretman etkilerinin doğru regresyon spesifikasyonu oluşturularak nasıl ölçüleceğine dair kılavuzlar (yani, bir regresyonda hangi ana değişkenler ve etkileşim değişkenleri kullanılacağı, hangi kontrol değişkenleri modele dahil edileceği) oluşturdum. Bu yönergeleri ampirik olarak test etmek için, futbolda maç içi geri bildirimler, maça özgü koşullar, takım özellikleri ve en önemlisi yönetici özellikleri hakkında ayrıntılı değişkenler içeren büyük bir veri seti oluşturdum. Ortaya koyduğum ampirik kanıtlar, futbol yöneticilerinin risk alma davranışlarının maç sırasında alınan geri bildirimlerden ve maç öncesinde gözlemlenen bilgilerden ne zaman ve nasıl etkilendiğini göstererek spor analitiği literatürüne katkıda bulunmaktadır. Ayrıca, bu tez, bilinen heterojen tretman etkileri ile üretilen sentetik verileri kullanarak iyi bilinen iki nedensel Makine Öğrenimi tekniğinin (Son zamanlarda popüler olan ve ortalama tretman etkilerini bulmaya odaklanan FLAME ve doğrudan heterojen tretman etkilerini bulmaya

çalışan Nedensel Orman) performanslarını değerlendirerek nedensel Makine Öğrenimi literatürüne katkıda bulunmaktadır.

Spor analitiğine ek olarak, eğitim verileriyle de çalıştım ve bilişsel olmayan bir beceri olan azmin öğrenciler için akademik başarıyı nasıl öngördüğünü gösterdim. Davranışsal bir azim ölçüsü oluşturmak için dijital bir öğrenme platformundan benzersiz bir veri kümesi kullandım ve davranışsal olarak ölçülen azmin, araştırmacılar tarafından geleneksel anketlerle ölçülen azim ölçüsüne kıyasla öğrenci performansının daha iyi bir prediktörü olduğunu gösterdim. Ayrıca, büyük verinin gücü sayesinde, makine öğrenimi algoritmalarının, herhangi bir yapısal model veya regresyon spesifikasyonu oluşturmadan bile akademik dayanıklılığı tahmin etmede iyi performans gösterdiğini buldum. Spor ve eğitimdeki vakalar üzerinde çalışarak elde ettiğim ampirik bulguların, nedensel çıkarım yapmak için geleneksel ve teoriye dayalı modellerle çalışan araştırmacıların Makine Öğrenimi ve büyük veriden sağlayabileceği faydaları açıkca ortaya koyduğuna inanıyorum.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor Assoc. Prof. Özden Gür Ali for her precious guidance and support throughout this thesis. It has been a great experience for me to study under her supervision.

I would like to thank Prof. Seda Ertaç Güler for her help, support, collaboration and valuable comments.

I owe special thanks to my dear parents, brother and wife for their continuous love and support.

I would also like to thank my friends Zeynel Harun Alioğulları, Ergün Kotan and Sündüz Divle for their friendship, encouragement and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

MODELING THE HETEROGENEOUS EFFECT OF RISK TAKING BEHAVIOR OF FOOTBALL MANAGERS ON MATCH OUTCOMES

1.1 Introduction

Sports do not only appear as an activity positively impacting people's well-being, they also emerge as one of the fastest growing industries. The global sports industry is worth up to about \$600 billion today (Figure A.1), and its long-term prospects appear strong with the current rate of growth that is faster than global GDP. Sports analytics market worldwide, in particular, has been reached to \$2.2 billion as of 2020 (EmergenResearch2021) and expected to continue growing with a CAGR of 21.8% in the upcoming years. Along with that, there has been a growing literature in sports management and behavioral sports analytics (Figure A.2 and A.3).

Main drivers of the growth of sports analytics market are increasing demand for performance tracking and analysis of sport agents (e.g. players, coaches). Thanks to rapid developments in data technologies such as player movement and body pose detection system, managers (e.g. coaches, professional team managers, federations) can access large volume of data, which allows them to assess large statistics, recognize important behavioral patterns and make better decisions. By utilizing behavioral data, training programs can be optimized, better players can be hired and managers can select best tactical systems, etc. Like in any industry, data now plays a crucial role in maximizing return of investments that any sport team has been doing.

Risk taking in sports, as an important type of decision making, has also taken great attention in the literature. Tournaments are used as an incentive mechanism in workplaces and sports to improve performance among participants and reward the one with the highest performance. On the other hand, individuals faced by a tournament can often not only choose their effort level but also the risk level of their strategy. Therefore, understanding the determinants of risk seeking behaviour and how they affect performance in tournaments has become an important concern. As a well-known and prevalent type of tournament, football has also been studied a lot in the risk taking literature. Risk level of a strategy in football is usually defined as how aggressive players in the field are. For instance, initial team formation (i.e. the number of forward and midfield players in field), or how the coach substitutes a player (e.g. replacing a midfield player with a forward one or vice-versa) are some of the main variables defined as risk taking behaviour [Bartling et al.2015, Bucciol et al.2019, Grund and Gürtler*2005].

To understand risk taking behaviours in sports, we first need to understand why individuals or teams have different attitudes towards risk in general, how people decide on the riskiness of their strategies in tournaments and how managers make risky decisions for their teams. When we look at the general findings of previous studies, we can classify the factors that could potentially predict risk taking behaviours in sports mainly into two subgroups: 1) Ex-ante information including individual (i.e. the manager) characteristics, team features and match characteristics, and 2) interim information that includes all types of feedback revealed during the game.

In this chapter, we will focus on the heterogeneity of managerial risk taking behaviors on game results. Our paper contributes to the literature focusing on determinants of managerial risk taking and its impact on performance in two ways: Firstly, this study will analyze several factors documented as determinants of risk taking in the literature with a large date-set, hence provide a more robust analysis about determinants of risk taking behaviour in football. Secondly, this study will try to investigate heterogeneity of the effect of risk taking on performance in detail by using diverse team, match and manager characteristics. To the best of our knowledge, this will the first study in football analytics literature for such purposes. Furthermore, this chapter will create a baseline for Chapter 2 where we will use causal machine learning techniques to investigate heterogeneous risk taking behavior and its impact on performance. Output of machine learning in Chapter 2 will be used to fine-tune the results of this chapter and eventually we will be using better regression models to model the relationship between risk taking behavior and performance in football.

Rest of the chapter is organized as follows: Section 1.2 summarizes the literature on the determinants of risk taking in sports and effects of risk taking on performance in sports. In Section 1.3, we introduce our data-set and provide descriptive statistics. Section 1.4 presents results and Section 1.5 concludes with a discussion.

1.2 Literature Review

Risk preferences has long been considered an important part of strategic choice and largely studied in the economics literature. Experimental studies conducted in laboratory settings largely focus on measuring attitudes towards risk, investigating the determinants of risk preferences and the relationship between those preferences and real outcomes [Alan et al.2017, Apicella et al.2008, Barseghyan et al.2013, Bellemare and Shearer2010, Charness et al.2013, Croson and Gneezy2009, Pondorfer et al.2017], while empirical studies working with the observational data have examined how risk taking behavior affects performance and outcomes in different fields [Adams and Waddell2018, Bromiley1991, Genakos and Pagliero2012, Sanders and Hambrick2007].

Risk taking in business has also drawn great interest among scholars [Bromiley1991, Desai2008, Mollah et al.2017, Naldi et al.2007, Sanders and Hambrick2007]. For instance, an earlier study of [Bromiley1991] investigates the determinants of organizational risk taking and tests the effect of risk taking on future performance. He defines risk as the variance in security analysts' forecasts of a company's income stream, which could be expected to be a direct consequence of risk taking strategies the company employed. According to his findings, risk taking is significantly predicted by poor performance and result in further poor performance. [Desai2008] studies U.S.

railroad firms in the period from 1978 through 2003. In contrast to [Bromiley1991], he examines managerial risk taking instead of organizational risk measured through income stream variation, and defines risk taking as capacity expansions. The idea behind his choice is that firms generally determine capacity according to demand predictions, but demand fluctuations can be highly unpredictable; therefore, the consequences of any capacity expansion could be uncertain and highly variate. His analyses suggest that higher operating experience, legitimacy and the age of the company moderate risk taking subsequent to performance shortfalls.

Risk taking is generally associated with investment decisions in business and defined as an action or strategy that increases the variation of return [Bromiley1991]; however, this definition is not limited to financial decisions. Almost any action that changes variation of outcome and consequently of payoff in a specific domain can be considered as a risk taking behaviour. For instance, [Wang et al.2013] investigate risk seeking behaviour in advertising industry and define risk taking as the willingness to tolerate uncertainty in performance outcomes related to novel and unconventional advertising campaigns. The authors argue that originality inherently comes together with risk and uncertainty since highly original campaigns might be poorly evaluated or even evoke negative market responses. Their study concludes that the advertiser's creative risk taking propensity increases campaign originality and performance. As an example from the outside of corporate industry, [Llewellyn et al.2008] examine risk taking among rock climbers by defining risk taking as the frequency and difficulty of high and medium risk rock climbing.

The previous studies in the literature have answered different questions using data from distinct fields; however, there is a common definition of risk taking across all domains: When an agent makes a risky decision or move, the probability of obtaining both a high and low payoff increases. In short, just like in the context of finance, risk taking behaviour in any domain implies an increase in the variation of payoffs.

1.2.1 Determinants of Risk Taking in Sports

Ex-Ante Information

Individual characteristics (i.e. manager characteristics in our context as the managers are assumed to be the ones making risky decisions on the field) are shown to have a role in explaining differences in risk taking behaviour. As an example of individual specific factors that determine risk taking, gender has drawn a considerable attention from scholars. Gender differences in risk aversion in experimental settings and observational data are well documented in the economics literature [Agnew et al.2008, Booth and Nolen2012, Charness and Gneezy2012, Croson and Gneezy2009, Ertac and Gurdal2012] . There are also several studies documenting gender differences in risk taking in sports [Böheim et al.2022, Böheim and Lackner2015, Böheim et al.2016, Feess et al.2016, Lackner and Sonnabend2020]. However, there is not many studies on the effect of gender on risk taking in football, as most of the agents (players and coaches) have been male in football so far. Women football is becoming more popular nowadays, however, data availability is still an issue.

The previous literature has also attempted to understand the relationship between age and risk taking behaviour. [Ruedl et al.2010] provide a real-life evidence and report that being above forty years old significantly predicts risk seeking among recreational skiers and snowboarders. In the context of football, there is no evidence provided yet in the literature for effect of age on risk taking behavior, which makes our study the first one analyzing such behavior.

Nationality and religiousness have been also analyzed as a determinant of risk taking behaviour. [Zinkhan and Karande1991] conduct the same experiment measuring risk preferences with both Spanish and American MBA students, and observe that the Spanish sample was less risk averse than the American sample. Similarly, [Bartke and Schwarze2008] study determinants of risk tolerance using a large data-set, German Socio-Economic Panel (SOEP). They focus on emigrants to Germany and find that nationality can not significantly predict risk preferences. On the other hand, they observe significant negative effect of religiousness on risk attitudes. Moreover, religious affiliation also has an influence on tolerance for risk: Muslims are more risk averse than Christians. As a field study focusing on risk taking behaviour on ski slopes, [Ruedl et al.2010] test whether nationality is a predictor of self-reported risk taking and report that there are no significant differences in risk attitudes across nationalities. Similar to age, in our study, nationality will be analyzed for the first time as a potential risk determinant in football.

Ability could also be expected to affect risk taking behaviour. If one is more confident about his ability, when he might make more risky decisions and strategies, believing the probability of success is more likely for him. This argument is supported by the evidence coming from different sports. Feeling confident about their abilities leads players to make more risky choices in rock climbing and skiing [Llewellyn et al.2008, Ruedl et al.2010]. It is not easy to define ability for managers in football as they are not the ones playing the game on the pitch, yet, ability of a manager can be defined as how experienced he is in terms of years and number of clubs or countries he worked for before.

The literature also provides some evidence regarding the effect of different individual tendencies on risk taking. For instance, [Ge2019] connects NFL players' off-field misconduct with their playing strategies in the field. More specifically, he shows that players with arrest record tend to take more aggression-related penalties during the game, suggesting that there could be inherent preferences toward risks or aggression. Another interesting study of [Cross et al.1998] investigates the relationship between gambling and risk taking behaviour among student athletes. Their results suggest that student-athletes who gambled are more likely to perform risk taking behavior than those who did not gamble, providing supportive evidence for correlation between risk seeking behaviours in different domains.

There is a scant literature on the determinants of risk taking behaviour in football. One of the factors that have been studied previously is the past performance. How a team performed in the previous matches could change how aggressive players the couch put in the field. [Bucciol et al.2019] use male soccer data covering eight seasons of the Italian premier league and demonstrate that prior defeats matter. Single and heavy defeats cause the coaches choose more offensive systems of play, whereas multiple defeats result in less risk seeking. It is important to note here that the authors make a further analysis to investigate the heterogeneity in the effect of prior outcomes on risk seeking behaviour. They find that coaches of top teams, those with average salary above 1,000,000 EUR, do not respond to past performances as much as others, and tend to follow more stable strategies by following preset seasonal objectives of their team.

Match characteristics also matter for risk seeking strategies in football. First of all, there are some studies reporting the effect of ex ante probability of winning the match on team formation. [Grund and Gürtler*2005] find that the difference in ranking, defined as the rank of the team minus its opponent's rank, is positively associated with switching into risky strategies. [Bartling et al.2015] analyze expectations from a different perspective by positioning them as reference points. According to their findings, when teams are behind the expected match outcome, measured by pre-match betting odds, then their coaches make significantly more offensive substitutions. [Schneemann2017b] analyzes both the rank of playing teams and the difference in goals at the same time and reports that couches of favorites significantly increases risk when their team is behind with two or more goals, whereas couches of underdogs prefer less aggressive team formation when their team in the lead. Some other examples analyzing effect of rank difference on risk taking from other sport categories include [Neuberg and Thiem2022] for handball and [Yaskewich2017] for car racing.

There also some studies in the literature trying to understand whether playing against a direct rival, a team with a similar ranking, affects managerial risk taking. [Grund and Gürtler*2005] focuses on the starting team formation and demonstrate that playing against a team with a similar ranking does not significantly change the aggressiveness of initial team formation. On the other hand, [To et al.2018] report a significant result that playing a game against a rival partner increases risk taking behaviour. To understand the underlying mechanism behind that result, they conduct additional laboratory experiments involving face-to-face competition and show that rivalry increases risk taking by increasing promotion focus (i.e, motivation to achieve gains) and physiological arousal.

As an interesting result, the place where the game is played also has an influence on the managerial risk taking. [Grund and Gürtler*2005] document that coaches start to the game with a higher number of forward players when they play at home. Similarly, [Böheim et al.2016], [Genakos and Pagliero2012] and [Schneemann2017b] have analyzed the impact of playing at home on risk taking behavior in different sports; however, only [Schneemann2017b] studying risk taking in football reports a positive coefficient for home. Playing at home has not any significant impact on risk taking in basketball [Schneemann2017b] and weight-lifting [Böheim et al.2016].

Interim Information

One could expect that getting a interim information (e.g., performance feedback) during a game can have a influence on managerial decision and actions. A theoretical work of [Nieken and Sliwka2010] provides a model about the role of interim performance feedback for subsequent risk taking behaviour in tournaments. In their study, a risk taking tournament occurs between two agents who choose among a risky and a safe strategy according to the variance of the outcome. They find a Nash equilibrium, in which the leading player chooses the safe strategy, while the one lagged behind opts for the risky strategy. An unpublished study of [Dobson et al.2008] find a similar result from a dynamic game-theoretic model of optimizing strategic behaviour by football teams and test their model empirically, using English football league data.

Empirical studies in the literature also corroborate theoretical findings about the impact of interim information on risk taking. The first variable in the literature as interim information is the current score during the game. An earlier study of [Lee2004] demonstrates that bottom-ranked players in poker tournaments take more risk as the incentive for risk taking is boosted by a larger expected gain. A similar logic applies to football. Since the ranking of the teams depends on the point they get from each match, the incentive for an additional goal depends on the current score which is calculated based on the difference in goals. [Grund and Gürtler*2005] study 306 matches of season 2003/2004 German Major League Soccer and report that the probability of risk taking substitutions decreases with the difference in goals. When a team is ahead, its coach tend to choose a less risky team formation to decrease the likelihood of suffering a goal. On the other hand, when the team is behind, the coach is more willing to substitute existing players with the more risky ones. Another supportive evidence is from the National Basketball Association (NBA) players. [Grund et al.2013] investigate how point differences between teams during games impact their subsequent risk taking behavior. They report that trailing teams increase their risk taking by using more three-point shots. A recent study of [Schneemann2017b] confirms this result by using the data of seven seasons of the UEFA Champions League (from 2009/10 to 2015/16) and documents that trailing teams are more likely to choose riskier strategies, whereas leading teams implement a less risky team formation. [Bartling et al.2015] combine the current score with ex-ante expectations and concludes that coaches take more risks if their teams are behind expectations. [Neuberg and Thiem2022] provide evidence from handball and finds that trailing teams are willing to take more risks. They further analyzes in-game heterogeneity and finds that degree of risk taking under a current deficit is even larger for underdog teams comparing to favourites. Another evidence regarding the impact of interim information on risk taking in tournaments comes from another sport, weight lifting competitions. [Genakos and Pagliero2012] show that professional athletes are more willing to take risks when ranked close enough to the first athlete but choose less risky strategies when ranked lower. Different than other studies, they find that risk taking exhibits an inverted-U relationship with interim rank. Moreover, [Böheim et al.2016] define risk taking in basketball as fraction of three-point shots and finds that point difference increases the likelihood of taking more three-point shots. In golf, there can be several interim information type such as yards to pin, hole difficulty, accum score and [McFall and Rotthoff2020] and [Ozbeklik and Smith2017] analyze effect of these in-game feedback on risk taking behavior.

Having a player taking a red card can also be considered as a kind of negative feedback. Losing a player increases the likelihood of conceding a goal; therefore, coaches might want to modify their team formation to reduce this probability. [Schneemann2017b] shows that losing a player due to a red card is negatively associated with risk taking.

Another kind of information feedback which a player or team face in multi-stage tournaments is the result of the first stage. [Mueller-Langer and Andreoli-Versbach2017] study knock outs from the UEFA Champions League and UEFA Europa League from 1955 until 2009 and do not find any significant differences in risk taking across having negative feedback in the first stage of the tournament. In other words, teams which lagged behind do not significantly increase their risk seeking behaviour in the field.

1.2.2 Effects of Risk Taking on Performance in Sports

Many of the aforementioned authors who have provided determinants of risk taking, have also analyzed whether the risk taking has a significant impact on performance. This is important for policy recommendation outlining the cases where risk taking indeed paysoff.

Starting with the literature on football, [Bucciol et al.2019] have shown prior defeats (i.e ex-ante information) make the coaches taking more risks in the following game. However, increased risk taking, does not pay off in terms of match outcome (not significant in predicting wins). A similar result from [Schneemann2017b] where performance is defined as goal scored and conceded, effect of the cumulative offensive score of each player (i.e risk taking) on the field at that time has no significant impact on performance. Here, we need to note that both authors have not interested in heterogeneity in effect of risk taking on performance, rather in their regressions, they analyzed the average effect of risk taking on performance.

A relevant paper for my study, where my definition of risk taking and performance are very similar, is from [Grund and Gürtler*2005]. They analyze the impact of risk taking substitutions on point advancement which is defined as the difference of points realized at the end of the match and points that are expected to be taken if there is no change in the score from substitution time until the end of the match. They conclude that switching to a strategy of higher risk is less successful than maintaining the initial one. One limitation with their research is that they have only 612 observations and their results may not be very robust. A contrasting finding comes from [Bartling et al.2015] defining performance as both a team's final goal difference and number of points. When they analyze the impact of offensive substitutions on performance, they conclude that the effect is negative; hence, players and coaches does not act in a fully rational way. These two studies show us that even if the effect of risk taking on performance has been analyzed in football literature, there is no consensus among them and the heterogeneity of the effect has not been questioned.

Some recent studies from basketball may give us clues about the heterogeneous impact of risk taking on performance. [Grund et al.2013] analyze whether increased fraction of three-point shots as a risk taking measure has an impact on game outcome. Their conclusion is that the impact is positive only when the point difference is substantial, and no impact when the point difference is relatively small. We can suspect that risk taking might pay off as a last resort given that probability of reverting the result is already very low and taking risk can only benefit the team. [Böheim et al.2016] focus on gender differences in risk taking (which is also defined as fraction of three-point shots) and its potential impact on performance. They observe that when score difference is high, higher fraction of three-point shots might pay off well for female where this impact is negative for male. They also show that there is no significant impact of risk taking on performance for both male and female player if the score difference is small.

There are some other studies from different domains investigating the heterogeneity of risk taking on performance. [Böheim et al.2022], in their causal evidence on gender differences in risk taking study in pole vault, report that male athletes perform worse on risk taking (i.e. attempts directly following a risky pass), while there is not a significant impact of risk taking on performance for female athletes. They state that the reason behind this observation can be excessive risk taking or overconfidence among male athletes. As another interesting study, [Neuberg and Thiem2022] analyze the impact of substituting the goalkeeper for an additional field player (i.e. risk taking behavior) on probability of game wins (i.e. performance) in handball. They observe heterogeneity in ex-ante expectations of the teams and conclude that risk taking is beneficial for underdogs whereas favorites lose significantly more games when they take more risks.

1.3 Data and Descriptive Statistics

Table 1.1: League Descriptions with the Number of Teams and Games

League	Country	Number of Teams	Number of Games per Year	Year
Premier League	England	20	38	2013-2020
Bundesliga	Germany	18	34	2013-2020
LaLiga	Spain	20	38	2013-2020
Ligue 1	France	20	38	2013-2020
Serie A	Italy	20	38	2013-2020
Süper Lig	Turkey	18	34	2013-2020

Our data consists of all matches played between 2013 and 2020 (8 seasons) for men's first divisions of the following 6 countries: Germany, France, Italy, England, Spain and Turkey. We have collected the data from *BeinSports*[1] that publishes the details and results of each game on their website. For player and manager related variables,

[1] www.beinsports.com

we have mainly used two websites: *TransferMarkt*[2] and *Wikipedia*[3]. Tournament rules for each league are similar as all countries above are members of UEFA. The leagues consist of 20 teams (see exceptions in Table 1.1) which are competing in a double round robin tournament.

In these tournaments, a typical match could end with 3 possible results: i. “Win” if the team scores more than the opponent and gains 3 points, ii. “Draw” if the team scores same number of goals as the opponent and gains 1 point, iii. “Loss” if the team scores less than the opponent and gains 0 point. After all matches played between teams, the team which accumulated the most points becomes the champion claiming the highest reward (i.e. highest monetary prize, being qualified to join UEFA Champions League which is the most prestigious and profitable contest in Europe). The tournaments have been designed in a way that not only the winner of the season, but the other teams are also rewarded and incentivized. As it slightly varies depending on the year and country, from 2^{nd} to 6^{th} finished teams also qualify for some prestigious European Cups and claim for a good amount of monetary prize. The teams who accumulated the least points (3 or 4 teams depending on the country and year) relegates to lower division of the same league. Therefore, even a team who does not have any chance to be placed in the top ranks has still an incentive to gain more points. Furthermore, football associations of the countries also put monetary prize for each game irrespective of the importance of the game for each team, eventually creating a continuous incentive scheme for teams.

[2] `www.transfermarkt.com`

[3] `www.wikipedia.org`

Table 1.2: Descriptive Statistics

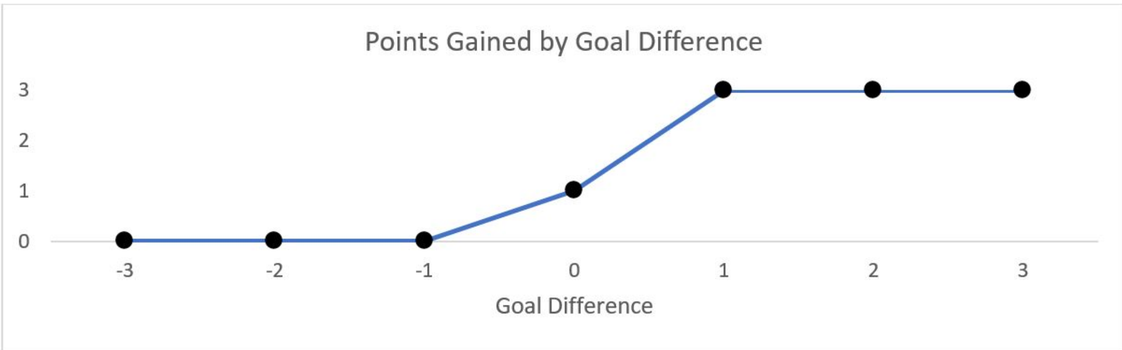
	N	Mean	SD	Min	P25	Median	P75	Max
A. OUTCOMES OF INTEREST:								
Risk Taking	73984	0.25	0.43	0.00	0.00	0.00	1.00	1.00
Point Change	73984	0.06	0.85	-3.00	0.00	0.00	0.00	3.00
B. INTERIM RESULTS:								
Goal Difference	73984	-0.03	1.54	-8.00	-1.00	0.00	1.00	8.00
Red Card	73984	0.05	0.23	0.00	0.00	0.00	0.00	3.00
Red Card Opponent	73984	0.06	0.24	0.00	0.00	0.00	0.00	3.00
Minute of the Game	73984	69.20	10.98	45.00	62.00	71.00	78.00	85.00
Opponent Risk Taking	73984	0.13	0.87	-5.00	0.00	0.00	0.00	5.00
C. MATCH FEATURES:								
Derby	73984	0.04	0.19	0.00	0.00	0.00	0.00	1.00
Home Match	73984	0.50	0.50	0.00	0.00	1.00	1.00	1.00
D. MANAGER CHARACTERISTICS:								
Manager Age	73984	50.17	7.26	28.58	45.07	49.76	54.52	73.06
Manager Foreign	73984	0.29	0.45	0.00	0.00	0.00	1.00	1.00
Manager Tenure	73984	1.57	2.21	0.00	0.35	0.88	2.09	26.55
E. TEAM CHARACTERISTICS:								
Current Ranking	73984	10.22	5.63	1.00	5.00	10.00	15.00	20.00
Match Importance	73984	0.46	0.50	0.00	0.00	0.00	1.00	1.00
Relative Strength	73984	-0.00	0.79	-2.93	-0.36	-0.00	0.36	2.93
F. PAST PERFORMANCE:								
Unexpected Consecutive Game Lost	73984	0.37	0.72	0.00	0.00	0.00	1.00	9.00
Unexpected Consecutive Game Won	73984	0.32	0.64	0.00	0.00	0.00	0.00	6.00
Last 5 Matches Average Point	73984	1.38	0.66	0.00	0.80	1.40	1.80	3.00
Humiliated Defeat	73984	0.05	0.22	0.00	0.00	0.00	0.00	1.00
Big Victory	73984	0.08	0.28	0.00	0.00	0.00	0.00	1.00

Notes: A full list of variables in our data-set can be found in Appendix A.3

After eliminating matches with missing or questionable data[4], we have ended up with 18,439 matches. As there are two teams in one match, total number of matches is 36,878. For each match, we do have minute by minute data such as current score, cards shown to each team, substitutions with incoming and outgoing players, if any. Therefore, the total number of observations at minute level is 3,319,020.

1.4 Results

Figure 1.1: Incentive Structure



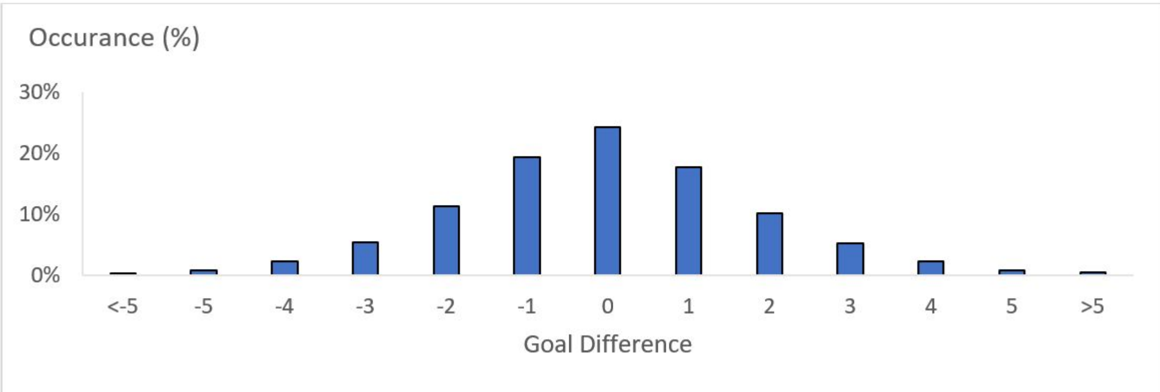
We first want to start by explaining some important concepts heavily used in our analysis. First of all, *"Goal Difference"* is defined as the difference between goals scored and conceded at any point in time during the game. Goal difference is the main feedback given to the manager as if the match ends without any change and the outcome of the match will be based on the current goal difference. As it can be seen in Figure 1.1, teams do not gain any point if they scored less than the team they conceded. For all the leagues considered, the number of points accumulated during the season is the main success factor. In case two teams have the same number of total points, then the points in the match played between those two teams are assessed and the better team takes the lead. Also, if in-between matches are resulted

[4]France Ligue 1 had ended early at week 28 due to COVID-19 situation in 2019-2020 season, we also have omitted some observations from seasons 2013-2014 and 2014-2015 as minutely data was not available from the sources we have used.

in the same point, the leader is determined according to the total goal difference throughout the season. Therefore, goal difference will be an important variable for our analysis.

Figure 1.2 shows that match results in terms of goal difference almost have a normal distribution where 24 percent of the games resulted with a draw. Second most observed results are having only one goal difference with 19% and 18% respectively.

Figure 1.2: Distribution of Goal Difference



Another term we will frequently use throughout our analysis is *"risk taking"*. Risk taking behavior can be observed in different ways in football. Starting team formation is decided by the manager and number of forward, midfield and defender players can be used as an indicator of risk taking behavior [Bucciol et al.2019]. Manager can also change the formation of the players during the game by either changing the position of the player or changing the formation by substituting players. The former is hard to observe in the data since it is not usually revealed by the match reports; therefore, the latter is heavily used in the literature as a revelation of risk taking behavior in football [Bartling et al.2015, Grund and Gürtler*2005, Schneemann2017b]. Since we do observe minute by minute substitution decisions taken by the managers in our data, we use substitutions as an expression of risk taking behavior to analyze the consequences of such decisions.

We note here that each substitution is characterized in terms of the outgoing and incoming players' position on the pitch in our analysis. If the positions are the same, then current formation of the team is not changing in terms of riskiness and we label it as "neutral" risk behavior. If a midfielder or forward player goes in instead of a defender, or a forward player goes in instead of a midfielder, then the team formation has been shifted toward offense; therefore, we label such substitution as "risk taking" behavior. Any substitution which is opposite of so-called risk taking behavior is regarded as "risk reducing" (see Table 1.3). In the next section, we will investigate factors behind risk seeking behavior in football and examine payoff-related effects of these behaviours.

Table 1.3: Possible Substitutions and Their Risk Taking Definitions

Substitution types		Incoming Player Position		
		Defender	Midfielder	Forward
Outgoing Player Position	Defender	Neutral	Risk taking	Risk taking
	Midfielder	Risk reducing	Neutral	Risk taking
	Forward	Risk reducing	Risk reducing	Neutral

Notes: Goalkeeper related substitutions are not considered here, given that it cannot be a strategic action for the manager (it is possible a player becomes goalkeeper when the goalkeeper has a red card and all substitutions have been already done, therefore it might be important whom to replace the goalkeeper, however, this is very rare with less than 0.01% of the cases).

Lastly, since one of our main assumption is that risk taking substitutions are mainly made to change the score on behalf of the trialing teams, we must properly define what success means following the substitution decision. Here, we have the following definition that is aligned with [Xu et al.2019]. Point change is defined as the difference between the point the team is supposed to get if the match ends with the goal difference at substitution time and actual point achieved at the end of the match. We can formulate it as:

$Point_{ijt} = 3 \quad \text{if } GoalDiff_{ijt} > 0$

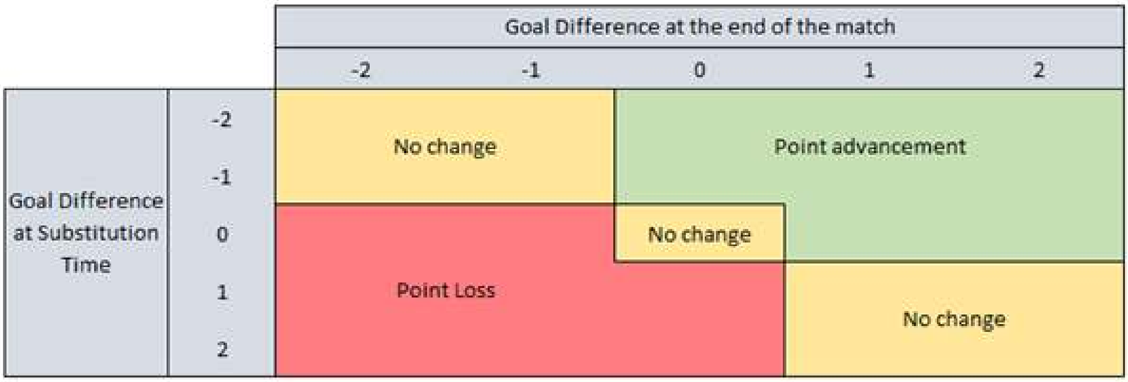
$Point_{ijt} = 1 \quad \text{if } GoalDiff_{ijt} = 0$

$Point_{ijt} = 0 \quad \text{if } GoalDiff_{ijt} < 0$

$PointChange_{ijt} = \triangle Point_{ijt} = Point_{ij,et} - Point_{ijt}$

where i = Match ID, j = Team ID, et = time at the end of the match, t = time at the substitution. All combinations of point change and substitution strategies are summarized in Table 1.4.

Table 1.4: Definition of Point Advancement



Notes: Goal differences of <-3 and >3 are not shown for simplicity. Score and point advancement rules are the same for those intervals as well. Green area can be regarded as positive point change, yellow as zero change and red area is negative point change.

1.4.1 Determinants of Risk Taking Behavior

In this part of the chapter, we will be analyzing the factors behind risk seeking behaviours in football. Our dependent variable is a binary indicator of making a risky substitution where i = Match ID, j = Team ID and t = minute of the game. It takes 1 where the manager substitutes a defender player with a forward or midfielder one, or a midfielder with a forward one. More formally, $RiskTaking$, is defined as whether there is a risky substitution has been made:

$$RiskTaking_{ijt} = 1 \quad \text{if Offensiveness}_{ij,p^{in}} - \text{Offensiveness}_{ij,p^{out}} > 0$$

$$RiskTaking_{ijt} = 0 \quad \text{if Offensiveness}_{ijt,p^{in}} - \text{Offensiveness}_{ijt,p^{out}} \leq 0$$

where $\text{Offensiveness}_{ijt,p^{in}}$ is defined for the incoming player who are substituted at time t for team i in match i and becomes 2 for Forward, 1 for Midfielder, 0 for Defender. Likewise, Where $\text{Offensiveness}_{ijt,p^{out}}$ is defined for the outgoing player who are substituted at time t for team i in match i and becomes 2 for forward, 1 for midfielder, 0 for defender.

A potential factor that could affect risk taking in football is the current score (i.e. goal difference) at the given minute of the game. We have already mentioned that goal difference is strongly related to payoffs; therefore, managers might want to adjust their strategies according to the current goal difference. Table 1.5 presents distribution of risk level of substitutions by current goal difference. We see that most of the substitutions are risk taking when the team is behind in score, possibly hoping to change the score in the subsequent minutes of the game. Likewise, teams leading the match are tended to reduce risk by putting more defensive players on the pitch instead of offensive ones. In case of draw (i.e. goal difference equals to zero), some managers still prefer taking risk, as there is an improvement opportunity in the score.

Table 1.5: Frequency of Substitution Types by Current Goal Difference

		Risk Decision		
		Risk reducing	Neutral	Risk Taking
Goal Difference	<-1	16%	52%	32%
	-1	12%	50%	38%
	0	19%	59%	22%
	1	27%	59%	14%
	>1	25%	62%	13%
	All	19%	56%	25%

Notes: All scores with a goal difference less than 1 is grouped under "<1". Similar grouping is done for ">1"

We note here that goal difference during the game is not evenly distributed (more observations towards goal difference are zero – see Figure 1.3). We will be focusing on those instances that are more frequently observed. Therefore, a goal difference of -2, -1 and 0 will be the main analysis set we'll be exploring. When goal difference is -3 or worse, it's almost impossible to reverse the score, hence it will be discarded from the analysis. Likewise, for positive goal difference instances, as it means manager is happy with the current score, we don't see any incentive to take additional risk to change the score. We still see in the data that managers take risky substitution when the team is leading might be for different reasons although it is significantly lower in numbers). Firstly, it might be stemmed from the noisy data as labeled positions sometimes can be misleading. For example, a manager can put a defensive midfielder in the game and the player is positioned in the defense instead of midfield, or a midfielder who supports the defensive play in a lesser capacity can be replaced with a forward player who can put pressure on the opponents by running more on the pitch, eventually increasing the defensive capacity of the team. We must admit that this is one of the limitations of the exercise as we can not see exactly how the incoming player will be instructed by the manager to play on the pitch.

Figure 1.3: Risk Taking Probability by Current Goal Difference

Notes: Negative goal difference indicates a losing situation with zero point, zero goal difference indicates a draw with 1 point and positive goal difference indicates a winning situation with 3 points

Table 1.6 presents the marginal effects from the logistic regression where the dependent variable is the risk level of substitution. The model behind our regression is:

$$\Pr(\text{RiskTaking} = 1 \mid \text{X}) = \frac{\exp(\gamma + \zeta X_{ijt} + \epsilon)}{1 + \exp(\gamma + \zeta X_{ijt} + \epsilon)} \tag{1.1}$$

where $RiskTaking$ represents the risk level of substitutions (1 for risk taking, and 0 otherwise) and X_{ijt} stands for our large set of covariates.

Table 1.6: Logit Regression - Predicting Risk Taking

	(1)
Goal Difference	-0.053***
	(0.00)
Red Card	-0.095***
	(0.01)
Red Card Opponent	0.062***
	(0.01)
Minute of the Game	-0.001***
	(0.00)
Relative Strength	0.020***
	(0.00)
Home Match	0.014***
	(0.00)
Derby	-0.004
	(0.00)
Current Ranking	0.001***
	(0.00)
Opponent Risk Taking	-0.012***
	(0.00)
Match Importance	-0.000
	(0.00)
Manager Age	-0.001**
	(0.00)
Manager Foreign	0.009**
	(0.00)
Manager Tenure	0.001
	(0.00)
Past Performance	✓
Time and Country	✓
N	73984

Notes: Estimates are average marginal effects after logit regression. Standard errors are reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01. The outcome variable a dummy variable which equals 1 if the manager makes a risky substitution. Controls include time dummies, country dummies and past performance variables (unexpected consecutive game lost, unexpected consecutive game won, last 5 matches-average point, humiliated defeat and big victory). The full version of the regression results can be found in Appendix A.9.

Our analysis confirms that current goal difference strongly predicts risk seeking substitutions when we control for all the potential correlates of risk taking substitutions in football. When the goal difference is negative, it means that the team is losing and managers would take further risk to revert the results. Our results are consistent with [Bartling et al.2015, Grund and Gürtler*2005, Schneemann2017b].

As an important feedback taken during the match, getting a red card is also a good candidate for being a correlate of risk taking behaviour in football. When a red card is shown to a player, his team has to play the remaining period of the game with one player less. Losing a player also causes an increase in the likelihood of conceding a goal; therefore, we may expect coaches to try to reduce this probability by making more defensive substitutions. On the other hand, in case a red card is shown to the opponent team, coaches may substitute a more defensive player with a more offensive one and attack to their opponent more aggressively to benefit from their weakness. Our regression results confirm the findings of [Schneemann2017b] that managers of teams shown to a red cars are less likely to make risky substitutions, while those whose opponents are shown a red card is more likely to opt for a risky substitution.

Another factor that could potentially affect managerial risk taking in football is the minute of the game. In the first minutes of the game, coaches know that there are a plenty of time and the result of the game will probably change. However, when the remaining time is very limited, coaches may need to have some kind of bold moves that could change the flow of the game and consequently the ending score. Therefore; having less remaining time could lead coaches to make more risky substitutions to change the current score. In our regression, minute of the game coefficient appears to be negative, implying that the managers prefer taking the risky decisions in the early stages of the game. This behavior will be analyzed in Chapter 2 in detail to investigate whether early risky decisions pay off well or not.

The previous literature has documented a positive relationship between ex-ante expectations and risk taking behaviour in football [Bartling et al.2015]. When a team is expected to win the match, then the coach seemingly want to his team to play more aggressively, therefore substitute a more defensive player with a more offensive one. In our case, we will use relative strength as a proxy for ex ante expectations. We define relative strength as difference between the average point of the team and opponent's average point taken over the last three years. Our analysis suggests that coaches with stronger teams tend to take more risks, which is consistent with the previous findings. Unlike [Bartling et al.2015], we do not use pre-match betting odds as ex-ante expectations because betting odds include all factors including relative strength, playing home, importance of the match, etc., and we want to analyze the impacts of these factors separately.

In football, playing at home is usually considered as a considerable advantage in football [Pollard2006]. On the other hand, in the matches played at home, coaches may feel more anxious about reactions of fans and audience, in case of loss; therefore, they might tend to make risky substitutions by putting more aggressive players on the pitch. This hypothesis is supported by our regression analysis: we observe that coaches follow more risky strategies when playing at home. Our conclusion is consistent with [Schneemann2017b].

Current rank has also been included in our regression to test whether the previous performance of the team has an influence on managerial risk taking in football. A positive coefficient suggests that teams ranked at the bottom rankings tend to take more risk, probably to put an effort to create a difference and change the current ranking.

The opponent strategies may also matter for risk taking decisions in football. We have examined opponent risk taking behaviors as another interim information, and test whether the actions of opponent managers have an influence on the decision of risk taking substitutions. We may expect a negative relationship between a team and its opponent's risky behaviours, since if the opponent is taking risk, it may make sense to switch to a safer strategy to stand against attacks. Our analysis confirms our hypothesis.

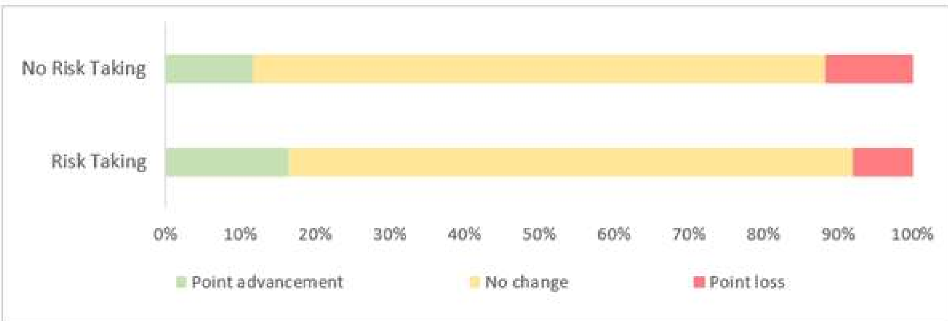
Lastly, we have investigated the relationship between manager characteristics and managerial risk taking in football. Age is an important factor that is found to be positively correlated with risk aversion in the experimental economics literature [Dohmen et al.2011, Rieger et al.2015]. One may expect a similar result for risk taking in football. We do have a statistically significant effect, which confirm the findings of the previous studies.

My study is the first one in football literature analyzing impact of nationality on risk taking behavior. The reason behind the hypothesis is that managers who are coaching a foreign team may feel less public and media pressure due to their actions; therefore, they could be relatively less anxious about payoffs. Hence, we expect to find an effect of being a foreign to a team on risk taking in football. Our analysis indeed shows that foreign managers are slightly more inclined to take risk.

1.4.2 Impact of Risk Taking on Performance

An important question in this chapter is whether risk taking pays off in football. To answer this question, we will examine the impact of risk taking behavior on point change which is the most important performance criteria in these 6 European League Competition.

Figure 1.4: Performance Change from Substitution to Final Whistle



When we look at Figure 1.4, at first glance, risk taking seems to increase probability of point advancement (i.e. positive point change). This is not surprising, as the definition of point advancement is creating a bias towards positive values.

Considering managers are taking more risk when goal difference is negative, they have nothing to lose in terms of points earned because they are already facing the worst outcome which is a point of zero. However, they can take further risk and gain more points. However, in case of goal difference is equal to zero or positive, risk taking can back fire and managers can face a point loss in the match comparing to the time they made the risky decision. What matters for our analysis is the heterogeneity of impact of risk decision on performance rather than whether the risk taking decision overall pays-off or not. Therefore, we will be using interaction terms heavily to investigate this heterogeneity.

$$PointChange_{ijt} = \alpha + \beta X_{ijt} + \theta RiskTaking \times X_{ijt} + \epsilon$$

Our outcome variable is defined as the point change from the time a substitution is made to the end of the match. We have already formally defined both our outcome and treatment variable (ie., point change and risk taking) in the previous section. Our covariates are defined as follows (see Appendix A.3 for the full set of variables and descriptions):

- Interim information that are observed at the time of substitution
  - Examples: current goal difference, red card status for each team etc.
- Match characteristics (ex-ante information) that are observed at the beginning of each game and not changing during the game
  - Examples: weather, home etc.
- Team characteristics (ex-ante information) that are observed at the beginning of each game and each team
  - Examples: team strength, past performance, match importance

- Manager characteristics (ex-ante information) that are observed for each manager. Since a manager can change teams during the season, these variables are defined for each match and team.
  - Examples: nationality, tenure, age

Table 1.7 presents coefficients from the OLS regression where the dependent variable is point change from substitution to the final whistle. We observe that risk taking is positively correlated with point change in football. The effect is significant at the 1 percent level when we control for all the potential correlates of point change in football. Coaches and managers seems to improve their final payoffs by changing the risk level of strategy in the whole sample. However, as explained earlier, an overall positive impact of risk taking on performance is not surprising since the definition of point change is usually zero or positive when risk taking is observed more. We believe that there may exist some subgroups in which risk taking pays more and some others in which risk taking does not pay off at all. Thanks to our large data set, we can search for heterogeneous effects of risk taking on point advancement. To do that, we have included both stand-alone variables and their interactions with the risk taking behavior in our regression.

The first observation in this table is that goal difference and red card seem to be negatively correlated with point change. This is expected as goal difference is a direct indicator of the current match status and a team with red card will perform worse than the opponent. As another expected result, playing at home brings an advantage to teams which is consistent with the previous study of [Pollard2006]. Similarly, coefficient of relative strength also appears to be positive, as the stronger teams are more capable of getting a good outcome.

The main interest of our analysis is interaction effects, since it gives us the clues about the heterogeneous impact of risk taking on performance. Starting with goal difference, risk taking may not pay off when the current score is not satisfactory. When coaches send more aggressive players to the pitch, the likelihood of conceding a goal would increase. A positive coefficient implies that risk taking does back-fire when the result is not satisfactory.

Another factor that might affect the relationship between risk taking and point change is the minute of the game. The time when a manager make a substitution may affect the outcome, as an early substitution can give enough room for team to revert the score on one hand, but it might worsen the result because it can increase the probability of conceding more goals in the remaining time. Our regression results suggest that risky decisions should be taken as early as possible. However, the regression does not tell us exactly when the substitution decision should be made because of the limitations coming with the structure of the regression. This kind of limitations will be addressed in Chapter 2 when we use results of causal machine learning models as the optimal cut-off points in generating the interaction terms.

Another expected result we observe is that when managers with stronger teams take risky decisions, it pays-off well, as the stronger teams are more able to revert the results in their favor. Similar to the minute of the substitution, we also do not know how much difference in team strength should lead the manager to take more risk. Another observation here is that match importance also appears to have a positive effect on the impact of risk taking on performance.

Lastly, manager characteristics may also be an important source of heterogeneity in the impact of risk taking on performance. Our study is the first one investigating impact of manager characteristics on the risk taking behavior in the context of football. However, we do not see a significant relationship between interactions of manager characteristics with risky decisions and outcome. We need to remind to the audience of this paper that Chapter 2 will be used to address these non-significant relationships, we will indeed see that if we use the right interaction terms with the right cut-off points, we will observe that managerial characteristics can also be a source of heterogeneity in the impact of risk taking on performance in football.

Table 1.7: OLS - Predicting Point Change

	(1)
Risk Taking	0.212***
	(0.07)
Goal Difference	-0.145***
	(0.00)
Red Card	-0.196***
	(0.01)
Red Card Opponent	0.147***
	(0.02)
Minute of the Game	-0.002***
	(0.00)
Relative Strength	0.117***
	(0.01)
Home Match	0.121***
	(0.01)
Derby	-0.043***
	(0.01)
Current Ranking	-0.036***
	(0.00)
Opponent Risk Taking	-0.031***
	(0.00)
Match Importance	0.001
	(0.01)
Manager Age	0.001**
	(0.00)
Manager Foreign	-0.040***
	(0.01)
Manager Tenure	-0.006***
	(0.00)
Risk Taking × Goal Difference	0.018***
	(0.01)
Risk Taking × Red Card	0.040
	(0.03)
Risk Taking × Red Card Opponent	0.025
	(0.03)
Risk Taking × Minute of the Game	-0.003***
	(0.00)
Risk Taking × Relative Strength	0.020*
	(0.01)
Risk Taking × Home Match	0.011
	(0.01)
Risk Taking × Derby	-0.008
	(0.02)
Risk Taking × Current Ranking	0.000
	(0.00)
Risk Taking × Opponent Risk Taking	0.000
	(0.01)
Risk Taking × Match Importance	0.020*
	(0.01)
Risk Taking × Manager Age	0.001
	(0.00)
Risk Taking × Manager Foreign	-0.002
	(0.02)
Risk Taking × Manager Tenure	0.002
	(0.00)
Past Performance	✓
Time and Country	✓
R-squared	0.087
Adj R-squared	0.087
N	73984

Notes: Estimates are obtained via Ordinary Least Squares regression. Standard errors are reported in parentheses. * p<0.10, ** p<0.05, *** p<0.01. The dependent variable is point change from the time of substitution to final whistle and takes values between -3 and 3. Controls include time dummies, country dummies and past performance variables (unexpected consecutive game lost, unexpected consecutive game won, last 5 matches-average point, humiliated defeat and big victory). The full version of the regression results can be found in Appendix A.10.

1.5 Discussion and Concluding Remarks

In this chapter, we have analyzed the determinants of risk taking behavior and its heterogeneous impact on performance in football. Our results on the determinants of risk taking behavior are consistent with the literature in general. To begin with, goal difference is the main indicator of risk taking consistent with the previous studies [Bartling et al.2015, Grund and Gürtler*2005, Schneemann2017b]. Similarly, having a red card reduces the probability of risk taking, which is also documented by [Schneemann2017b]. Playing home encourages managers to take more risk, as winning the game matters more in front of fans, which is also shown in the previous studies [Pollard2006, Schneemann2017b]. Managers of stronger teams are also inclined to take more risk which is consistent of findings of [Bartling et al.2015].

Different than the literature, we have used nationality of the managers for the first time as an indicator of risk taking. Our result indeed showed that foreign managers are more inclined to take risk, probably because they feel less pressure to take such decisions. Also, we have used opponent risk taking for the first time in the football literature and not surprisingly, managers are taking less risk when the opponent manager take risk, probably because the opponent manager is taking this decision as the current score is not in their favor, hence playing a safer strategy can help retain the score.

We highlight here that our analysis of understanding determinants of risk taking is merely getting ready for understanding the heterogeneous impact of risk taking on performance. That is why, unlike the other authors cited here, we have not run different regressions to find the factors that predict risk seeking behaviours. Our main goal is to investigate the heterogeneous impact of risk taking on performance. Our results show that risk taking pays-off well when it is taken at the early stages of the game, and taken by stronger teams comparing to the opponents. Also, taking risk for more important matches also usually pays-off well. None of the previous studies we have cited in the literature has investigated heterogeneity in their analysis, their main analyses focus on the overall impact of risk taking on performance. This makes our study unique in the literature; however, the strength of our study is not that we have performed many robust regression analyses, rather, we will use the results of this chapter as a baseline for Chapter 2 which investigates heterogeneous impact of risk taking on performance more carefully.

Chapter 2

MODELING THE HETEROGENEOUS EFFECT OF RISK TAKING BEHAVIOR OF FOOTBALL MANAGERS ON MATCH OUTCOMES WITH CAUSAL MACHINE LEARNING METHODS

2.1 Introduction

In Chapter 1, we had used econometric models to investigate heterogeneous effects of risk taking behaviour on performance in football. In particular, we used interaction of variables to identify heterogeneity of the treatment effect. Among all potential covariates, of which the interaction with risk taking behavior is believed to be significant, only relative strength, minutes of the game, match importance and goal difference in the current score were found significant. Furthermore, we have found no significant relationship between managerial characteristics and the treatment effect of risk taking. We have to consider the limitations of our method and search for more plausible methodologies to evaluate validity of these results and provide any further advices to managers.

One of the main limitations of the econometrics tools we had employed in the previous chapter (i.e. OLS regressions) is that it is very difficult to identify the subgroups where the effect of the treatment on outcome significantly differs, since the regressions and the specified known interactions we had run only provide linear effects. Without identifying those subgroups, it would be difficult to perform any empirical analysis and suggest any decision-making mechanism for managers. For example, minute of the game is a numeric variable and interaction with risk taking behavior was found significant and negative. However, it is not necessarily true that making risky substitution any time t is better than time greater than t. Linear regression tells us that there is an overall linear relationship between minute of the game and effect of risk taking on performance, but this might not be true for specific minutes of the games. To understand the effect at different values of each covariate, one can manually create specific variables by taking only certain values of the variable, however, given the total number of covariates, this becomes a tedious work for the researcher who tries to do multiple hypothesis testing.

Apart from the interpretability of the regression analysis, there are two concerns with the method we have used in Chapter 1. Firstly, in any traditional econometric model, researchers need to know beforehand which variables create heterogeneity and decide on the specification before running any regression. For a subject with well-established literature, researchers can directly propose her hypotheses related to heterogeneous effects since the theory is solid. In our case, however, where there are not many papers analysing effect of risk taking on performance in football (especially considering managerial characteristics as potential heterogeneity variables), it was difficult to find the right variables among more than 30 variables that could possibly contribute to the heterogeneity of the effect of risk taking on performance. The second problem is that traditional econometric methods are ill-suited to the current big data reality, where the number of covariates and interaction terms can outnumber the observations.

These limitations pushed us into searching new methods to get more comprehensive results regarding the heterogeneous effect of risk taking on performance in football. The previous studies in the econometrics literature (see Section 2.2) have attempted to solve similar problems by using causal machine learning techniques. Although the first studies on causal ML techniques have emerged lately around 2014 and the related literature is relatively scant, there are many ongoing and exciting developments made around 2018. In this chapter, we will employ two popular methods in causal ML literature: FLAME, as a recent and alternative matching method proposed by [Wang et al.2021] and Causal Forest method, that is a specific causal machine learning method developed by [Wager and Athey2018]. Our expectation from the causal ML techniques is twofold: Firstly, we would like to identify new variables that are not found significant in contributing to the heterogeneous treatment effect in traditional econometric modeling in Chapter 1 due to the linear-relation restriction. Secondly, we would like the causal ML techniques to identify sub-populations (e.g. treatment effect is substantial when variable X is greater than x) in a data set with 30+ variables and with no prior knowledge about the potential impact of these variables on the heterogeneity of the treatment effect. Once we have these variables and cut-off points at hand, we will employ a final econometric model by including only the variables with given cut-off points that are suggested by the causal ML techniques. If the final econometric model is better in identifying heterogeneity comparing to our model in Chapter 1, we would say that causal ML can be a beneficial toolkit for econometric practitioners in finding the heterogeneous treatment effects. In other words, we will be using Causal Machine Learning techniques to find the right structure of the econometric model to estimate heterogeneity of the treatment effect.

The rest of the paper is structured as follows: Section 2.2 gives a detailed literature review. Section 2.3 describes the methods and results, Section 2.4 compares the models and links the discussion to Chapter 1. Section 2.5 discusses the methods in a synthetically generated data and explains the differences between causal ML models, and Section 2.6 concludes the chapter.

2.2 Literature Review

In this section, we will be summarizing the literature evolving around causal inference starting from econometric methods up until recent causal ML techniques. Since our paper is the first study using causal ML techniques to find heterogeneous treatment effect in sports, we will be mainly focusing on causal ML techniques.

We first start from the definition of causal effect: causal effect is simply defined as the amount by which an outcome variable $[Y]$ is changed due to exposure to a treatment [W]. Assume a binary treatment variable $W \in \{0, 1\}$. Define potential outcomes as Y_i^1 and Y_i^0 – potential outcomes under treatment and no treatment respectively. Then, TE (i.e. treatment effect) for the i^{th} observation is defined as $\triangle \mathrm{Y}_i = Y_i^1 - Y_i^0$. The fundamental problem of causal inference states that only one of Y_i^1 or Y_i^0 can be observed, hence true $\triangle \mathrm{Y}_i$ would never be calculated. Yet, we can still learn some characteristics of ATE under certain assumptions [Neyman1923, Rubin1974].

Average treatment effect (τ) is defined as $\mathbb{E}[Y_i^1 - Y_i^0]$ under the following conditions [Cox1958]:

- *Stable Unit Treatment Value Assumption (SUTVA):* Potential outcome on one unit should be unaffected by the particular assignment of treatments to the other units.

- *Random Assignment of Treatments:* $W_i \perp \{Y_i^1, Y_i^0\}$

Unconfoundedness and the Propensity Score

For any causality related problem, RCTs (i.e randomized controlled trials) provide the best causal inference when they are allowed and feasible. Even in the absence of RCTs, when observational studies are usually used, quality of data is assessed based on how good observations approximate RCTs. RCTs are designed to estimate average effect of a treatment such as a medicine, political intervention and economic decision. Estimation of average treatment effects in RCTs has been discussed around the potential outcomes model proposed by [Neyman1923] and [Rubin1974].

Unconfoundedness discussion arises when we go beyond a single RCT. A typical example is that when we have two RCTs, simple aggregation of data might be misleading (sometimes called *Simpson's paradox*, [Wagner1982]). Problem is that aggregated data would not be RCT as treatment ratio and effectiveness might differ and simple aggregation will give misleading results by simply ignoring these differences. A simple solution is aggregating difference-in-means estimators, which works as long as we can identify finite number of groups/sub-populations. Exact number of groups (i.e. dimensionality of covariate space) does not affect the accuracy; however, when covariate space is continuous, we might need additional assumptions and methods to estimate ATE.

Unconfoundedness assumption implies that $\{Y_i^1, Y_i^0)\} \perp W_i | X_i$. This assumption is difficult to hold in practise as one needs to condition on a continuous random variable. [Rosenbaum and Rubin1983] suggest propensity score to make the assumption more usable.

- Propensity score: $e(x) = \mathbb{P}[W_i = 1 | X_i = x]$
- Note that under unconfoundedness $\Rightarrow \{\mathrm{Y}_i^1, Y_i^0\} \perp W_i | e(x_i)$

Here propensity score solves an important problem. It is enough to control for propensity score (which is one dimensional) instead of multi-dimensional covariates. Then, once we can partition our data into groups with equal propensity scores, those observations will be comparable and average treatment effect can be estimated for each match (propensity stratification is proposed by [Rosenbaum and Rubin1983]).

Alternatively, one can remove confounding bias by creating "pseudo-populations" where treatment and covariates are independent. Idea is to add more weight to observations believed to be underrepresented in the population. When we use propensity score as a denominator for each observation, the method is called *"inverse probability weighting (IPW)"*. Propensity score can be estimated with a non-parametric regression (e.g. logistic regression).[1] It has been shown that the estimated propensity score corrects for local variability in the sampling distribution of the W_i.

$$\hat{\tau}_{IPW}^* = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{e\left(X_i\right)} - \frac{\left(1 - W_i\right) Y_i}{1 - e\left(X_i\right)} \right)$$

[1] For propensity score estimation, logistic regression is the most common method. Other non-parametric methods like boosted CART and generalized boosted models have also been proven to be effective [Lee et al.2010, McCaffrey et al.2004, Setoguchi et al.2008].

It would be beneficial to compare IPW estimator (non-parametric estimation with multiple covariates) with standard OLS linear regression approach (where the parametric model controls for covariates). Regression might be acceptable if the specification has been made correctly.

Augmented IPW

Despite its simplicity and usability, one problem with IPW is that it might underperform a baseline that estimates treatment effects separately for each value of x in X and aggregates them (especially relevant when covariate space is discrete). The question here is whether there are ways of building asymptotically optimal treatment effect estimators under unconfoundedness. There are two ways of estimating the treatment effect. One is by using IPW and second is conditional response surfaces which is defined as $\mu(1)(x) - \mu(0)(x)$ where $\mu(1)(x)$ and $\mu(0)(x)$ are conditional expected values of outcome under the treatment and control respectively. Augmented IPW combine both strategies and benefit from advantages of both. As initially proposed by [Robins et al.1994], augmented IPW (or AIPW) can be charactered as:

$$\hat{\tau}_{AIPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)} \right)$$

AIPW has some advantages over the other methods we had previously mentioned. As you can deduce from the equation above, AIPW is consistent if either $\mu_{(w)}(x)$ is consistent or $\hat{e}(x)$ is consistent. Because of this double robustness, the method is also called "doubly robust estimator". Additionally, it is asymptotically optimal among all non-parametric estimators in a strong sense and can provide confidence intervals.

Matching

We have already discussed how causal inference is adversely affected when treatment and control groups are different. This can be easily avoided in RCTs where the researcher carefully selects the groups, however, both in observational studies and also in experimental studies where full randomization is not possible, one usually needs to correct for this difference between comparison groups.

Matching can be broadly defined as any algorithm aiming to balance the distribution of covariates in treatment and control groups. There are two common ways of applying matching methods. One of these ways is matching observations based solely on covariates and not using outcome variable at all [Reinisch et al.1995, Stuart2010]. This could be beneficial if there is an additional cost of collecting the outcome. The other matching method uses outcome data primarily to reduce bias in the estimation of treatment effect.

[Cochran and Rubin1973], [Rubin1973], and [Rubin and Thomas1996] are the first studies to discuss matching algorithms in the context of estimating a treatment effect. Many potential issues such as how big should be the control reservoir to get good matches, definition of good matches, how to loosely define close-enough match etc., have been widely discussed by [Althauser and Rubin1970]. Also, main steps in implementing matching methods can be found in Appendix B.1.

Matching methods have some advantages over regression methods. First of all, matching methods are not very sensitive to the functional form of X (interactions, polynomials, etc.). Secondly, it is easier to check whether a certain matching algorithm is successful at creating similar matches or not. These methods are also very useful when there is an imbalance between the treatment and control group since they create similar matches for each observational unit in the control and treatment group. Moreover, comparing two similar units in the control and treatment groups are more intuitive and easier to explain.

There are also some disadvantages of matching methods compared to regressions. Regression methods can perform better when the treatment variable is continuous. Regressions provide us with the effect of all variable (not just the treatment) and more importantly, they allow us to estimate interactions of treatment with covariates without conducting any additional analysis.

Regression Discontinuity Design

A good example for regression discontinuity design is from education. If we want to estimate the effect of a scholarship on a student's GPA, we see treatment (i.e giving a scholarship) is not random, as it is usually given to more successful students. Assume that there is a test and threshold where only students exceeding the thresholds will be awarded with scholarship. Then, the idea of regression discontinuity design is that students around the threshold will be similar and giving scholarship according to this threshold will be as good as a random assignment. Then, any jump in GPA for students close to the threshold but some with scholarship and some not, will be shown as treatment effect.

Identification via continuity gives us:

$$\tau_c = \lim_{z \downarrow c} \mathbb{E}\left[Y_i \mid Z_i = z\right] - \lim_{z \uparrow c} \mathbb{E}\left[Y_i \mid Z_i = z\right]$$

Where c is the cutoff point and Z_i is the running variable that is believed to be correlated with the outcome.

One of the main differences of such a design from IPW and other previous RCT related discussions are that here unconfoundedness holds trivially as the treatment is deterministic (exceeding a certain threshold will make it eligible for treatment). In other words, treatment and control observations are different by definition and that is also why one can not apply any matching type of methods here.

Estimation is usually done via local linear regression applied on each side of the boundary. This idea of using RDDs was introduced by [Thistlethwaite and Campbell1960]; however, more identification-based frameworks have been proposed by [Hahn et al.2001].

Methods for Panel Data

We have not discussed a more complex data structure such as longitudinal data with time periods yet. Following a simple sampling model can be defined for estimating treatment effect assuming constant effect:

$$Y_{it} = Y_{it}(0) + W_{it} \times \tau, \text{ for all } i = 1, \ldots, n, t = 1, \ldots, T$$

where τ is interpreted as a constant treatment effect that we aim to estimate. There are two main issues with such specification:

- Treatment effect is constant for all units in all time periods, which might not necessarily be true.
- Outcome at time t is only affected by the treatment at t, which is also restrictive. This assumption does not hold especially in healthcare since the treatment affects the patient's health not only for time t but also afterwards.

Although there are methods relaxing restrictions above, let us focus on a simple two-way model to take care of the time dimensionality for now.

$$Y_{it} = \alpha_i + \beta_t + W_{it}\tau + \varepsilon_{it}, \quad \mathbb{E}[\varepsilon \mid \alpha, \beta, W] = 0$$

Here we assume each time period have a fixed effect and any deviation is due to noise. Although this specification is very restrictive, it might be beneficial especially when t is finite and small (e.g. only two time periods). Then OLS estimator gives us a closed-form solution such as:

$$\hat{\tau} = \frac{1}{|\{i : W_{i2} = 1\}|} \sum_{\{i:W_{i2}=1\}} (Y_{i2} - Y_{i1}) - \frac{1}{|\{i : W_{i2} = 0\}|} \sum_{\{i:W_{i2}=0\}} (Y_{i2} - Y_{i1})$$

This difference in differences estimator is firstly proposed by [Card and Krueger1993] comparing employment outcomes in New Jersey and Pennsylvania where the former increased minimum wages and later did not.

[Abadie et al.2010], and [Arkhangelsky et al.2019] generalized the two-way model by using interactive panel models with synthetic controls. One advantage of synthetic controls is to re-weight the unexposed units.

Instrumental Variables Regression

Endogeneity is a serious problem for econometricians, and we have not addressed endogeneity above yet, as we mainly assumed unconfoundedness. Endogeneity may arise due to i. two-way causality – changes in the dependent variable can also lead to a change in the explanatory variable, ii. omitted variables which are potentially correlated with dependent and independent variables, and iii. non-random measurement errors which is not our primary concern for IV regression.

As a classical example of confoundedness, one can estimate causal effect of price on demand. One main problem of this analysis is that due to dynamic structure of supply-demand equilibrium, price and demand are affecting each other simultaneously. Therefore, one can not simply use OLS regression to understand the effect of prices on demand.

[Angrist et al.2000] estimate a demand problem where treatment is price of fish, and the outcome is demand. They used weather conditions as an instrument to mediate the effect. Intuition is that fish supply is lower under bad weather hence the prices are higher, but there is no direct correlation between the weather condition and fish demand.

By using IV, we have simply two equations:

$$Y_i = \alpha + W_i\tau + \varepsilon_i, \quad \varepsilon_i \perp Z_i$$

$$W_i = Z_i\gamma + \eta_i$$

Here, important to note that Z_i(IV) is uncorrelated with ε_i. Therefore, if one can solve the second equation and predict the values of treatment, and replace in the first equation, then they would have had a consistent estimation. This seems very powerful because IV claims that consistent estimation can be done even without unconfoundedness. There are 3 key assumptions that the IV must satisfy: First of all, IV must be exogenous (i.e. not correlated with the error term). Secondly, IV must be correlated with the treatment variable. Lastly, the effect of IV on the dependent variable must be mediated only by the treatment variable.

In case there are more than one instrumental variable identified, one can simply convert them into a univariate instrument, therefore high-dimensional IV space is not a real problem. Perhaps the most difficult part of IV regressions is to make sure that the instrument is powerful. [Bound et al.1995] showed that this is indeed a serious problem in practise. They put forward an example where IV is pure noise and two stage OLS is no longer providing the true causal effect. In theory, one can also apply non-parametric IV regression. However, it involves solving the integral equations, and one should be careful with regularization, therefore it's not an easy task in practise. [Newey and Powell2003] discussed under what conditions the non-parametric approach provides consistent estimates.

There is a long history of IV regressions starting with [Amemiya1974], [Chamberlain1987] and some others. [Belloni et al.2012] used lasso to estimate the first stage regression. An important note here is that according to [Pearl2009], two-stage regressions with instrumental variables is another form of "structural equations model" which will be covered in the next section.

Structural Equation Modeling

Firstly, we have to say that structural equation modelling (SEM) is not a certain statistical method. It is rather a framework that combines different techniques and methods. It is mainly useful when there is a system of relationships between variables rather than a single dependent variable explained by a set of independent variables. Usually, that multi-faceted complex set of relationships would leave considerable errors but we will see how SEMs are good to make corrections for such error of measurements.

SEM can be thought as a path analysis using latent variables. Remember that latent variables are those that are not directly observable. However, we can still use observable indicators to measure latent variables. Idea is that a measured variable is a combination of true score and measurement error (systematic error + random error). This equation is unidentified since there is only one equation. Therefore, we can not separate true score and error. Then, we can define multiple equations with multiple latent variables.

[Pearl1995] introduced a method, called "do calculus", to solve the multiple equations at once. [Pearl2009] also provided an overview more recently. There are some alternative approaches to representing complex linear equations as directed acyclic graphs [Robins1986, Spirtes et al.1993].[2]

Estimating Treatment Heterogeneity

We have so far discussed only ATE. However, there might be situations where the average treatment effect is close to zero, but the treatment effect is positive or negative for certain sub-populations.

Conditional average treatment effect (CATE) is defined as:

$$\tau(x) = \mathbb{E}\left[Y_i^1 - Y_i^0 \mid X_i = x\right]$$

Important to note that CATE is different than the unit treatment effect in a way that it is still average effect but for certain covariates or sub-populations.

When we estimate the CATE as the difference between these two regression estimates:

$$\hat{\tau}_T(x) = \hat{\mu}_{(1)}(x) - \hat{\mu}_{(0)}(x)$$

This approach is simple and still consistent, yet it comes with certain disadvantages. First of all, when there is an imbalance between the treatment and control groups, regularization used for each regression will differ and CATE function will not be stable. [Künzel et al.2019] provide a good example where CATE function is

[2]See [Imbens2020] and [Pearl and Mackenzie2018] for additional empirical studies.

not constant due to scarcity of treatment observations.

Secondly, validating treatment heterogeneity is not straightforward because of the fundamental problem of causal inference. However, there are still some approaches proposed by several authors. For instance, [Wager2020] discusses whether having a numerical difference between the cross-validated losses is a problem, and prefers the estimator with expected test set error (i.e. mean squared error). On the other hand, [Van der Laan et al.2007] proposes model aggregation via stacking or pruning a complex model of treatment heterogeneity. Another approach heavily used in the literature is using a conditional average treatment effect estimator $\hat{\tau}(x)$ to guide subgroup analysis. Another recent approach for validating treatment heterogeneity is fitting a partially linear model [Athey and Wager2019, Chernozhukov et al.2018b].

To sum up, estimating treatment heterogeneity can be solved with a non-parametric regression, but regularization bias is a real problem to be tackled. [Athey and Imbens2016], [Imai and Ratkovic2013], and [Tian et al.2014] have proposed methods like lasso or regression trees to modify statistical learning tools. Double machine learning in the context of orthogonalization (Neyman-orthogonal) proposed by [Chernozhukov et al.2018a] play a key role in non-parametric causal inference. Moreover, some additional works on this topic have been presented by several authors. [Athey et al.2019] developed causal forest implementation, [Ding et al.2019] estimated treatment heterogeneity in 33 randomized trials under strict randomization inference, [Hahn et al.2020] and [Künzel et al.2019] utilized propensity score, [Nie and Wager2021] provided excess error bounds for heterogeneous treatment effect estimation via non-parametric kernel regression with the R-learner, and [Zhao et al.2017] implemented post-selection inference for treatment heterogeneity using R-lasso.

FLAME

FLAME is an alternative matching method under the potential outcomes framework with a binary treatment for large datasets which uses only discrete covariates [Wang et al.2021]. FLAME (Fast, Large-scale, Almost Matching Exactly) provides us almost-exact matches that treatment and control units are exactly matched on important covariates [Gupta et al.2021].

Let us first briefly mention the FLAME procedure itself. The main task of the algorithm is to determine a sequence of variable (covariate) selection indicators $\left\{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^d\right\}$. Once we have the sequence of variable selection indicators, observations can be easily matched.

The algorithm first initializes the selection indicator to include all covariates. In that way, all possible exact matches are made at the beginning, and then excluded from the leftover data. Then, at each iteration, FLAME determines the selection indicator for that iteration by minimizing the prediction error on the hold-out training set, described by the below equation. We note here that the FLAME algorithm drops only one covariate in each iteration and never uses it again in matching the remaining observations. Therefore, we can say that FLAME algorithm does not guarantee to solve the Full Almost-Matching-Exactly (Full-AME) Problem described by [Wang et al.2021], rather it provides an efficient approximation to the problem.

$$\mathrm{PE}_{\mathcal{F}_{\|\boldsymbol{\theta}\|_0}}(\boldsymbol{\theta}) = \min_{f^{(1)} \in \mathcal{F}_{\|\boldsymbol{\theta}\|_0}} \mathbb{E}\left[\left(f^{(1)}(\mathbf{x} \circ \boldsymbol{\theta}) - y\right)^2 \mid t = 1\right] + \min_{f^{(0)} \in \mathcal{F}_{\|\boldsymbol{\theta}\|_0}} \mathbb{E}\left[\left(f^{(0)}(\mathbf{x} \circ \boldsymbol{\theta}) - y\right)^2 \mid t = 0\right] \tag{2.1}$$

FLAME matches the remaining units exactly on the covariate set determined by the newly found selector and excludes matched units from the remaining data. The procedure continues until either prediction error from removing an additional covariate is very large, or all the data are matched.[3]

There are mainly two benefits of FLAME over other matching algorithms. First, it learns a distance rather than using a pre-specified distance, which is a weighted

[3]The Python code for FLAME is available at `https://www.github.com/almost-matching-exactly/`. An introduction to the project with links to the code is also found at `https://almost-matching-exactly.github.io/`. You can also find a short overview of FLAME in [Gupta et al.2021] or [Wang et al.2021].

Hamming distance for matching based on a hold-out training set. In this way, FLAME drops the most irrelevant covariates and always keeps enough covariates to obtain high quality matches and consequently high quality CATE estimation. Secondly, FLAME is very fast with large datasets since it handles very well both the datasets that has been pre-processed and fits in the memory and the datasets which are too large to fit in memory. In the former, it uses bit vectors, and in the latter, it works through highly optimized built-in SQL group-by operators that can operate directly on the database. The use of database systems allows the matching algorithm run through parallel executions, which provides us with considerable efficiency.

One limitation of FLAME comparing to DAME [Gupta et al.2021], is that FLAME drops the covariates iteratively, therefore not necessarily providing the optimal set of retained covariates. Secondly, FLAME takes only categorical variables as inputs that might create an additional burden on the researcher with additional data engineering and it is not always possible to know how to categorize the variables.

Causal Forest

The causal forest is a method of Generalized Random Forests, developed by [Athey et al.2019] and known as recursive partitioning. To understand the causal forest mechanism, it would be better to recall how random forest [Breiman2001] and a single decision tree differ. A single decision tree tries to maximize homogeneity of outputs in each leaf and also maximizing heterogeneity of outputs between leaves. Since a single decision tree could over-fit easily because of its greedy algorithm, random forest algorithms are used to alleviate over-fitting by ensembling the multiple trees where both the sample and the features are randomized.

A similar relationship can be found with causal trees and causal forests as well. Causal trees are aiming to represent treatment effect heterogeneity and splitting criterion is such that differences of treatment effect heterogeneity of each leaf are maximized. A causal tree can be pruned by cross-validation, but the performance of the tree is evaluated based on treatment effect heterogeneity. Similar to how we move from decision trees to random forests, a causal forest can be constructed by ensembling a number of causal trees where each causal tree is different because of sampling randomization. To create a causal forest from causal trees, one needs to estimate a weighting function, and then estimate the CATE by using those estimated weights to solve a local generalized method of moments (GMM) model [Athey and Imbens2019]. To prevent over-fitting, causal forests use only honest causal trees where half the data is used to estimate the tree structure, and the other half is used to estimate the treatment effect in each leaf, providing the sub-samples with heterogeneous treatment effect [Jacob2021]. The treatment effects are predicted as the difference in the average outcomes between the treated and the control observations of the estimating sub-sample in terminal leaves.

Since causal forests use honest trees, the estimator for the variance of the estimates is asymptotically normal. Therefore, reliable confidence intervals for the estimated parameters can be built [Athey and Wager2021]. One main difference of causal trees is observed when it comes to performance evaluation. Although we usually use mean squared errors for performance evaluation in a typical predictive task (e.g. decision trees), this is not possible in causal forest as we do not observe the actual treatment effects (i.e. fundamental problem of causal inference). However, since the bias disappears asymptotically, the causal forest estimates are consistent and asymptotically Gaussian, suggesting that valid confidence intervals are ensured if we have the estimator for the asymptotic variance.

2.3 Methodology and Results

In this chapter, we use the same data set we previously used in Chapter 1. This time, we will implement causal ML techniques to estimate heterogeneous treatment effects of risk taking on performance. If we succeed to find the variables that contributes to heterogeneity of treatment effect, then we will use these variables to employ a final linear regression to validate importance of using causal ML techniques for econometric model practitioners.

Recall that our data-set consists of all football matches played between 2013 and 2020 (8 seasons) for men's first divisions of the following 6 countries: Germany, France, Italy, England, Spain and Turkey and includes 30+ variables on ex-ante information (e.g. match and team characteristics, manager characteristics etc.) and interim information (e.g. minute of the game, current goal difference etc.). A detailed explanation on these variables and some descriptive statistics has been given in Chapter 1.

Different than Chapter 1 where we have used 73984 observations for regression analysis, we're randomly splitting our data set into two: half of the data (36992 observations) is labeled as $CausalMLData$ and it will be used in the Machine Learning applications in this chapter. The other half, $RegressionData$ is reserved for the final section of this chapter and used for final regression model for validation of the causal ML models. A more detailed discussion on this "honesty" condition will be made later in the final regression section.

$$PointChange_{ijt} = \alpha + \beta X_{ijt} + \gamma RiskTaking \times X_{ijt} + \epsilon$$

Our outcome variable is defined as the point change from the time a substitution is made to the end of the match. We can formulate the outcome variable as:

$Point_{ijt} = 3 \quad \text{if } GoalDiff_{ijt} > 0$

$Point_{ijt} = 1 \quad \text{if } GoalDiff_{ijt} = 0$

$Point_{ijt} = 0 \quad \text{if } GoalDiff_{ijt} < 0$

$PointChange_{ijt} = \triangle Point_{ijt} = Point_{ij,et} - Point_{ijt}$

where i = Match ID, j = Team ID, et = end time, t = substitution time.

Our treatment variable, $RiskTaking$, is defined as whether there is a risky substitution has been made:

$RiskTaking_{ijt} = 1 \quad \text{if Offensiveness}_{ij,p^{in}} - \text{Offensiveness}_{ij,p^{out}} > 0$

$RiskTaking_{ijt} = 0 \quad \text{if Offensiveness}_{ijt,p^{in}} - \text{Offensiveness}_{ijt,p^{out}} \leq 0$

where $\text{Offensiveness}_{ijt,p^{in}}$ is defined for the incoming player who are substituted at time t for team i in match j and becomes 2 for Forward, 1 for Midfielder, 0 for Defender. Likewise, Where $\text{Offensiveness}_{ijt,p^{out}}$ is defined for the outgoing player who are substituted at time t for team j in match i and becomes 2 for forward, 1 for midfielder, 0 for defender.

2.3.1 FLAME

An ideal approach to estimate casual effects in an observational setting is to match each treatment observation to an identical control unit. That is, we find two observations sharing every attribute except for treatment status. If we can find such pairs, then we can estimate how their outcomes differ and what is the effect of treatment on the outcome of interest for this specific group, so identify conditional average treatment effects.

When we work with large data-sets and observe many characteristics of treated and control units, finding identical twins (referred to as exact matches) with the same features becomes an uneasy task. High dimensional covariate spaces make exact matching almost impossible (e.g. in our data set, only 1% of the observations are exactly matched given all 30 variables). To deal with this problem and match observations in large data-sets in an acceptable way, previous studies have proposed different dimension reduction approaches (e.g. propensity score, FLAME etc.).

Using propensity score as matching criteria to define how closely related two observations are, might be beneficial in a high dimensional data-set; however, it also brings about two important issues. Firstly, matching methods preferably should be interpretable, since causal effect estimations conducted after matching need to provide information regarding which observations benefits from treatment most and why one group gets better off with treatment while others do not. Unfortunately, the methods using pre-specified distance metrics including propensity scores do not provide us with a set of covariates which is important for identifying treatment effects. Secondly, the quality of matches matter a lot. We do not want to compare two completely different units; matched pair should be similar enough to each other to estimate heterogeneous treatment effects accurately. There is no doubt that the more variables you have, more distant becomes the two matched points (matched based on propensity score) in terms of other covariates.

Taking these concerns into consideration, FLAME (Fast Large-scale Almost Matching Exactly) seems to be a good alternative to matching algorithms using distance metrics to reduce dimensions of large data-sets. FLAME learns a distance metric for matching using a separate training data set and then computes high quality almost-exact matches for high-dimensional categorical data-sets. Moreover, FLAME is relatively fast compared to other causal ML techniques and provides robust results in large data-sets.

We have modified the parameters of the method determining what covariate should be dropped next. We have tried both Ridge Regression and Decision Tree options to see if there is any difference due to the model selection. No significant change in the order of covariates to be dropped has been observed, hence we will be reporting the Ridge Regression results below.

When we execute FLAME, the algorithm retains four covariates (relative strength, goal difference, derby match and match importance) and drops all the other variables. These four variables are the most relevant ones which are relatively more correlated with our output variable (i.e. point change). Moreover, the most irrelevant covariates which are not significantly correlated with the output, are generally dropped before the slightly correlated variables (see from Table 2.1).

Table 2.1: FLAME - Dropping Order

Variables	Dropping Order
Relative Strength	Kept
Ranking Round	6
Home Match	9
Red Card	8
Red Card Opponent	5
Goal Difference	Kept
Manager Tenure	2
Manager Foreign	4
Manager Age	7
Minute of the Game	1
Opponent Risk Taking	3
Derby	Kept
Match Importance	Kept

FLAME provides unit level treatment effects after calculating difference between average outcome under treatment and control groups for the matched groups. However, it does not give us directly what covariates are the source of heterogeneity of treatment effect on outcome.

2.3.2 Explainable Boosting Machine

A relevant study for us, [Gur Ali2022], has used Generalized Additive Models plus Interactions (GA^2M) to make their black-box causal ML ensemble interpretable. They use the (GA^2M) output to modify the theory driven model with additional terms of heterogeneity and impose sign and monotonicity constraints with their Generalized Linear Model with Qualitative Constraints (GLM_QC) method.

Similar to [Gur Ali2022], We have employed Explainable Boosting Machine (EBM) to explain unit level treatment effects by using set of covariates as potential sources of heterogeneity. EBM is a tree-based boosting Generalized Additive Model (that is originally proposed by [Hastie and Tibshirani1987]) that is called a glassbox model as it predicts outcome with high accuracy while being still classified as explainable model. Underlying GA^{2}M models are proposed by [Lou et al.2013], and having only univariate terms and selected number of two-ways interactions (thanks to their FAST algorithm) makes the model easily interpretable by the researchers. This is the main reason why we have chosen EBM to explain unit level treatment effects. We both want to make a prediction with high accuracy while requiring the algorithm to give us a set of one-way and two-ways interactions that will be used as potential sources of heterogeneity and being used in the final regression model later.

GA^{2}M is implemented as Explainable Boosting Machine (EBM) in Microsoft's Interpret package [Nori et al.2019].[4] Unit level treatment effects that are found by causal ML algorithm is used as target variable and all covariates are used as explanatory variables. Please note that all causal and Explainable AI models we have been using in this chapter (FLAME, Causal Forest, and EBM) have used $CausalMLData$ with 36992 observations.

Key outputs provided by EBM algorithm are 1) Global explanations that explain the entire model behavior and provide list of features that are important to contribute to the model and 2) Local explanations where the individual predictions can be explained by the features. For the purpose of our analysis, we are interested to see the global explanations as the important features will be interpreted as potential sources of heterogeneity.

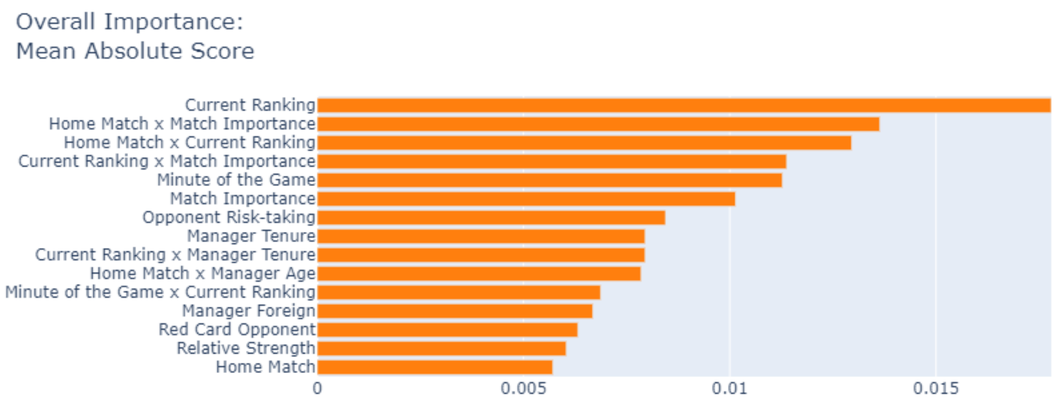
2.3.3 EBM Interpretation of FLAME HTE

When we look at the overall importance chart given by Figure 2.1 , we see that current ranking, home match and match importance are top 3 variables that are either appeared as univariate or two-ways interaction terms. It is surprising that

[4] All details regarding the ExplainableBoostingRegressor can be found in the following website: `https://github.com/interpretml/interpret`

variables such as goal difference, relative strength or minute of the game do not appear as the most important features. This might be explained by the low accuracy ($R^2 = 0.12$) of the EBM algorithm. We should note that FLAME is a matching algorithm with a main focus of pairing treatment and control units and provide reliable ATE calculations. Since HTE (i.e. heterogeneous treatment effect) is not the direct focus of FLAME algorithm, having a low accurate predictions from EBM might be understandable. This leads us to go into the second causal ML algorithm, Causal Forest, where the algorithm directly predicts HTE.

Figure 2.1: Global Explanations by FLAME and EBM



Notes: Mean absolute score is used to rank variables for feature importance. Since EBMs are additive models, the ranking provides what feature has the largest impact on the predictions. For each feature and observation, absolute score of the prediction is taken and average of all observations become the mean absolute score for that given feature. Number of observations equal to 36992 as we used only $CausalMLData$.

Note that FLAME dropped most of the variables and retained only 4 variables for the exact matches. One hypothesis would be that there might be a variable that has heterogeneous treatment effect but being dropped due to the low correlation with output. We will be testing this hypothesis at the end of this chapter once we also see the results of Causal Forest.

2.3.4 Causal Forest

FLAME, as a matching algorithm, does not directly aim to find treatment effect heterogeneity. Instead, matching algorithms focus on finding high quality matches where the treatment and control groups can be compared, eventually leading to estimate a reliable average treatment effect. On the other hand, a causal forest is built from causal trees, where the causal trees learn a low-dimensional representation of treatment effect heterogeneity directly. The splitting criterion optimizes for finding splits associated with treatment effect heterogeneity. Therefore, the end goal is to find leaves where the treatment effect is constant within the leaf but is different from other leaves (i.e. maximum inter-leaf treatment effect). Causal forest requires unconfoundedness (i.e all potential confounders are observed), and similar to Chapter 1, we're also assuming unfoundedness in this chapter.

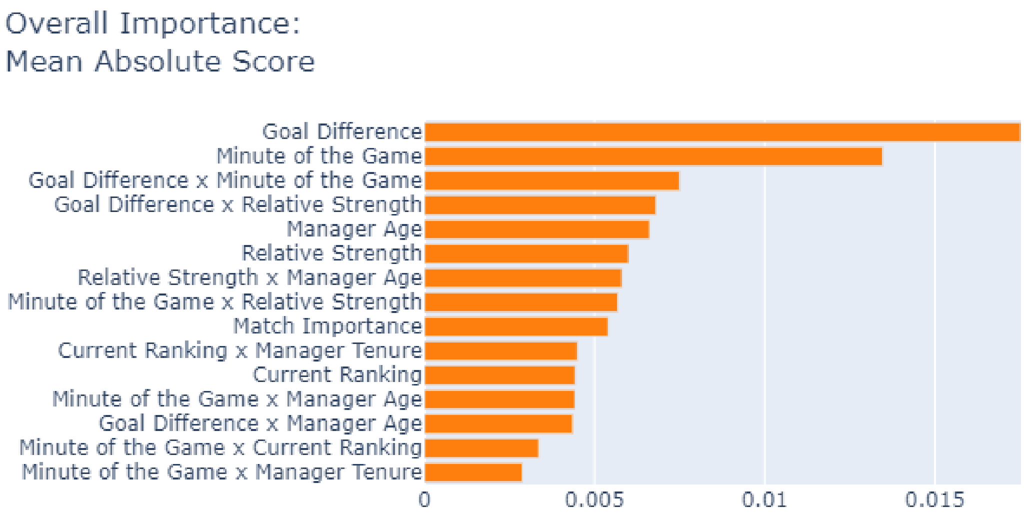
To run Causal Forest model, we have used *CausalForestDML* package from *EconML* which is a python package developed by the Microsoft Research team [Battocchi et al.2019]. Difference of *CausalForestDML* than the original causal forest proposed by [Athey et al.2019] is that they also leverage double machine learning methods based on residualization of the treatment and outcome variables. Double machine learning methods [Chernozhukov et al.2018a] are especially effective when all potential confounders are observed (i.e. unconfoundedness), but high-dimensionality of the covariate space makes classical statistical approaches ill-suited to measure the effect on the treatment and outcome. We have already discussed the potential reasons of high-dimensionality problem in the beginning of this chapter. Double machine learning method alleviates this problem with two prior predictive tasks: 1) Predicting the outcome from the controls, then finding the residuals and 2) Predicting the treatment from the controls and then finding the residuals again. Cross-validated Multitask LASSO has been used for the residualization tasks which is the default option suggested by the package. To implement the Causal Forest, we use the honesty condition and split the data-set into two subsets as one being used to decide where to place the split, and the other subset used to estimate the within-leaf treatment effect [Jacob2021]. In this specification, we have defined Y (output) as point change, T (treatment) as risk taking behaviour (which is a binary indicator of an offensive substitution), W as all covariates that could have determined the choice of treatment, and simultaneously having a direct effect on the outcome, and finally X as a subset of W that we think can be a heterogeneity dimension of the effect of treatment on the outcome. In other words, we have used all the variables used in the regressions of Chapter 1 in W, and we only used those variables that are interacting with risk taking for defining X.[5]

2.3.5 EBM Interpretation of Causal Forest HTE

Causal Forest model does not also directly give us the sub-populations with heterogeneous treatment effects. Therefore, similar to what we did in FLAME, we have also employed EBM to explain unit level treatment effects estimated by Causal Forest. Please note that we have used the same EBM specification (e.g. package used, parameters, number of covariates and observations etc.) with FLAME for comparison. For output interpretation, we also start with global explanations that explain the entire model behavior. Figure 2.2 shows that Goal Difference, Minute of the Game, Relative Strength and Manager Age are the top variables contributing to prediction of HTE. This is more consistent with the findings of Chapter 1 where Goal Difference, Minute of the Game and Relative Strength were also found to be significant in predicting the outcome and HTE. Finding more consistent results with causal forest comparing to FLAME might be because of 1) EBM by Causal forest model is more accurate with $\mathrm{R}^2 = 0.23$ where it was 0.12 for EBM by FLAME and 2) Causal forest is directly aiming to find the HTE unlike FLAME where it just matches the treatment and control groups. Therefore, we are selecting Causal Forest as the main causal machine learning model in this paper and we will use causal forest results in the rest of the chapter.

[5]More details regarding the *CausalForestDML* method can be found in `https://econml.azurewebsites.net/spec/estimation/dml.html`

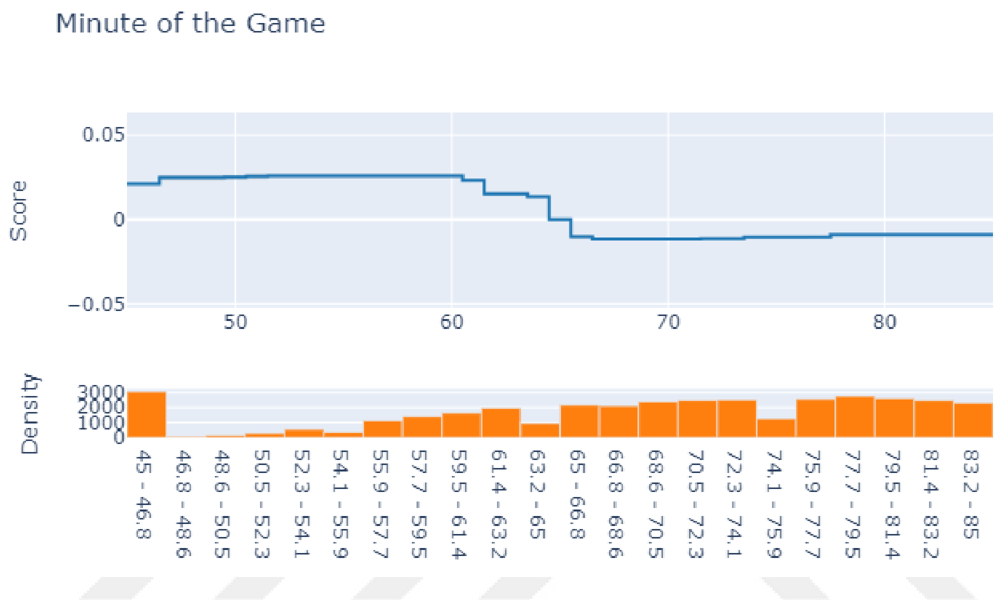
Figure 2.2: Global Explanations by Causal Forest and EBM



Notes: Mean absolute score is used to rank variables for feature importance. Since EBMs are additive models, the ranking provides what feature has the largest impact on the predictions. For each feature and observation, absolute score of the prediction is taken and average of all observations become the mean absolute score for that given feature. Number of observations equal to 36992 as we used only $CausalMLData$.

In addition to global explanations, *show* method of EBM also provides inspections on individual functions f that represents the impact of each individual feature on the target variable. In other words, we are able to see the impact of a feature on treatment effect with all given values of the feature. This gives us a perfect opportunity to transform the variables into binary variables where the cut-off point is determined by the biggest jump in explaining the impact of that feature. For example, we look at Figure 2.3 for understanding how minute of the game is impacting score (which is the treatment effect in EBM). We see that there is a positive yet constant effect before minute 60 whereas the positive impact is gradually decreasing until minute 67 and becomes negative yet constant afterwards. Therefore, the graph suggests us that the interval [60,67] is really critical for us as the impact changes sign during these 7 minutes. As a principle in transforming the variable, we have determined the cut-off point in the biggest jump in the graph, so in this case we are creating a variable of "Minute of the game < 64" as the biggest jump appears in minute 64.

Figure 2.3: EBM Additive Terms for Minute of the Game Variable



Consistent with the literature, we have created transformed variables for the optimal cut-off points suggested by EBM. You can see all individual feature analysis in Appendix B.2. Our transformed variables are 1) "Trailing team" if goal difference is less than 1, 2) "Minute of the game <64", 3) "Strong team" if relative strength is higher than 0.7, 4) "Contender team" if ranking round is less than 5, 5) "Tenured manager" if manager tenure is less than 3 years, 6) "Manager Age > 53" and 7) "Important match" if var is greater than zero.

2.4 Final Regression Specification and Results

It may not be appropriate to list the factors which contribute to heterogeneity by only looking at the results of these different models since causal ML methods like FLAME and Causal Forest are recently developed. Even though they are theoretically valid and reliable, their applications to empirical problems are still in the process of test and validation. Moreover, due to the nature of causal ML algorithms, their performance can not be directly measured (i.e. fundamental problem of causal inference). On the other hand, we would like to exemplify that what we had learned from causal ML methods can be used to improve our initial econometric model. Therefore, in this section, we are going to use a linear regression model with interaction terms similar to what we did Chapter 1, however, this time, we will let Machine Learning models guide us finding the right specification of the regression model.

There has long been debate about the principles and guidelines in determining the right specification of a regression model. Researches often try to collect as many as covariates to control for all confounders in the data. However, regression models are ill-suited to increased dimensionality and more data samples are required to remedy the high dimensionality problem [Greenewald et al.2021, Linden et al.2016]. On the other hand, [Angrist and Pischke2009, Angrist and Pischke2014] have discussed another problem called as "bad control". This problem may arise when the researcher is undeliberately adding new variables to the regression and this can create discrepancies between expected and realized coefficients of the regression. A solid literature and theory behind the topic and development of relevant hypotheses can help the researcher to add variables more carefully. However, in our research, understanding heterogeneous risk taking effects on match results, we may suffer from “bad control" variables as we have too many variables to be tested and the literature is scant in providing recommended set of variables to be added.

Therefore, we have to be very careful in selecting which variables will be interacted with risk, whether we will also create two and three ways interactions, and which variables will be kept as control variables. The problem becomes even harder when you think about the number of potential controls with two ways interactions. For example, for a regression with 20 variables, total number of two-ways interactions is 190 and simply adding 190 extra variables to the regression is definitely something we should be avoiding.

A recently published relevant paper, [Cinelli et al.2021], provides us good guidelines in determining good and bad control variables. The authors discuss 18 different scenarios of causal relationship and provide recommendation whether one should include or exclude the variable for each scenario (they call it “model" to be exact). Based on our understanding of the guidelines given by all the authors, [Angrist and Pischke2009, Angrist and Pischke2014, Cinelli et al.2021, Greenewald et al.2021, Gur Ali2022, Imbens and Rubin2015, Linden et al.2016], we will be following the below guidelines to decide which main and interaction effects will be included in our final regression model:

1. Include those main and interaction effects that are common causes of treatment effect and outcome (i.e. confounders).

2. Include those main effects and interactions that are determinants of the outcome or HTE (i.e. defined as neutral control by [Cinelli et al.2021]).

3. Exclude those main effects and interaction effects that are determinants of the treatment but not the Outcome or HTE (i.e. bad controls). Inclusion of these variables may increase the variance of the estimates, and might make some of them insignificant.

Until now in this chapter, our focus has been predicting heterogeneous risk taking (i.e treatment) effect on point change (i.e. outcome). We have employed FLAME and Causal Forest algorithms to find the unit level treatment effects, then we have used these treatment effects as target variable and found out what predicts treatment effects by employing EBM algorithm on top of FLAME and Causal Forest. Then, we have decided Causal Forest + EBM provides more accurate and insightful estimates; therefore, we saved these results for later use in the regression analysis. For naming convention, let's label our Causal Forest + EBM model as $HTEPredModel$.

Table 2.2: List of Features Found Significant by EBM under Different Models

	$HTEPredModel$	$RiskPredModel$	$OutcomePredModel$ ($RiskTaking = 0$)	$OutcomePredModel$ ($RiskTaking = 1$)
	(1)	(2)	(3)	(4)
Features	Causal Forest + EBM	Random Forest + EBM	Random Forest + EBM	Random Forest + EBM
Goal Difference	✓	✓	✓	✓
Minute of the Game	✓	✓		✓
Manager Age	✓	✓		
Relative Strength	✓	✓	✓	✓
Match Importance	✓			
Current Ranking	✓		✓	✓
Red Card Opponent		✓		
Manager Tenure		✓		✓
Red Card		✓		
Home Match			✓	✓
Opponent Risk-taking		✓		
Goal Difference × Minute of the Game	✓			
Goal Difference × Manager Age	✓			
Goal Difference × Relative Strength	✓	✓	✓	✓
Goal Difference × Current Ranking		✓	✓	✓
Goal Difference × Home Match		✓	✓	
Minute of the Game × Manager Age	✓			
Minute of the Game × Relative Strength	✓			
Minute of the Game × Current Ranking	✓			✓
Minute of the Game × Manager Tenure	✓			
Manager Age × Relative Strength	✓			
Relative Strength × Current Ranking			✓	✓
Relative Strength × Home Match			✓	
Current Ranking × Manager Tenure	✓			
All other main effects & interactions				

Before we go into final regression specification, we would like to add three Machine Learning models to be used in the regression specification. First two models (labeled as $OutcomePredModel$) are aiming to predict outcome (i.e. $PointChange$) separetely for risk taking and non-risk taking observations by using the same covariate set we have used in $HTEPredModel$. For such purposes, we have used $RandomForestRegressor$ from sklearn library in Python. As hyperparameters, we have selected number of estimators as 100, maximum depthness as "None", minimum number split as 2 and minimum number of samples in each leaf as 50. Then, we have used the predictions of Random Forest as target variable and employed EBM to find out what features are contributing to predicting the outcome. For the second supplementary model (labeled as $RiskPredModel$), we have used $RandomForestClassifier$ from sklearn library in Python with the same hyperparameters we have used in $OutcomePredModel$ to predict risk taking behavior with the same set of covariates. Similarly, we have used risk predictions as target variable in EBM and found out list of features contributing to risk prediction. You can find a summary table below for comparison of results of all machine learning models we have used so far (see Table 2.2). We will be discussing the usage of these results in the next section.

When we translate above guidelines for our regression analysis, we have decided to 1) include a variable if it is a determinant in either ("$RiskPredModel$" and "$HTEPredModel$") or ("$RiskPredModel$" and "$OutcomePredModel$"), 2) include a variable if it is a determinant of "$OutcomePredModel$" (no need to check other models), 3) exclude variables that become determinant of only "$RiskPredModel$", but not for others. See below Table 2.3 for exact specification of the final regression model.

We should remind the readers that all regression analyses have been done with the other half of the original data ($RegressionData$) with 36992 observations. We do believe that this honesty property will prevent any bias we might have introduced in understanding the model specification and applying the regression model.

Table 2.3: Final Regression Model Specification

Features	(1) Interacting with Risk	(2) Added as Control Variables	(3) Variable Transformation
Goal Difference	✓	✓	Trailing team
Minute of the Game	✓	✓	Minute of the Game ≤ 64
Manager Age	✓	✓	Manager Age > 53
Relative Strength	✓	✓	Stronger Team
Match Importance	✓	✓	Important Match
Current Ranking	✓	✓	Contender Team
Red Card Opponent			
Manager Tenure	✓	✓	Tenured Manager
Red Card		✓	
Home Match		✓	
Opponent Risk-taking		✓	
Goal Difference × Minute of the Game	✓		
Goal Difference × Manager Age	✓		
Goal Difference × Relative Strength	✓	✓	
Goal Difference × Current Ranking		✓	
Goal Difference × Home Match	✓	✓	
Minute of the Game × Manager Age	✓		
Minute of the Game × Relative Strength	✓		
Minute of the Game × Current Ranking	✓		
Minute of the Game × Manager Tenure	✓		
Manager Age × Relative Strength	✓		
Relative Strength × Current Ranking		✓	
Relative Strength × Home Match	✓	✓	
Current Ranking × Manager Tenure	✓		
All other main effects and interactions			

2.4.1 Final Regression Model (Proposed Specification)

Table 2.4: Final Regression Model with Proposed Specification

	(1)
Risk Taking	0.049
	(0.03)
Trailing Team	0.561***
	(0.02)
Minute of the Game <= 64	0.029***
	(0.01)
Manager Age > 53	-0.000
	(0.01)
Stronger Team	0.137***
	(0.03)
Important Match	-0.037***
	(0.01)
Contender Team	0.229***
	(0.02)
Tenured Manager	-0.003
	(0.01)
Home Match	0.139***
	(0.02)
Trailing Team × Stronger Team	0.144***
	(0.03)
Trailing Team × Contender Team	0.122***
	(0.02)
Trailing Team × Home Match	0.019
	(0.02)
Stronger Team × Contender Team	-0.039
	(0.03)
Stronger Team × Home Match	-0.025
	(0.03)
Trailing Team × Risk Taking	-0.039
	(0.03)
Minute of the Game <= 64 × Risk Taking	-0.288***
	(0.05)
Manager Age > 53 × Risk Taking	-0.000
	(0.05)
Stronger Team × Risk Taking	0.019
	(0.07)
Important Match × Risk Taking	-0.017
	(0.02)
Contender Team × Risk Taking	-0.079**
	(0.04)
Tenured Manager × Risk Taking	0.044
	(0.04)
Trailing Team × Minute of the Game <= 64 × Risk Taking	0.342***
	(0.05)
Trailing Team × Manager Age > 53 × Risk Taking	0.026
	(0.05)
Trailing Team × Stronger Team × Risk Taking	-0.099
	(0.06)
Trailing Team × Home Match × Risk Taking	-0.019
	(0.02)
Minute of the Game <= 64 × Manager Age > 53 × Risk Taking	0.010
	(0.04)
Minute of the Game <= 64 × Stronger Team × Risk Taking	0.123*
	(0.06)
Minute of the Game <= 64 × Contender Team × Risk Taking	0.199***
	(0.05)
Minute of the Game <= 64 × Tenured Manager × Risk Taking	0.005
	(0.05)
Manager Age > 53 × Stronger Team × Risk Taking	0.013
	(0.06)
Stronger Team × Home Match × Risk Taking	0.043
	(0.06)
Contender Team × Tenured Manager × Risk Taking	-0.028
	(0.06)
Time and Country	✓
R-squared	0.13
Adj R-squared	0.129
N	36992

When we include/exclude control variables according to the 3 guidelines we have adopted above, we can see the results in Table 2.4. Our first observation is that risk taking does not have significant impact on performance. This is different than what we had in the regression in Chapter 1 where risk taking was positively impacting the point change. One possible explanation is that when we add all relevant main effects and interactions with other control variables, then the coefficient of risk taking becomes insignificant because the impact is shifted towards controls and interactions with risk. Coefficients of main effects are not surprising. Being a strong team or contender or playing at home brings an advantage and positively affects the outcome. An interesting result is that being an experienced or tenured manager does not have a direct impact on performance.

When we evaluate the interactions with risk taking behavior, we first see that early risk taking (which is characterized as "Minute of the game < 64") is not a good strategy as the coefficient of variable interacted with Risk Taking is negative. However, two-ways interactions with Risk Taking tell us there are conditions when early risk taking might be helpful. If the risk taking teams are trailing which means they need at least one goal to win the match, or if they are contender which means that they are competing for the top positions in the league, then early time risk taking is helpful for the manager. In other words, early risk taking is unnecessary except for trailing and contender teams. Another observation is that risk taking for contender teams is unnecessary unless it is taken in the early stage of the game.

Another difference from the initial regression in Chapter 1 is that we do not observe any strong significant heterogeneous treatment effect for relative strength (i.e. strong team) and goal difference (i.e. trailing team). One possible explanation is that addition of (trailing x strong) interaction without the risk term might have caused (trailing x strong x risk taking) to be insignificant. This means that impact of making any substitution for trailing and strong teams is so impactful that whether the substitution is risky or not, does not bring any additional effect. This is a clear example of how including or excluding specific control variables can change the insights we are extracting from the analysis.

2.4.2 Final Regression Model (No Additional Control)

In the proposed regression model, we have included specific control variables based on a set of guidelines. We have also observed that coefficients of some control variables are big and significant, which might be oppressing some other variables and make them insignificant. In this section, we would like to see the regression results when we removed all additional two-ways control variables suggested by those supplementary machine learning models. In other words, we will be including the main effect variables, and only adding those interaction terms with risk taking behavior to the regression. Selection of risk taking interactions is again done by the Causal Forest + EBM model (i.e. $HTEPredModel$).

When we look at the results in Table 2.5, the observation is that risk taking becomes significant again with a positive impact on point change. Therefore, we can conclude that what made effect of overall risk taking behavior insignificant in the proposed model was not adding new risk interaction terms, it was the new control variables we have introduced by supplementary machine learning models. For the main effects, we do have almost the same results with the proposed model. It means that the magnitude of main effects might be too large to be influenced by including/excluding new control variables.

For one-way and two-ways interactions with risk taking behavior, we observe a richer set of insights that can be extracted from the regression analysis. On top of those insights we have generated in the proposed model, we also now see negative impact of risk taking for trailing and this impact becomes positive when risk is taken before minute 64 for strong and trailing teams. We have to note that richer set of insights does not necessarily mean that this is a better regression. What we truly show is 1) how removing the control variables has impacted the results and 2) we do still see some robust insights (i.e. impact of early minute changes or contender teams) with the changed control variables.

Table 2.5: Final Regression Model (No Additional Control)

	(1)
Risk Taking	0.068**
	(0.03)
Trailing Team	0.623***
	(0.01)
Minute of the Game <= 64	0.030***
	(0.01)
Manager Age > 53	-0.000
	(0.01)
Stronger Team	0.153***
	(0.02)
Important Match	-0.037***
	(0.01)
Contender Team	0.279***
	(0.01)
Tenured Manager	-0.004
	(0.01)
Home Match	0.146***
	(0.01)
Trailing Team × Risk Taking	-0.079**
	(0.03)
Minute of the Game <= 64 × Risk Taking	-0.289***
	(0.05)
Manager Age > 53 × Risk Taking	-0.003
	(0.05)
Stronger Team × Risk Taking	-0.073
	(0.06)
Important Match × Risk Taking	-0.016
	(0.02)
Contender Team × Risk Taking	-0.054
	(0.04)
Tenured Manager × Risk Taking	0.046
	(0.04)
Trailing Team × Minute of the Game <= 64 × Risk Taking	0.341***
	(0.05)
Trailing Team × Manager Age > 53 × Risk Taking	0.029
	(0.05)
Trailing Team × Stronger Team × Risk Taking	0.124**
	(0.06)
Trailing Team × Home Match × Risk Taking	-0.007
	(0.02)
Minute of the Game <= 64 × Manager Age > 53 × Risk Taking	0.010
	(0.04)
Minute of the Game <= 64 × Stronger Team × Risk Taking	0.113*
	(0.06)
Minute of the Game <= 64 × Contender Team × Risk Taking	0.216***
	(0.05)
Minute of the Game <= 64 × Tenured Manager × Risk Taking	0.004
	(0.05)
Manager Age > 53 × Stronger Team × Risk Taking	0.013
	(0.06)
Stronger Team × Home Match × Risk Taking	0.016
	(0.05)
Contender Team × Tenured Manager × Risk Taking	-0.026
	(0.06)
Time and Country	✓
R-squared	0.128
Adj R-squared	0.127
N	36992

2.4.3 Final Regression Model (Variation of Cut-off Points)

By looking at the results of our proposed regression, one could claim that heterogeneity of the risk taking can be also observed for other cut-off points different than the ones found by the causal ML algorithms (e.g. minute of the game being less than 52 instead of 64). If this claim is true and the minute of the game is still a significant heterogeneous factor for other cut-off points like minute 47, 52 or 57, then we would not say that causal ML algorithms indeed help us to find the right cut-off points in transforming the variables for finding the sources of heterogeneity. To check this suspicion, we have run alternative regression analyses as a robustness check. Based on the results given in Table B.1 in Appendix, we can see that our interaction terms lose their significance in most cases when we change cut-off points for goal difference, minute of the game or relative strength. We can conclude that there are indeed only few cut-off points that can provide us with heterogeneous treatment effects, and causal ML algorithms (especially EBM) is good at finding the true ones. We believe that the robustness check confirms that causal forest + EBM model can guide us finding the true source of heterogeneity without running hundreds of different regression combinations to find the best regression when we have tens of potential candidates of variables at hand.

2.5 Synthetic Data and Experiments

We have already compared the results of different causal ML methods in the previous section. However, in this section, we would like to conduct a further analysis to understand why our methods give us different results, which models seem to be more plausible and whether these differences are specific to our data-set. We would like to remind that when we execute FLAME, we see that $EventMinute$, one of the important variables we hypothesized as an important source of heterogeneity, is eliminated during the iterations. Our hypothesis is simple: Risk taking in the last minutes of the match is expected to have different effects on payoffs and scores compared to early strategic changes. Also, we note here that the other causal ML algorithm (i.e. Causal Forest paired with Shapley) report this variable as an important covariate that could result in heterogeneous treatment effects.

One of the most important features of FLAME is that observations are matched on the set of the most relevant covariates which have a strong predictive power over the outcome. In our data, it is shown that $EventMinute$ could predict the treatment effect; therefore, it is an important source of heterogeneity. However, the relationship between $EventMinute$ and effect of treatment on outcome might not be linear or $EventMinute$ is not a direct predictor of the outcome. Hence, FLAME might not retain $EventMinute$ as an important covariate and does not match observations using $EventMinute$. We believe that this could be the reason behind we can not find any heterogeneous treatment effect related to $EventMinute$. That is, our matched pairs are different in terms of $EventMinute$.

We go further and examine why FLAME drops this variable in the early stages. We find that the objective function which is minimized while choosing relevant covariates may have a role in this elimination process. The set of covariates $\boldsymbol{\theta}$ is determined by how well these covariates can predict outcomes. Specifically, the prediction error $\mathrm{PE}_{\mathcal{F}_k}(\boldsymbol{\theta})$ is defined with respect to a class of functions $\mathcal{F}_k := \left\{f : \{0,1\}^k \to \mathbb{R}\right\}$ ($1 \leqslant k \leqslant d$) as:

$$\mathrm{PE}_{\mathcal{F}_{\|\boldsymbol{\theta}\|_0}}(\boldsymbol{\theta}) = \min_{f^{(1)} \in \mathcal{F}_{\|\boldsymbol{\theta}\|_0}} \mathbb{E}\left[\left(f^{(1)}(\mathbf{x} \circ \boldsymbol{\theta}) - y\right)^2 \mid t = 1\right] + \min_{f^{(0)} \in \mathcal{F}_{\|\boldsymbol{\theta}\|_0}} \mathbb{E}\left[\left(f^{(0)}(\mathbf{x} \circ \boldsymbol{\theta}) - y\right)^2 \mid t = 0\right]$$

where the expectation is taken over x and y, when $\|\cdot\|_0$ is the count of nonzero elements of the vector. That is, $\mathrm{PE}_{\mathcal{F}_{\boldsymbol{\theta}_0}}$ is the smallest prediction error we can get on both treatment and control populations using the features specified by $\boldsymbol{\theta}$.

To investigate whether this error in the elimination of covariates is specific to our data, we will some run experiments by creating synthetic data. Our data generating process can be defined as:

$$y = \sum_{i=1}^{5} \alpha_i x_i + T \sum_{i=1}^{5} \beta_i x_i + T \sum_{i=1}^{3} \gamma_i z_i + \epsilon$$

Here, $T \in \{0, 1\})$, $\alpha_i \sim N(10s, 1)$ with $s \sim$ Uniform $\{-1, 1\}, \beta_i \sim N(1.5, 0.15)$ and $\epsilon \sim \mathcal{N}(0, 0.1)$. This contains linear baseline effects and treatment effects.

For each experiment, we first generate 20,000 units (10,000 control and 10,000 treated). Similar to [Wang et al.2021], for relevant covariates, $1 \leqslant i \leqslant 5$, $x_i \sim$ Bernoulli(0.5). We also include 10 irrelevant covariates which are neither correlated with the outcome nor related to treatment effect ($\alpha_i = \beta_i = 0$). These variables are also categorical: For $5 < i \leqslant 15$, $x_i \sim$ Bernoulli(0.2) in the control group and $x_i \sim$ Bernoulli(0.8) in the treatment group.

Different from [Wang et al.2021], we introduce three relevant variables (z_1, z_2, z_3) which has not baseline effect but a strong determinant of treatment effect. $z_1, z_3 \sim$ Categorical$(0.2, K)$ where $K \in \{-2, -1, 0, 1, 2\}$ and $z_2 \sim$ Bernoulli(0.5). While $\gamma_2 \sim N(1.5, 0.15)$, we have the following constant treatment effects (γ_1, γ_3) for z_1 and z_3 respectively:

$$\gamma_1 = \begin{cases} 2 & \text{if } z_1 = -2 \\ 0 & \text{if } z_1 = -1 \\ -2 & \text{if } z_1 = 0 \\ 0 & \text{if } z_1 = 1 \\ 2 & \text{if } z_1 = 2 \end{cases}$$

$$\gamma_3 = \begin{cases} 2 & \text{if } z_3 = -2 \\ -1 & \text{if } z_3 = -1 \\ -1 & \text{if } z_3 = 0 \\ -1 & \text{if } z_3 = 1 \\ -1 & \text{if } z_3 = 2 \end{cases}$$

Table 2.6 provides a correlation analysis between synthetically created variables and the outcome for treatment and control groups separately. Since FLAME attempts to minimize the prediction errors (i.e. $\mathrm{PE}_{\mathcal{F}_k}(\boldsymbol{\theta})$) separately for treatment and control groups during the covariate elimination process, we first check the correlation between input and output for each group. As expected, we do observe a high correlation (with a mean of 0.43 after 100 experiments) between $x_1 - x_5$ variables and the outcome for both treatment and control groups. High correlation suggests that these variables are not expected to be dropped by FLAME as they are highly predictive of the outcome. On the other hand, $x_6 - x_{15}$ variables are not found to be correlated with the outcome for neither treatment nor control groups. This is also not surprising given that the intention of creating these variables were to check if irrelevant variables are being dropped by FLAME or not. Although our expectation from FLAME on how to treat x_i variables is clear, what is not obvious is what to expect from $z_1 - z_3$ variables as they were not subject of interest by [Wang et al.2021]. These variables are exceptionally created, with a pre-defined treatment effect (i.e. γ_i) and without any baseline effect on outcome (i.e. $\alpha_i = 0$).

Table 2.6: Mean and Standard Deviation of Correlation after 100 Experiments

	Correlation with Outcome	
	(1)	(2)
	Treatment	Control
A. Variables with both baseline and treatment effect:		
$1 \leqslant i \leqslant 5$	0.43	0.44
	(0.05)	(0.05)
B. Variables with neither baseline nor treatment effect:		
$5 < i \leqslant 15$	0.01	0.01
	(0.001)	(0.001)
C. Variables with only treatment effect:		
z_1	0.003	0.001
	(0.001)	(0.001
z_2	0.03	0.01
	(0.02)	(0.02)
z_3	0.09	0.01
	(0.02)	(0.02)

Notes: This table presents the mean coefficients of Pearson correlation between each covariates and the outcome of interest separately for the treatment and control group. Standard deviations are given in parentheses.

Table 2.7 presents the results of different experiments and reports which covariates FLAME drops and in which order it drops them. We can observe that during almost all executions, “purely irrelevant variables" are successively eliminated before “relevant covariates" which have both baseline and treatment effect. Moreover, our “relevant covariates" are not dropped when we determine our stopping criteria as a specific number of iterations. However, when it comes to the covariates which has only treatment effect, FLAME drops one of our important covariates, which is potential a source of heterogeneity, z_1, before most of the purely irrelevant variables in all executions. On the other hand, z_2 and z_3 at least are eliminated after purely irrelevant variables. The reason behind this result could be the direct consequence of the covariate and treatment effect functions. Binary and L-shape treatment effects create a slight correlation between the outcome and these covariates even if we do not define a proper baseline effects for them in our data generating process.

Our results show that FLAME may not notice the value of covariates predicting not the outcome but the treatment, and eliminate them. We want to highlight here that FLAME perfectly retains the most relevant variables and drops the most irrelevant variables before our special covariates. Matched groups are similar enough to reliably estimate ATE and $CATE$. However, researchers implementing FLAME should be careful about directly using covariates selected by FLAME for matching observations to identify heterogeneous effects. If you have some doubts about having covariates which have only treatment effects, then you may consider implementing other causal ML techniques such as Causal Forest. Our analyses over the synthetic data show that Causal Forest can be more successful at identifying the source of heterogeneity when covariates do not have baseline effects.

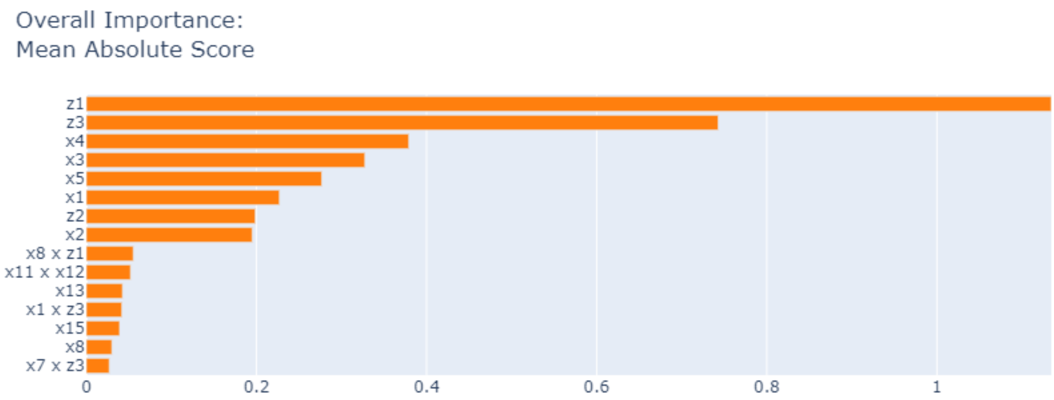
Figure 2.4 and 2.5 show that Causal Forest and FLAME find similar results for the relevant (having both baseline and treatment effects) and irrelevant covariates (having neither baseline nor treatment effects). That is, $x_1 - x_5$ variables are the only x_i variables that appear in the decision tree of both techniques. Among our exceptional variables, $z_1 - z_3$, all three of them are likely to appear in the decision tree of Causal Forest, however, z_1 is never found as a significant variable in the decision tree of FLAME. This leads us to conclude that when it comes to the covariates having only treatment effects, Causal Forest is more successful at detecting their relationship with the treatment.

Table 2.7: Dropping Order of Covariates in FLAME

	N_T= 10K, N_C= 10K		N_T= 10K, N_C= 10K		N_T= 10K, N_C= 30K	
	(1)	(2)	(3)	(4)	(5)	(6)
	Early stop at 10th iterations	No stop until all matched	Early stop at 13th iterations	No stop until all matched	Early stop at 13th iterations	No stop until all matched
A. Variables with both baseline and treatment effect:						
x_1	Kept	Kept	Kept	Kept	Kept	Kept
x_2	Kept	11	Kept	15	Kept	Kept
x_3	Kept	Kept	Kept	17	Kept	Kept
x_4	Kept	12	Kept	14	Kept	Kept
x_5	Kept	13	Kept	16	Kept	Kept
B. Variables with neither baseline nor treatment effect:						
x_6	6	7	5	6	6	6
x_7	7	6	4	7	11	7
x_8	1	1	10	11	7	5
x_9	8	3	9	5	4	9
x_{10}	2	2	1	4	5	2
x_{11}	9	8	8	8	10	10
x_{12}	10	9	11	3	1	8
x_{13}	3	4	2	9	8	11
x_{14}	5	5	6	2	3	3
x_{15}	4	10	7	10	9	1
C. Variables with only treatment effect:						
z_1	Not defined	Not defined	3	1	2	4
z_2	Not defined	Not defined	12	12	12	12
z_3	Not defined	Not defined	13	13	13	Kept*

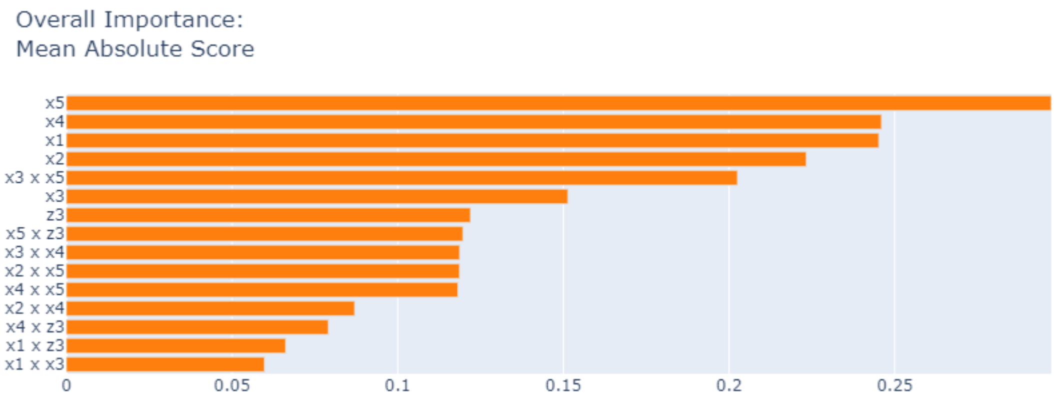
Notes: z_3 in column 6 was kept since all the observations were matched at iteration 12 and the algorithm stopped.

Figure 2.4: Global Explanations by Causal Forest and EBM with Synthetic Data



Notes: Mean absolute score is used to rank variables for feature importance. Since EBMs are additive models, the ranking provides what feature has the largest impact on the predictions. For each feature and observation, absolute score of the prediction is taken and average of all observations become the mean absolute score for that given feature.

Figure 2.5: Global Explanations by FLAME and EBM with Synthetic Data



Notes: Mean absolute score is used to rank variables for feature importance. Since EBMs are additive models, the ranking provides what feature has the largest impact on the predictions. For each feature and observation, absolute score of the prediction is taken and average of all observations become the mean absolute score for that given feature.

An additional analysis has been done to see how good FLAME and Causal Forest are at detecting the true treatment effect at the unit level. When we check the scatter plots in Figure 2.6 and 2.7, we see that both algorithms are consistently predicting well around the true treatment effect where the errors usually occur when the new z_i variables take positive or negative values.

Figure 2.6: Treatment Effect - FLAME

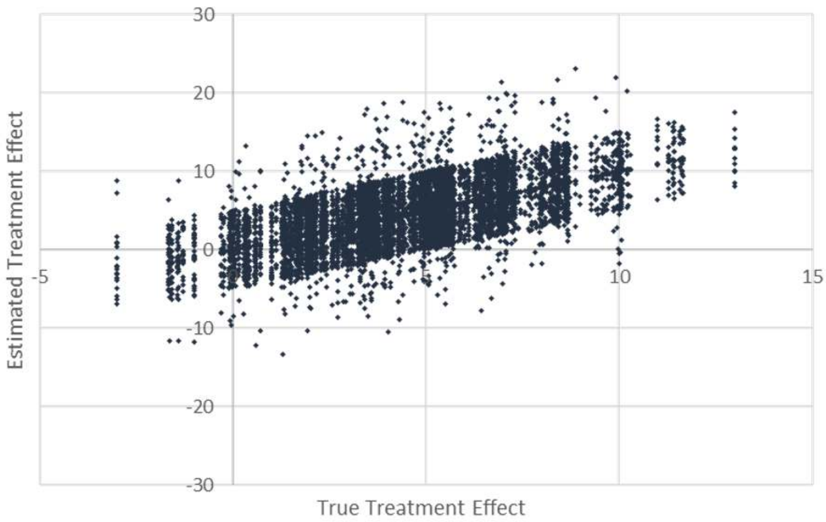
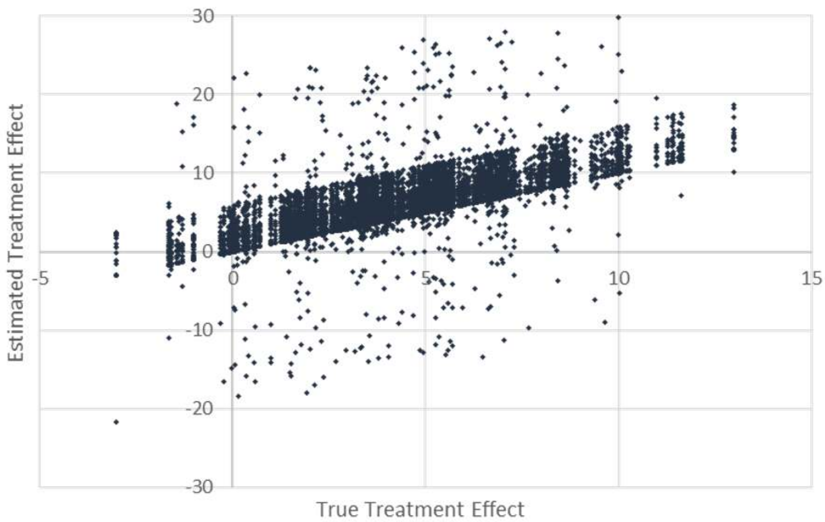


Figure 2.7: Treatment Effect - Causal Forest



2.6 Discussion and Concluding Remarks

Traditional econometric models have been widely used to solve the empirical problems in natural and social sciences for a long time. For the topics which are theoretically well-established and not suitable for collecting large amount of data, it makes sense to use traditional econometric tools. We have examined a research question which has been largely studied in different fields except for sports. It is very important to study the impact of risk taking on performance in football as well as in other sectors like finance and education.

I believe that there is a dilemma in the use of econometric tools. Unconfoundedness assumption forces us to include all the potential covariates into the regression to prevent omitted variable bias. However, in theoretically not well-established topics like football, there are plenty of potential covariates. Unfortunately, there is very narrow literature on the heterogeneous effects of risk taking on performance in football, and none of these studies includes the variables (e.g., manager characteristics, match features) which we believe to have a role in heterogeneity. When we try to include all these variables, the results still do not seem to be reliable since traditional econometric tools are ill-suited to big data reality.

I wanted to point out that causal ML methods could help researchers working with traditional econometric tools. Causal ML methods are strong in figuring out non-parametric relationships compared to traditional econometric methods; therefore, causal ML methods can lead us to look at which variable when we search for heterogeneity. FLAME and Causal Forest have become recently popular in estimating heterogeneous treatment effects. Using these methods, we obtained clues which can help us how to interact existing variables. When we updated our initial regression based on these insights, we found more intuitive and consistent results.

In addition to discussion about contribution of causal ML models, we have also touched an important topic which is what control variables should be included or excluded from the regression analysis. We used machine learning to eliminate bad controls and keep good control variables. Alternative regression models with different control group inclusion rules have shown that the researcher should be very careful in selecting what variable should be treated as bad or good control as it can significantly change the regression results.

Our study illustrates that researchers working with econometric tools should collaborate more with those working on causal ML methods. After the validation of these ML methods in applied topics, the methods could be added to Python or R as a package and the researchers who are not experienced with ML could also implement them. Similarly, researchers using ML methods can also utilize econometric models to make their results more plausible for the general audience in academia.

There are some limitations of our study. Firstly, for all the regression models, we assume unconfoundedness, which is an important limitation in a theoretically not well-established field. Another limitation is that we analyzed the impact of risk taking by linking each substitution decision to the final performance separately. Each change may have an effect on the output at the end and we want to measure it. However, each substitution can also be related to previous and later substitutions. For example, a manager who previously made a risky substitution can be less likely to take another risk in the next substitution. Substituting a midfield player with a forward player is a rare move since there are not many available forward players among substitute players. Furthermore, managers can change the risk level of their strategy without substituting any player. Managers can change how a player behaves on the pitch and how players pass each other, etc. Our data does not provide us with such strategic changes; therefore, we had to ignore them and focus on the changes in risk taking that result from substitutions.

Overall our results suggest that there are many opportunities allowing collaboration of causal ML methods and traditional econometric tools. Our study just benefits from one of these opportunities and try to provide some policy implications in football. We hope that these kind of collaborative studies increase a lot in the future and help business people, policy makers and researchers to obtain more reliable and comprehensive results.

Chapter 3

GRIT AND ACADEMIC RESILIENCE DURING THE COVID-19 PANDEMIC

3.1 Introduction

Grit is an important non-cognitive skill that has been shown to predict achievement outcomes including educational attainment, academic performance, and retention [Duckworth et al.2011, Eskreis-Winkler et al.2014]. In the school context, grit has a significant role on outcomes over and above cognitive skills, and can have large multiplier effects, given that students who give up early have fewer chances to recover. Recognizing this, in many countries, educational policymakers and NGOs implement programs to foster grit in the school environment, especially with "growth mindset" interventions that teach students the malleability of skills through effort and perseverance [Dweck2006, Paunesku et al.2015].

In addition to its association with performance at school, the non-cognitive skill of grit can also be important in predicting resilience in the distance learning context during the Covid-19 pandemic. The challenges brought about by the reduced access to teachers and peers in remote learning have likely impaired the support system for students, changed the interaction of teachers, students and learning content [City et al.2009], and made self-regulated learning [McCombs1986] more central. Motivating oneself when tackling potentially difficult learning tasks can be more challenging with remote learning, in the absence of direct interactions with the teacher and peers. Indeed, recent evidence shows that even in "best-case scenario" countries with favorable conditions, students have experienced learning losses with remote instruction during the pandemic [Engzell et al.2021].

In this paper, we use unique data from an online educational platform in the United Arab Emirates to explore whether the non-cognitive skill of grit predicts performance changes in math and science subjects during the coronavirus period, among a sample of 5th-9th graders. Specifically, we utilize pre-pandemic measures of grit, coming from a survey as well as actual behavior in the digital learning platform, to predict performance changes during the pandemic.

How to measure non-cognitive skills accurately is a major issue that has interdisciplinary relevance. In much of the literature, grit has been measured by the Duckworth grit scale [Duckworth et al.2007]. Measured this way, it has been shown to correlate with a wide range of outcomes such as retention and school performance, although there is also work that has found null effects (e.g. [Zisman and Ganzach2021]). However, survey measures tend to be prone to demand effects or social desirability effects [Bertrand and Mullainathan2001], which may be particularly concerning in an adolescent student sample. Experimental economists have recently sought to construct behavioral measures of grit, using external, dynamic tasks that require real effort (e.g. [Alan et al.2019]). While such measures have been shown to have strong predictive power, implementing and using them at a large scale may be difficult in the regular classroom environment, as they tend to require the use of material rewards and implementation by external researchers. In this respect, using naturally occurring data from a digital learning platform enables us to obtain a unique measure of "behavioral grit" that would be difficult to obtain in the traditional learning environment. Specifically, we are able to see "after-hours" studying and practice by the students, capturing sustained voluntary effort, which allows us to define a measure of "revealed" grit in terms of the response to performance setbacks. Using this measure along with a well-known survey measure in the same sample, we are thus able to make the contribution of comparing the predictive power of survey measures of grit, and the behavioral measure of grit from platform data, both coming from the pre-Covid period, on student performance and behavior during the pandemic.

Our data confirms that the pandemic has led to significant declines in performance in math and science subjects, in line with studies showing that learning has been disrupted during the Covid-19 period. We find that the well-known survey measure of grit [Duckworth and Quinn2009] has limited power in predicting performance in our sample, especially during the coronavirus period, and importantly, it does not predict the declines in performance. We hypothesize that revealed grit may instead be a better predictor of the performance response. We hypothesize that students who had higher grit before the pandemic period register lower declines in academic performance during Covid-19, controlling for baseline diagnostic scores. We indeed find that revealed grit, unlike or stronger than survey-based measures: 1) predicts both pre-Covid and after-Covid performance on the digital platform in math, 2) predicts the change in performance, with grittier students registering lower declines in performance in math and science. Finally, we show that pre-Covid behavioral data in the digital platform contains significant information to explain academic resilience: up to 77% of variance in declines in performance in math and science can be explained using machine learning.

Our findings highlight a novel effect of the well-studied skill of grit—that grit can also predict a positive response to a unique type of shock to the educational environment, coming from the switch to fully remote learning. In this sense, we provide new evidence for why fostering grit in student samples is important for achieving better learning outcomes. Another implication of our results is that digital learning platforms can provide valuable information that can be used for customizing student learning, through AI-based systems that can make use of student performance as well as revealed non-cognitive skills through platform behavior. That is, such platforms, by design, can be used, to not only deliver learning content but also to collect valuable behavioral information that can provide a foundation for customized learning for students with heterogeneous levels of non-cognitive skills.

The rest of the paper is organized as follows. In Section 2, we put forward the setup and research question, and explain our grit measure. Section 3 presents results, and Section 4 concludes.

3.2 Setup, Data and Research Questions

Our data come from a widely used digital learning platform in the United Arab Emirates.[1] In the pre-pandemic period, the platform was used for blended learning in K-12 public school classrooms, delivering core curriculum. After the pandemic, with the switch to complete distance learning, the platform became the sole environment for teaching and learning in the country, in conjunction with communication tools such as *Zoom*. The data used in the paper is partly coming from diagnostic performance tests and a survey conducted by Alef Education on a sample of middle school students in the United Arab Emirates, as part of their research and evaluation activities.[2] The data-set we use in the paper comes from a subset of students, for whom survey measures of grit and performance data from out-of-platform diagnostic tests before the pandemic period are available, in addition to behavioral data coming from the digital platform.

3.2.1 The Behavioral Grit Measure

There are several types of activities for students to engage in on the platform. For each subject (math and science), students can watch instructional videos, study digital course materials, and take online assessments to check their understanding of the material. The platform records every activity, from when the student started the activity to when it ended, how many assessments the student attempted, as well as all other login/logout attempts etc. Here, we define two important concepts

[1] The data are provided by Alef Education, a global K-12 education technology company that provides students and teachers with core curriculum and learning analytics through blended learning settings. The platform also enables students to apply and transfer their newly-acquired skills, and practice lessons for further skill mastery, with the aim of creating an effective instructional and classroom model.

[2] Consent from schools to implement the diagnostic test and survey, and from the Department of Education and Knowledge (ADEK) on working with the collected data for the research were obtained. All data were provided to the authors as secondary, anonymized data.

that will be used in the current paper for analysis. We define a student's "platform performance" as the average score of all assessments taken by the student within a specific period of time. These scores, and hence their averages, range from 0 to 100. We define performance metrics as above for each student and subject (math and science) in the pre/post-Covid periods separately. We also calculate the differences in score from pre- to post-Covid per subject, in order to see what might have impacted the change in performance.

The second variable we use is the number of attempts for a test made by each student. The intensity of engagement with the digital platform during school hours is usually determined by the teacher. For this reason, we focus on out-of-school hours, where it is completely up to the student to use the platform for learning or practice. Given that out-of-school hours activity is voluntary, students differ in how many times they attempt an assessment task. The number of times a student approaches an assessment task is defined as the "attempts", which we use to define gritty behavior.[3]

There are many ways to conceptualize how digital learning data can be used to measure grit. Our grit measure is based on the student's response to a performance that deviates from her earlier average performance. We interpret this fall as a "setback", and the student's positive response to this setback (higher effort exerted out of school hours following the setback) as "grit". Specifically, we define our main grit variable as follows. For each student and for every day in our sample ("Day t"), we calculate the difference between the student's average performance in the past week (during the days $t-7$ to $t-1$) and performance on Day t. If today's score is sizably lower than last week's average score, that is, if there is an at least 5-point score decline from the previous week, we define this incidence as a setback where a behavioral grit response may be observed.[4] We define the gritty response as an

[3]In calculating the variables for the pre- and post-Covid periods, there are several assumptions we make, and related definitions, which are discussed in the Appendix.

[4]If performance does not decline sizably or increases on a given day, the grit variable for that day is not defined. Also, since there is little room for performance improvement for very high

increase in effort, which we capture by the number of attempts the student made in their formative tests on the platform during out-of-school hours following the setback.[5] Specifically, for each setback event, we calculate the difference between the number of out-of-school attempts on the following day ($t+1$), and compare this with the previous week's average attempt ($t-7$ to $t-1$). The behavioral grit variable takes the value of 1 if there is an increase in attempts, and 0 otherwise. Once we have the grit measure for a given day t, we take an average over all days in our sample, for both pre- and during-Covid periods.[6] Thus, we have, for each student, what percentage of the time out-of-school attempts increased after a sizable score decline. We interpret this as a measure of behavioral grit, which we calculate separately for math and science subjects. We then compare our behavioral grit measure with the well-known survey grit measure, based on [Duckworth and Quinn2009]'s "short grit scale".[7]

3.2.2 Data

We use data coming from 1920 students, in 17 schools and 229 classes, out of which 1279 are female and 641 are male. [8] The students are middle- and high-schoolers,

performing students (whose daily score average is consistently higher than 90), these students are omitted from the analysis.

[5]We do not include weekends in the calculations as the number of attempts are significantly lower. When Thursday is day t, then Sunday becomes $t+1$.

[6]In our sample, we have 56 days of observations for the pre-Covid and 84 days of observations for the during-Covid period. Since we compare with an average from the previous week, we omit first-week observations both from pre- and during-Covid periods. Note that before the start of the "Covid period" in our data, there was a 16-day break, where average activity on the platform is 90% less than both the pre- and post-Covid periods. Therefore, we have 49 effective days for pre-Covid and 77 days for during-Covid.

[7]The questions are given in the Appendix.

[8]The sample we use in this paper reflects the set of students, out of a representative sample of all Abu Dhabi public school students, who completed both the diagnostic test (mandatory) and the grit survey (optional). There may be selection in the sample in the sense that the lower tail in terms of motivation have likely been left out, but our results on the effect of grit may be considered as conservative in this sample.

and the data include five grades: grades 5 to 9.[9] We have access to objective scores from a diagnostic test on math and science as well as a survey, which were all run before the pandemic, in Fall 2019. In addition, we have platform data from the period January 10 to March 5, 2020 (8 weeks), which we define as the pre-Covid period. The post-Covid period goes from 22 March, 2020 to June 13, 2020 (12 weeks).[10]

In additional analysis of revealed preference data on non-cognitive skills, we implement an elastic regression using 400 variables from the digital platform. These variables come from "click data" recorded on the platform, and include a detailed set of observables about students' platform activity, such as accessing different types of instructional content for each subject (e.g. videos, reading material, exercises, tests) as well as the timing and duration of these activities. The machine-learning model reduces the size of coefficients to zero when they do not explain much variance, allowing the consideration of a wide range of potential explanatory variables. To reduce risk of confusion, we use principal components analysis (PCA), a technique for reducing dimensionality of data-sets by creating new uncorrelated variables that successively maximize variance and minimize information loss. We use PCA on all of the features from the digital platform. We then report the elastic regression including only the variables selected by elastic regression.

3.3 Analysis and Results

3.3.1 Performance Changes and Covid-19

As expected, the Covid-19 period has led to an overall decline in performance, in both math and science subjects. Table 3.1 shows that in math, there is an about 11 points decline in the average score on the platform, and a similar decline of 10 points

[9]We have 391, 262, 488, 600 and 179 students in grade 5 to 9, respectively. Note that during our data period, there are only two students changing school, and these have been omitted from the analysis. In case of class changes, we take the last class that the student was enrolled in in the academic year as the relevant class.

[10]March 22 is the day when anti-Covid measures were taken in the UAE and home-schooling officially started. We omit the March 5 to 22, 2020 period, as the schools were closed and there were very limited attempts on the platform.

in science, for both genders. These declines are significant at the 1% level ($p < .0001$ for math and science in t-tests).

3.3.2 Grit Measures and School Performance: Pre- and Post-Covid

Table 3.2 shows that there is a strong correlation between behavioral grit and student performance measured in their overall platform scores in math, controlling for baseline diagnostic test scores, which can be considered as an objective baseline performance. One standard deviation of behavioral grit is associated with 4.6 points higher math scores prior to Covid relative to an average of roughly 50, which is equivalent to 10% relative to the mean. The association is significant ($p < 0.05$). For science, grit is not strongly associated with test scores prior to Covid.[11] Girls in this sample perform 20 points higher on math tests, and 28 percent higher on science tests ($p < 0.05$). The respective diagnostic score, as expected, is also positively associated with platform performance ($p < 0.01$ for both subjects). After Covid commenced, grit is associated with both math and science test scores ($p < 0.01$) at a substantial 15 points in both subjects. Of note, the gender gap shrinks by about half and is no longer statistically significant at conventional levels.

In comparison, Table 3.3 shows that survey grit is only associated with one of these four measures of student performance–0.8 points in science prior to Covid ($p < 0.05$). Notably, the association with gender is no longer statistically significant in comparison to Table 3.1. This indicates that there may be strong gender differences in self-reported grit. Moreover, science performance may be picking up certain characteristics of students that are associated with self-reported grit, however this association is small and disappears as a significant factor in the post-period.

[11] Grit's stronger association with math is expected, given that students tend to find math subjects particularly difficult, and consistently with this, grit has been shown to have most relevance in math performance (e.g. [Alan et al.2019]).

3.3.3 Grit Measures and Change in Performance Post-Covid

Next, we explore non-parametrically the relationship between behavioral grit and academic resilience (change in overall test scores, as used above, during Covid-19). A significant predictive association for behavioral grit can be seen in binned scatter plots in Figures 3.1 (math) and 3.2 (science), which condense the information from a scatter plot by partitioning the x-axis into bins and calculating the mean of y within each bin.

Table 3.4 shows, in regression form, that surveyed grit is essentially not correlated or slightly negatively correlated with resilience during Covid-19, measured as the performance change, defined as overall test scores (as above) after Covid minus overall test scores pre-Covid. In contrast, behavioral grit change is associated with 0.1 point greater resilience for math and science, respectively.

The lack of similar predictive associations for survey grit can also be seen in Figures 3.3 and 3.4, which present the non-parametric relationship between survey grit and academic resilience in math and science test scores, respectively.

We should note that there may be other ways to define "behavioral grit" using such data. Our results are largely robust to a set of variations in the assumptions we make in the definition; for example, if we look at effort not only the day after, but two days after the setback, or if we define "setback" to be only when the student registers a "bottom 20%" performance based on his/her historic performance distribution.[12]

To examine total variance explained by behavioral grit measured in the digital platform, we use an elastic regression model in which we include all clicks recorded by the platform, pre-processed into 400 categories. The detailed set of observables we have for this test include students accessing different types of instructional content on the platform for each subject (e.g. watching a video, accessing course-related reading material, doing exercises, starting/ending tests to check understanding), number of times each content is accessed, and the time spent on the platform. To aid interpretability, we first use principal components analysis on these other

[12]Results available upon request.

variables. In Table 3.5, we report the multivariate regression including only the variables selected by elastic regression. The results indicate that 77% of variance in resilience is explained by the behavioral data measured in the platform. The data, while inadequate to definitively infer causality, thus show the potential for high-dimensional revealed preference measurement of non-cognitive skills. Note that before the elastic regression model, we orthogonalize each explanatory variable to behavioral grit measure by projecting each successive variable on the behavioral grit variable and taking residuals. Then the elastic regression model is run by using the uncorrelated variables (i.e. residuals).

3.4 Discussion

The coronavirus pandemic has brought a unique set of challenges for education. The learning loss incurred because of the pandemic by students in grades 1 to 12 is estimated to translate to about 3 percent lower income over the entire lifetime [Hanushek and Woessmann2020]. In this paper, we show that an individual non-cognitive skill, grit, can predict how large of a performance loss students will experience due to the pandemic.

Research on the effects of Covid-19 has shown that personality traits can predict behavioral responses such as hoarding or compliance with guidelines [Zettler et al.2022], as well as with coping responses to the pandemic [Volk et al.2021]. [Johnston et al.2021] show that Covid-19 has led to declines in mental health, and this decline is not predicted by financial resources but by the non-cognitive skill of self-efficacy. We show that in the educational setting, the non-cognitive skill of grit predicts performance declines. This finding suggests that in addition to explaining performance, grit is also relevant for response to negative shocks in the educational setting, for instance distance learning during the Covid-19 pandemic.

Our behavioral measure of grit is based on identifying performance declines, and utilizing the unique advantage offered by the data recorded in voluntary after-school hours activity on the platform to measure the revealed effort response. In this sense, our measure captures the "perseverance of effort" aspect of the construct of grit. Our results are therefore consistent with [Credé et al.2017] and [Ponnock et al.2020], who show that it is the perseverance facet of grit (in contrast to consistency of interest) that has explanatory power over academic performance.

Our findings suggest that programs that aim to build grit early on in the educational environment(e.g. through promoting growth mindset, see for example [Dweck2006] and [Yeager et al.2019]), can also have positive effects on resilience and the academic performance response in times of crisis. To the extent that such crises also exacerbate educational inequalities [Murat and Bonacini2020], teaching grit could reduce the widening gap between students coming from different social strata. The results are also useful for designing AI-based remote learning technologies —identifying students' non-cognitive skills based on their behavior in learning software can be beneficial in creating customized/personalized content for each student and potentially improve achievement. Moreover, features that teach growth mindset and promote grit can be incorporated into digital learning environments, to support self-regulated virtual learning. We expect that such strategies will be especially relevant with the growing use of AI-based online learning technologies. Finally, our results underscore the importance of constructing revealed preference measures of non-cognitive skills and how self-reported survey data can be complemented with high-dimensional revealed preference data for predicting educational outcomes.

Tables

Table 3.1: Average Scores Pre- and Post-Covid, by gender

	Math Score			Science Score		
	(1)	(2)	(3)	(4)	(5)	(6)
	Pre- Covid	Post-Covid	Change	Pre- Covid	Post-Covid	Change
All Students	59.7	50.0	-9.5	65.4	54.6	-10.9
Female	60.1	50.1	-9.8	65.0	55.1	-10.0
Male	58.9	49.9	-9.0	66.2	53.7	-12.5

Table 3.2: Behavioral Grit and Pre- and Post-Covid Scores

	Pre-Covid		Post-Covid	
	(1)	(2)	(3)	(4)
	Math Score	Science Score	Math Score	Science Score
Behavioral Grit	4.6**	2	15.4***	15.2***
	(0.04)	(0.41)	(0.001)	(0.001)
Male	-20.3**	-28**	-10.7	-15.2
	(0.03)	(0.02)	(0.29)	(0.11)
Diagnostic Score	0.04**	0.005***	0.04**	-0.009***
	(0.001)	(0.38)	(0.001)	(0.14)
R^2	0.48	0.54	0.38	0.43
N	1738	1714	1738	1714

Notes: Coefficients are reported from OLS regressions where the dependent variables are pre/post-Covid math and science scores. P-values are reported in parentheses. Class-school dummies and grade are added as control variables. Very high performer students (average score > 90) for each subject are excluded from the analysis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.3: Survey Grit and Pre- and Post-Covid Scores

	Pre-Covid		Post-Covid	
	(1)	(2)	(3)	(4)
	Math Score	Science Score	Math Score	Science Score
Survey Grit	0.5	0.8**	0.5	0.5
	(0.17)	(0.04)	(0.25)	(0.26)
Male	-19.7*	-6	-8.7	-14.3
	(0.05)	(0.53)	(0.42)	(0.18)
Diagnostic Score	0.04***	0.005	0.04***	0.01*
	(<0.001)	(0.4)	(<0.001)	(0.05)
R^2	0.48	0.54	0.33	0.36
N	1738	1714	1738	1714

Notes: Coefficients are reported from OLS regressions where the dependent variables are pre/post-Covid math and science scores. P-values are reported in parentheses. Class-school dummies and grade are added as control variables. Very high performer students (average score >90) for each subject are excluded from the analysis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.4: Behavioral Grit, Survey Grit, and Pre- and Post-Covid Performance Change

	(1) Change in Math Score	(2) Change in Science Score	(3) Change in Math Score	(4) Change in Science Score
Change in Behavioral Grit			0.17***	0.1**
			(<0.001)	(0.01)
Survey Grit	0.03	-0.2		
	(0.94)	(0.64)		
Male	11.1	-8.2	11	35.4**
	(0.36)	(0.49)	(0.30)	(0.01)
R^2	0.32	0.31	0.35	0.34
N	1738	1714	1738	1714

Notes: Coefficients are reported from OLS regressions where the dependent variables are changes in math and science scores. P-values are reported in parentheses. Class-school dummies and grade are added as control variables. Very high performer students (average score > 90) for each subject are excluded from the analysis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

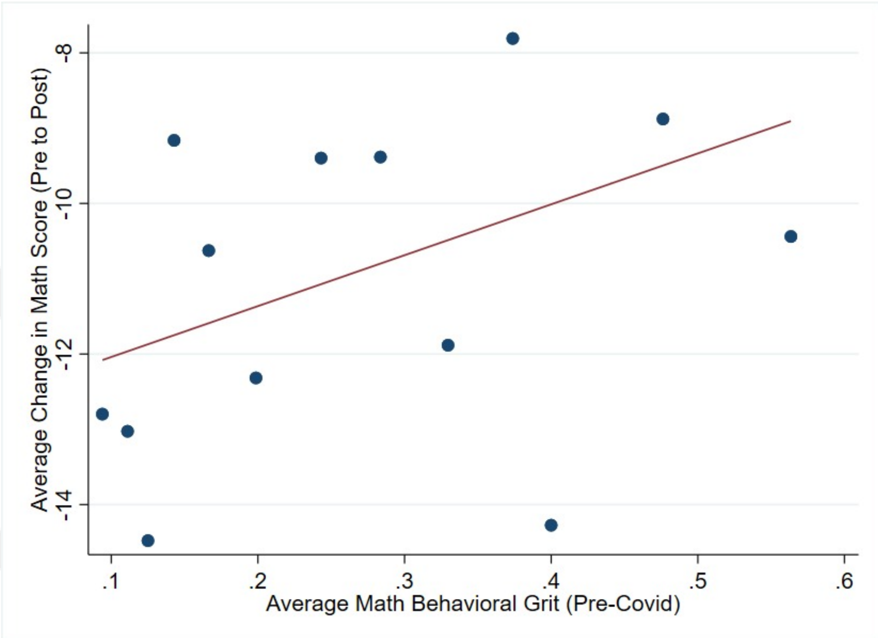
Table 3.5: Elastic Regression Results

	(1) Change in Performance (Math and Science)
Behavioral Grit	0.73***
	(<0.001)
pc2	-11.22***
	(<0.001)
pc7	7.41***
	(<0.001)
pc14	-5.23***
	(<0.001)
pc10	5.06***
	(<0.001)
pc5	-4.34***
	(<0.001)
pc11	4.478***
	(<0.001)
pc18	-4.077***
	(<0.001)
pc1	4.04***
	(<0.001)
pc8	3.81***
	(<0.001)
pc17	2.72***
	(<0.001)
pc13	2.56***
	(<0.001)
pc9	2.1***
	(<0.001)
pc12	1.374**
	(0.024)
pc15	-0.71
	(0.31)
pc16	-0.567
	(0.35)
pc19	0.43
	(0.42)
R^2	0.77
N	1434

Notes: Coefficients are reported from OLS regressions where the dependent variable is average score. P-values are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

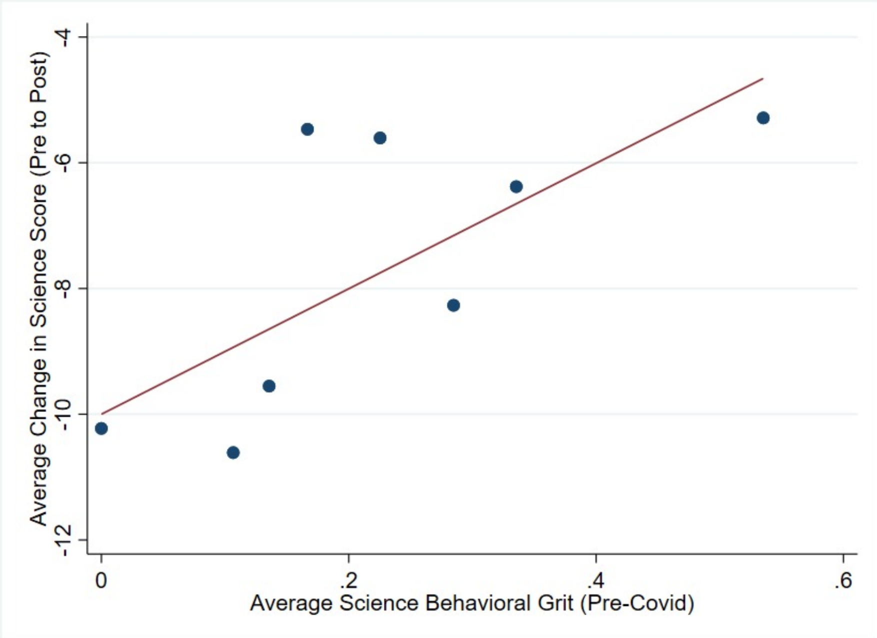
Figures

Figure 3.1: Pre-Covid Behavioral Grit and Academic Resilience in Math



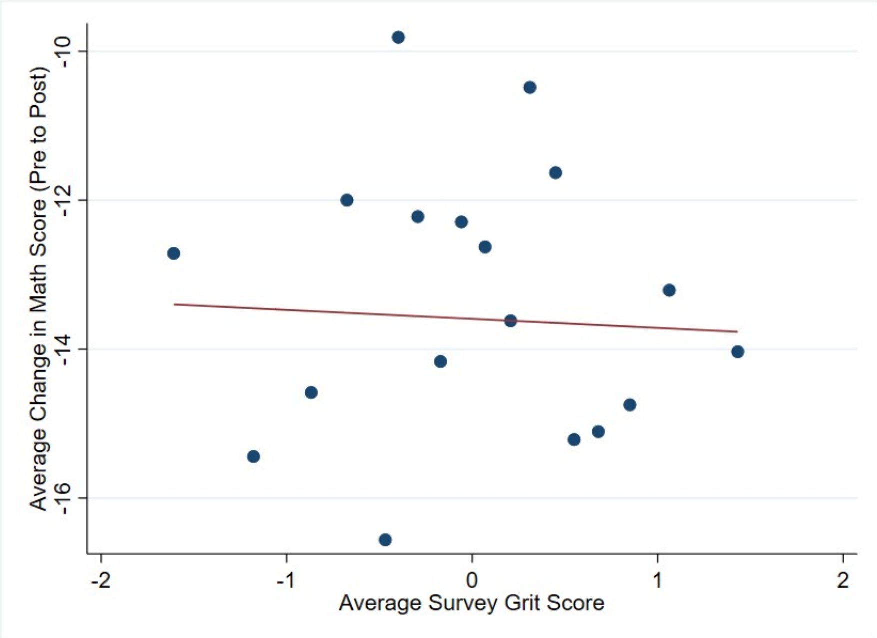
Notes: Binned scatterplots of correlation between behavioral grit and change in math scores

Figure 3.2: Pre-Covid Behavioral Grit and Academic Resilience in Science



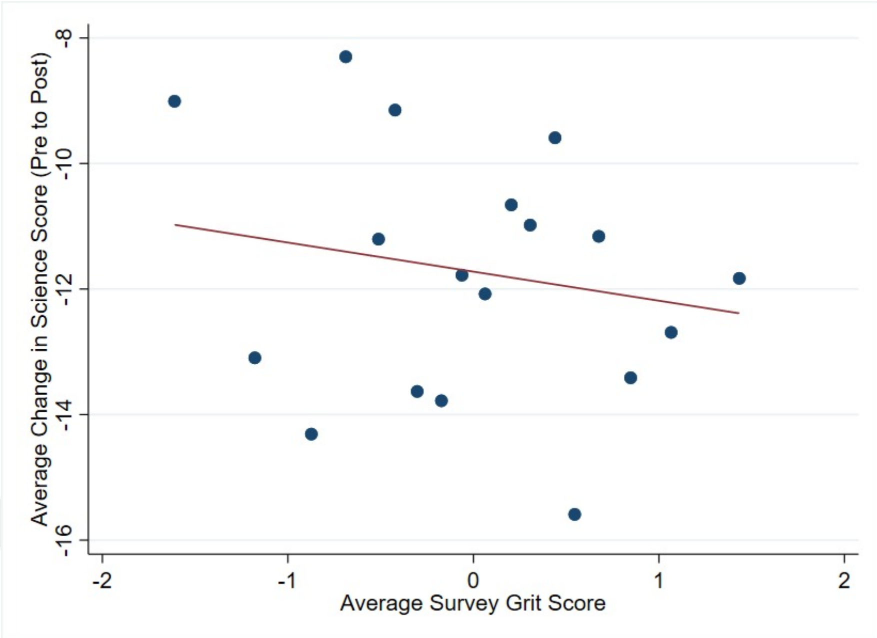
Notes: Binned scatterplots of correlation between behavioral grit and change in science scores

Figure 3.3: Survey Grit and Academic Resilience in Math



Notes: Binned scatterplots of correlation between survey grit and change in math scores

Figure 3.4: Survey Grit and Academic Resilience in Science



Notes: Binned scatterplots of correlation between survey grit and change in science scores

Appendix A

MODELING HETEROGENEOUS RISK TAKING BEHAVIOR OF FOOTBALL MANAGERS

A.1 Additional Literature: Behavioural Analytics in Football

Dynamic competitive settings, in particular, can be defined as a competitive environment where competitors get feedback about the performance of their counterparts before the end of the competition. Such feedback might create psychological pressure on each side and eventually impact their subsequent performance. An extensive literature review has been done to understand in what form such psychological pressure can be defined and how team, player and manager characteristics play a role in dealing with such pressure. Although this study is not directly about understanding pressure and its impact, understanding what has been done on the literature is important as it will give us clues about the determinants of risk behaviour and its impact on outcome. For example, if such pressure exists and we observe its impact on risk taking, then we have to find ways to measure the level of pressure and its impact on risk taking behavior.

Psychological pressure to create a first mover advantage

[Apesteguia and Palacios-Huerta2010] analyze penalty shoot-outs and examine whether psychological pressure creates any first-mover advantage. Their rationale of focusing on penalty shout-outs is that it creates a perfect a randomized natural experiment and any possible endogeneity problem is naturally avoided. Professionals are aware of this psychological impact and rationally react by systematically benefiting them. [Kocher et al.2012] extend the sample size (number of matches with penalty shootouts) and conclude that there is no significant first-mover advantage. An opposite conclusion is from [Feri et al.2013] who document that second movers perform significantly better under psychological pressure. Both papers had a remark on heterogeneity of psychological pressure on the performance of the subjects. [De Luca et al.2015] take nationality of players as a source of heterogeneity and report that Southern European football players are more successful in penalty kicks than their British colleagues, similarly European football players being less successful than British players. First mover advantage has been also analyzed in the context of two-legged football matches. [Flores et al.2012] show that Clubs playing the first leg at home have a better chance of converting any given negative first-leg outcome into eventual success.

Home advantage and fan pressure

[Pollard2008] provides a comprehensive literature review for home advantage in football and summarizes effects of crowd, travel, familiarity, referee bias, territoriality, specific tactics, rule factors and other psychological factors. The most obvious factor behind home advantage is the crowd effect created by dominant fans. However, the magnitude of this impact varies based on crowd size, density, intensity of support and proximity to the field [Böheim et al.2019]. There are contradictory results on the effect of distance traveled which is believed to create a disadvantage for the away team. Another reason behind home advantage would be playing at a familiar stadium in familiar conditions, and authors find that familiarity is another factor supporting home advantage. A common question being asked is whether referees are more biased towards home teams, and different studies agree on the presence of such bias. [Besters et al.2019] provide supportive evidence on the existence of referee bias for home team players. [Neave and Wolfson2004] also conclude that psychological factors are evident in explaining home advantage in team sports. [Pollard2006] reports country specific differences in home advantage. For example, Balkan countries (especially Bosnia and Albania) have more home advantage than average, however it is lower than countries in northern Europe. [Gürkan et al.2017] also provide supporting evidence for home advantage from Turkish league). On the contrary, [Thomas et al.2004] claim that home advantage exists, but there is a loss of advantage recently in playing at home vs away (i.e. playing at home has become less advantageous due to fan pressure). Due to Covid-19, matches were played without the presence of fans, which created a perfect opportunity for researchers to test hypothesis for home advantage and potential effects of fan pressure. [Tilp and Thaller2020] indeed show that Covid-19 created a disadvantage for home teams in the absence of their fans. Analysis of [Sors et al.2020] support these findings with a different data set (all pre-Covid matches played without spectators). [Besters et al.2019], and [Lago-Peñas et al.2016] go further and discuss how team quality affects the impact of playing at home.

Situational pressure and psychological momentum

Factors such as playing at home vs. away, teams' relative strengths and importance of the match to teams are known ex-ante to the players or coaches, and it is relatively easier to analyze their impact on behavior or performance. There are also factors occurring during the game and believed to create a situational pressure or psychological momentum for players. Scoring time, for instance, is a widely analyzed factor. [Meier et al.2020] analyze games from top five European football leagues between the 2013/14 and 2017/18 seasons, and find that if scoring happens just before half time, then the scoring team benefits more than scoring at an earlier time. This finding has been supported by [Greve et al.2020] with a similar methodology applied on a different data set. On the contrary, another recent study of [Gauriot and Page2018] analyze different time slots before the half time and do not find any significant psychological momentum created by scoring before the half time.

Another situational pressure factor studied largely in the literature is being the one who scores first. [Liu et al.2021] collect match results from Chinese Football Super League and show that scoring first creates an edge for the scoring team both in terms of the game result and goal difference. They also point out that the later the goal is scored, a better chance for the scoring team. [Martínez and González-García2019] support the importance of scoring first by analyzing the results of World Cup and UEFA Euro (using a smaller number of observations tough). [Ibáñez et al.2018] analyze Spanish women's football league and reached out to the same conclusion. Although advantage of scoring in the last minutes is obvious since there is almost no time for the opponent to respond, [van Ours and van Tuijl2011] investigate whether there are country specific characteristics in scoring at the final stage of the matches. According to their analysis, there are some countries dealing better with the last-minute stress and become more successful in scoring at the end.

In contrast to the effect of scoring first or the time of scoring on performance, there is a scant literature on the impact of yellow or red cards on the performance and behavior. [Červenỳ et al.2018] examine how sending off a player impact the goal scoring rates in FIFA World Cup matches. Their finding is intuitive: Teams who lost their player are less likely to score whereas the opponent finds more chances to score afterwards.

Learning from prior mistakes and responding to setbacks

A frequently asked question is whether players or teams learn from their mistakes, or given a setback, whether they respond well or not. [Haenni et al.2016] investigate demotivational effect of losing in repeated competitions. Analyzing a huge data set from tennis players (45k players), he concludes that players who lost before hesitate to play the next match and wait longer for another confrontation. As opposed to losing, [Iso-Ahola and Dotson2016] study whether consecutive winnings are important for achieving long run success and how attributes such as confidence, perceived competence, and internal (ability-skill) characteristics play a role in success. There are contradictory results on the relationship between setbacks and learning. [Rauter et al.2018] propose team reflexivity as a moderator between team-experienced setbacks and team learning. Setbacks can create further discouragement [Gill and Prowse2012] and responses to rival's effort might be even more negative. [Lackner et al.2020] also conclude that the effort is negatively impacted by the ability of the rival, further supporting the discouragement effect. In the context of football, [Bucciol et al.2019] analyze how risk-taking behavior of the coaches are changing when there is a setback, and their conclusions are consistent with the previous literature. It mostly depends on experience in the past.

Injury can be considered as a critical setback that not only affects players' psychical fitness, but also demoralizes the player. Each team puts their newly acquired players, before the final signature, into a medical test and checks for all potential or actual injuries player might have. Injury history is always an issue managers take into account when deciding about a player's future in the club. [Ardern et al.2013] investigate what positive psychological factors can be associated with the likelihood of returning to pre-injury level of fitness after the injury. Motivation, confidence and low fear are the type of characteristics that make a player more likely to recover soon. On the other hand, [Devantier2011] figures out the negative psychological factors impeding recovery process and time for male soccer players in Denmark. According to their results, injury history and coping with adversity are the most impactful negative psychological factors and should be considered by managers and medical staff. [Hardy1992] summarizes the literature on psychological stress in sport in general, and documents how psychological factors can impact performance, injury likelihood and recovery speed from injury. Three of the most explanatory psychological factors are fear of failure, social evaluation and loss of internal control over one's environment. More literature on psychological factors related to injuries can be found in [Junge2000], [Podlog and Eklund2005], [Podlog and Eklund2007], and [Wiese-Bjornstal et al.1998].

Role of expectations in dealing with psychological pressure

In this part of the literature review, we will link prospect theory and behavioral decision-making literature to psychological pressure concept and discuss how certain expectations can mediate this relationship.

A well-established concept in the prospect theory and behavioral decision-making literature is that people evaluate outcomes as gains or losses depending on a reference point. Reference dependence plays a crucial role in explaining people's preferences or decision taking behaviors. One strand of the literature focuses on what determines the reference point. One hypothesis is that expectations shape the reference points and there is a fair amount of evidence from the literature. [Bewley2021] interviews business executives, labor leaders, and professional recruiters and reports that workers compare current earnings to previous earnings and wage cuts hinder worker morale. Workers dislike falling short of their previous earnings as it constitutes expectation based reference point. [Mas2006] finds that, in the months following a New Jersey police officer's loss in final-offer arbitration over salary demands, arrest rates and average sentence length decline and crime reports rise. In a large, multinational company, [Ockenfels et al.2015] study how bonus payments affect managers' satisfaction and performance. They find that pay-outs falling short of individual bonus targets (which is a likely expectation-based reference point) decrease work satisfaction among managers. [Markle et al.2018] provide additional support for expectation-based reference points from marathon runs and [Chikish2019] from skating competitions.

One limitation of defining reference points with expectations is that expectations are hard to observe on the field [Abeler et al.2011]. [Bartling et al.2015] accept pre-play betting odds as ex-ante expectations and show that players' and coaches' behavioral outcomes are reference dependent. By defining the loss frame as a team when it is behind its reference point, they have shown that players in a loss frame are more likely to breach the rules of the game, and coaches are more aggressive in substitution (making riskier substitution decision to revert the result). [Lee2020] studies reference dependent preferences among supporters. [Fry et al.2021] mention the importance of managing expectations and how these expectations creates a negative or positive impact on the game outcome.

Given that expectations are important in behavioral decision making, there is also a literature on the relationship between psychological factors and expectations.

[Arrondel et al.2019] analyze the penalty kicks during French cup competitions and find that the performance is negatively affected by value at stake (how important is the current penalty kick for expected game outcome). Another study [McFall and Rotthoff2020] analyzing shots made in PGA TOUR, closes the gap in research for the link between performance and reward. They conclude that the more rewarded the outcome of final shot is, less likely the player is to successfully make the shot.

Manager choices and psychological pressure

Another ongoing debate on the football pitch is how much freedom and self-control a manager (i.e. also used as coach) should provide to players to boost their performance. [Navia et al.2019] investigate the level of self-control that should be given to players under pressure. Their results confirm increased anxiety when taking the penalty under presence (which is defined as performance being evaluated by the coaching staff). They observe that players perform better when they are self-controlled. [Boto-Garcìa et al.2020] use self-confidence of managers as an explanatory variable to predict performance and find that performance is positively affected by managers' confidence and risk tolerance. [Schneemann2017a] focuses on to what extent interim results can explain changes in managers' risk strategy during the game. They conclude that favorites and underdog managers respond differently to interim feedback. Underdog managers tend to choose safer strategies when leading, and favorites further take more risks when behind. However, they could not find a significant relationship between the risk taking strategy and match outcome.

Experience is another factor that help managers to perform better under pressure. [Detotto et al.2018] show if managers have international experience (outside of currently working country) or had previous experience with the current club as a player, they perform better under pressure. Similarly, performance is also adversely affected by lack of managerial experience in general. [Duarte et al.2014] support the positive impact of managerial experience on teams' performances under pressure; however, this relationship does not necessarily hold for players. According to [Potrac et al.2002], having an established social bond between the manager himself and players further increases the likelihood of success, which may be easily observed among managers who had previously played for the same club as a player. Expectations from the manager, current performance of the team and overall experience are the factors that impact a manager's future in the club [Humphreys et al.2016]. An interesting question one could ask here is how changing the manager during the season due to his low performance can affect performance. [Flores et al.2012] show that the change further deteriorates the results. On the contrary, another study argues that sacking the manager might pay-off well by pleasing the supporters and getting back the crowd support [de Dios Tena and Forrest2007].

A.2 Summary of Literature

Table A.1: Literature Review - Football

				Determinants of Risk-taking		Effect of Risk on Performance	
Domain	Paper	Risk Definition	Method	Ex-ante Information	Interim Information	Performance Definition	Effect
Football	[Bartling et al.2015]	Substitutions	OLS and linear probability model		Goal difference(-), loss frame (-)	Match result (win, draw, loss)	Negative
Football	[Bucciol et al.2019]	Initial formation	Probit	Home, prior defeats (+), wins, unexpected results, absolute salary, relative salary (+), tenure		Match result (win, draw, loss)	Not significant
Football	[Grund and Gürtler*2005]	Substitutions	Probit	Home, difference in ranking (-)	Goal difference(-)	Score and point advancement(-)	Negative
Football	[Mueller-Langer and Andreoli-Versbach2017]	% of wins/lose comparing to draws	OLS		First leg results		
Football	[Schneemann2017b]	Substitutions	OLS and Logit	Home(+), favorite	Goal difference (-), red card (-), minute (+)	Match result (goal difference)	Not significant

Table A.2: Literature Review - Other Domains

				Determinants of Risk-taking		Effect of Risk on Performance	
Domain	Paper	Risk Definition	Method	Ex-ante Information	Interim Information	Performance Definition	Effect
Basketball	[Böheim and Lackner2015]	Fraction of three-point shots	Probit	Home, match importance (-)	Point difference (+ for male, - for female), minutes left (-)	Match win probability	Positive for female, nagative for male
Basketball	[Grund et al.2013]	Fraction of three-point shot	OLS- Logit		Point difference	Game outcome	Positive when point difference is susbtantial
Car Racing	[Bothner et al.2007]	Probability of crashing the car	Logit		Crowding from above and below (+)		
Car Racing	[Yaskewich2017]	Accident proneness	Linear probability model, Logit	Pick-A-series rule change (-), age, experience, point rank (-), intensity of competition			
Handball	[Neuberg and Thiem2022]	substituting the goalkeeper for an additional field player	linear probability model - OLS	Home, rank difference (-)	Goal differential (-), timing of the match (+)	Success of a possession, match result (win/draw/lose)	Positive for underdogs, negative for favorite
High Jump/ pole vault	[Böheim et al.2022]	Share of passes	difference-in-difference model	Gender		Fail after pass	Negative for male, insignificant for female

Table A.3: Literature Review - Other Domains 2

				Determinants of Risk-taking		Effect of Risk on Performance	
Domain	Paper	Risk Definition	Method	Ex-ante Information	Interim Information	Performance Definition	Effect
Golf	[McFall and Rothoff2020]	Going for the green (gambling) with his second shot	OLS	Unit purse (-), superstar effect (+), tournament rank, difference in rank, player characteristics	Yards to Pin (+)		
Golf	[Ozbeklik and Smith2017]	Percentages of holes conceded	Probit		Hole difficulty, rank difference(-), accum score (-)		
Poker	[Eil and Lien2014]	Propensity to fold	Cox likelihood regressions	Experience (+)	Losing within a session (+)		
Poker	[Lee2004]	Absolute value of chip change	OLS		Expected gains and losses in the game		
Weight-lifting	[Genakos and Pagliero2012]	Weight announced to lift	OLS	Home	Interim rank (+), success in previous attempt (+)		

A.3 Variables

Table A.4: Team and Match Features

Variables	Description
Relative Strength	Difference in strength of competing teams. It is the one minus the ratio of total points gathered by each team in the last 5 years.
Derby	Whether the match is recognized a derby match by the authorities. Barcelona vs. Real Madrid or Manchester United vs. Manchester City are some examples.
Current Ranking	Current ranking of the team in the league ranging from 1 to number of teams competing in the league
Match Importance	1 if both teams are either ranked at top 6 positions in the league or ranked at bottom 6 positions
Home Match	Playing at home vs. away

Table A.5: Manager Characteristics

Variables	Description
Manager Age	Age of the manager
Manager Tenure	Average tenure in all the clubs he worked as manager
Manager Foreign	Whether the manager is foreign or local

Table A.6: Interim Feedback Variables

Variables	Description
Minute of the Game	Minute of the substitution ranging from 1 to 90
Opponent Risk-taking	Whether the opponent already made a risky substitution
Red Card	Total red cards shown to the team
Red Card Opponent	Total red cards shown to the opponent team

Table A.7: Past Performance Variable

Variables	Description
Unexpected Consecutive Game Lost	Number of consecutive matches where they lost the game unexpectedly
Unexpected Consecutive Game Won	Number of consecutive matches where they win the game unexpectedly
Humiliated Defeat	Whether the last match result was a humiliating defeat
Big Victory	Whether the last match result was a big victory
Last 5 Matches-Average Point	Average points in the last 5 match

Table A.8: Time and Country Variables

Variables	Description
Year	Year of the match
Bundesliga (Germany)	Whether the match is played in Germany league
Ligue 1 (France)	Whether the match is played in France league
Premier League (England)	Whether the match is played in England league
LaLiga (Spain)	Whether the match is played in Spain league
Seria A (Italy)	Whether the match is played in Italy league

A.4 Additional Figures

Figure A.1: Market Size of the Global Sports Market from 2011 to 2018 (in Billion U.S. Dollars)

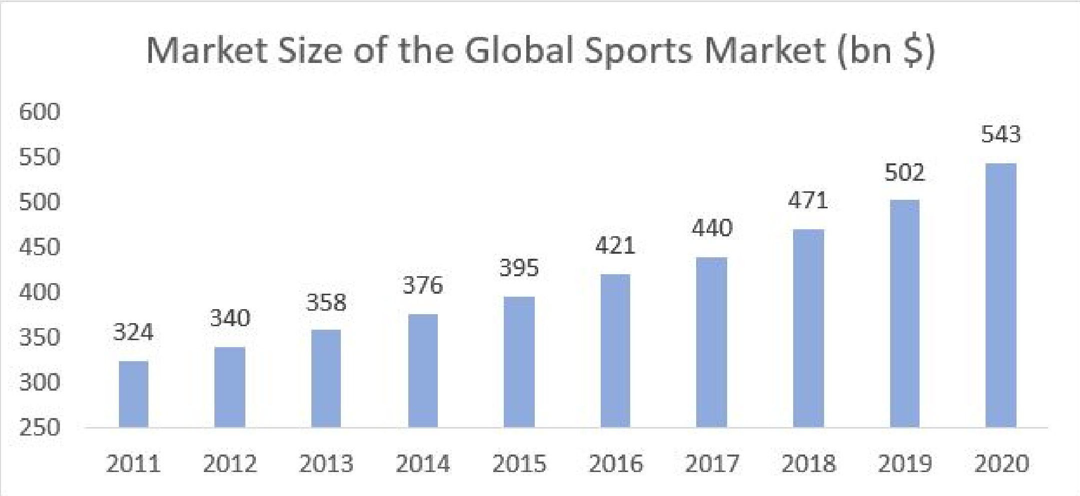


Figure A.2: Publication Trend in Sports Management Literature

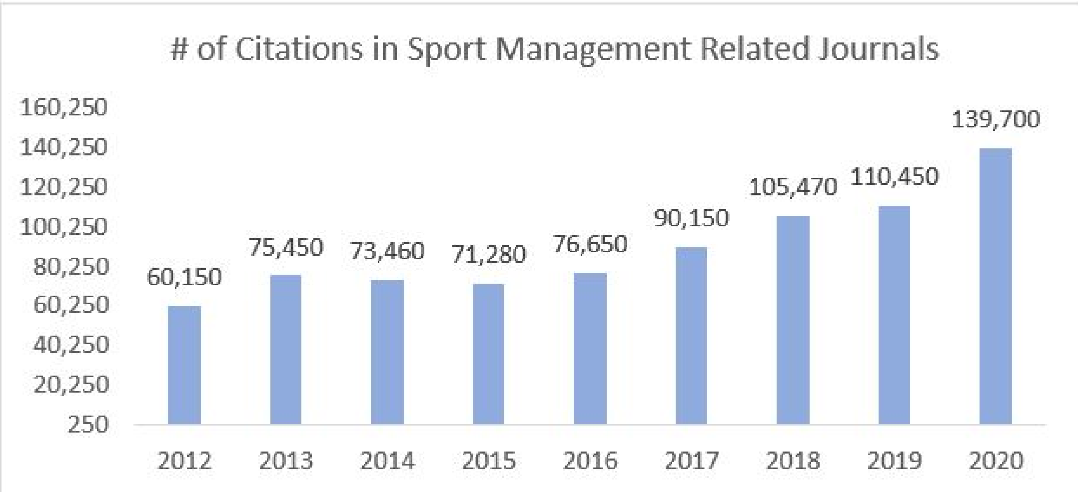
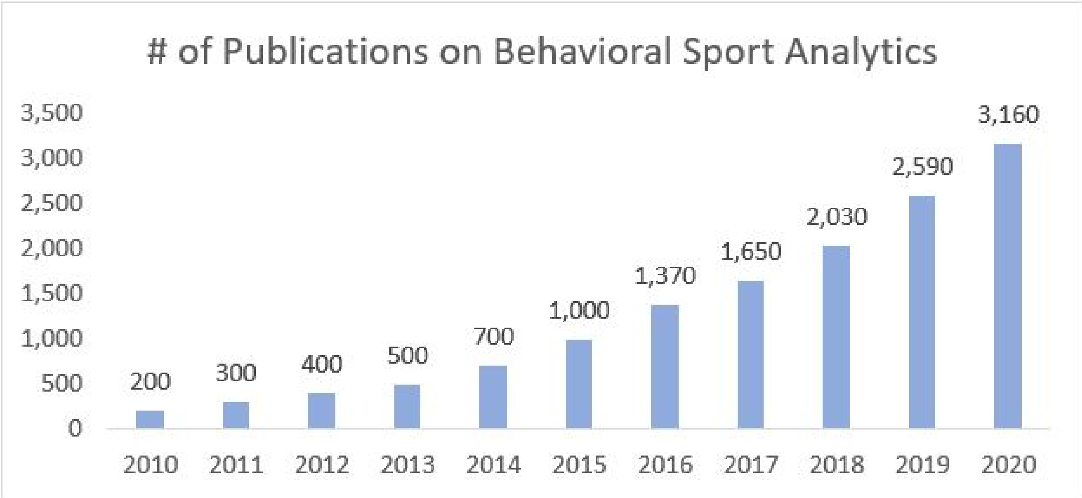


Figure A.3: Publication Trend on Behavioral Sports Analytics



A.5 Additional Analysis

Table A.9: Logit Regression - Predicting Risk Taking - Long

	(1)
Goal Difference	-0.053***
	(0.00)
Red Card	-0.095***
	(0.01)
Red Card Opponent	0.062***
	(0.01)
Minute of the Game	-0.001***
	(0.00)
Relative Strength	0.020***
	(0.00)
Home Match	0.014***
	(0.00)
Derby	-0.004
	(0.00)
Current Ranking	0.001***
	(0.00)
Opponent Risk Taking	-0.012***
	(0.00)
Match Importance	-0.000
	(0.00)
Manager Age	-0.001**
	(0.00)
Manager Foreign	0.009**
	(0.00)
Manager Tenure	0.001
	(0.00)
Unexpected Consecutive Game Lost	-0.002
	(0.00)
Unexpected Consecutive Game Won	0.004
	(0.00)
Last 5 Matches-Average Point	-0.001
	(0.00)
Humiliated Defeat	0.012
	(0.01)
Big Victory	0.002
	(0.01)
Time and Country	✓
N	73984

Notes: Estimates are average marginal effects after logit regression. Standard errors are reported in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$. The outcome variable a dummy variable which equals 1 if the manager makes a risky substitution. Controls include time and country dummies.

Table A.10: OLS - Predicting Point Advancement - Long

	(1)
Risk Taking	0.212***
	(0.07)
Goal Difference	-0.145***
	(0.00)
Red Card	-0.196***
	(0.01)
Red Card Opponent	0.147***
	(0.02)
Minute of the Game	-0.002***
	(0.00)
Relative Strength	0.117***
	(0.01)
Home Match	0.121***
	(0.01)
Derby	-0.043***
	(0.01)
Current Ranking	-0.036***
	(0.00)
Opponent Risk Taking	-0.031***
	(0.00)
Match Importance	0.001
	(0.01)
Manager Age	0.001**
	(0.00)
Manager Foreign	-0.040***
	(0.01)
Manager Tenure	-0.006***
	(0.00)
Unexpected Consecutive Game Lost	0.004
	(0.00)
Unexpected Consecutive Game Won	0.016***
	(0.01)
Last 5 Matches-Average Point	-0.151***
	(0.01)
Humiliated Defeat	-0.009
	(0.01)
Big Victory	-0.009
	(0.01)
Risk Taking × Goal Difference	0.018***
	(0.01)
Risk Taking × Red Card	0.040
	(0.03)
Risk Taking × Red Card Opponent	0.025
	(0.03)
Risk Taking × Minute of the Game	-0.003***
	(0.00)
Risk Taking × Relative Strength	0.020*
	(0.01)
Risk Taking × Home Match	0.011
	(0.01)
Risk Taking × Derby	-0.008
	(0.02)
Risk Taking × Current Ranking	0.000
	(0.00)
Risk Taking × Opponent Risk Taking	0.000
	(0.01)
Risk Taking × Match Importance	0.020*
	(0.01)
Risk Taking × Manager Age	0.001
	(0.00)
Risk Taking × Manager Foreign	-0.002
	(0.02)
Risk Taking × Manager Tenure	0.002
	(0.00)
Time and Country	✓
N	73984

Notes: Estimates are obtained via Ordinary Least Squares regression. Standard errors are reported in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$. The dependent variable is point change from the time of substitution to final whistle and takes values between -3 and 3. Controls include time and country dummies.

Appendix B

MODELING HETEROGENEOUS RISK TAKING BEHAVIOR OF FOOTBALL MANAGERS WITH CAUSAL MACHINE LEARNING METHODS

B.1 Main Steps in Implementing Matching Methods

1. Define closeness criteria (how you decide whether a selection is good enough to be included)
   - Variables to include:
     - [Shadish et al.2008] point out the importance of large set of predictors of convenience
     - Exclude variables that may have been affected by the treatment of interest [Greenland2003]
   - Distance measures:
     - Exact
       - [Imai et al.2014] state that exact matching is ideal in many ways, yet it is difficult to achieve especially when X is high dimensional
     - Mahalanobis
       - The Mahalanobis distance can work quite well when there are relatively few covariates [Zhao2004]
     - Propensity score (discussed above)
     - Linear propensity score
       - [Rosenbaum and Rubin1985], [Rubin and Thomas1996], and [Rubin2001] agreed on linear propensity score as it is effective in

reducing the bias.

2. Apply the matching method based on the chosen closeness criteria

   - Nearest neighbour matching
     - Optimal matching
       * This method considers the overall set of matches when choosing individual matches, hence minimizes a global distance measure [Rosenbaum2002]
     - Selecting the number of matches
       * This method gets multiple good matches for each treated individual, called ratio matching [Rubin and Thomas2000, Smith1997]
     - With or without replacement
       * Can we use controls as matches more than one? This might be beneficial when there are few control observations with respect to treatment.
   - Subclassification, Full Matching, and Weighting
     - Subclassification
       * Best example is to use quantiles for propensity scores. [Lunceford and Davidian2004] states that more sub-classes are needed as the sample size increases.
     - Full Matching
       * Minimizes the distances between each treated individual and each control individual within each matched set.
     - Weighting adjustment
       * Propensity scores are used as inverse weights in the estimation of the ATE. Benefits of propensity scores are explained above. One drawback is that the variance can be very large if the weights are extreme.

- Assessing Common Support: This issue is relevant when the distribution of treatment and control groups are different (unlike the situation in propensity score or K-nearest).

3. Assess the quality of matching

4. Estimate the treatment effect

B.2 Additional Analysis

B.2.1 EBM- Additive Terms for Causal Forest

Figure B.1: EBM Additive Terms for Age

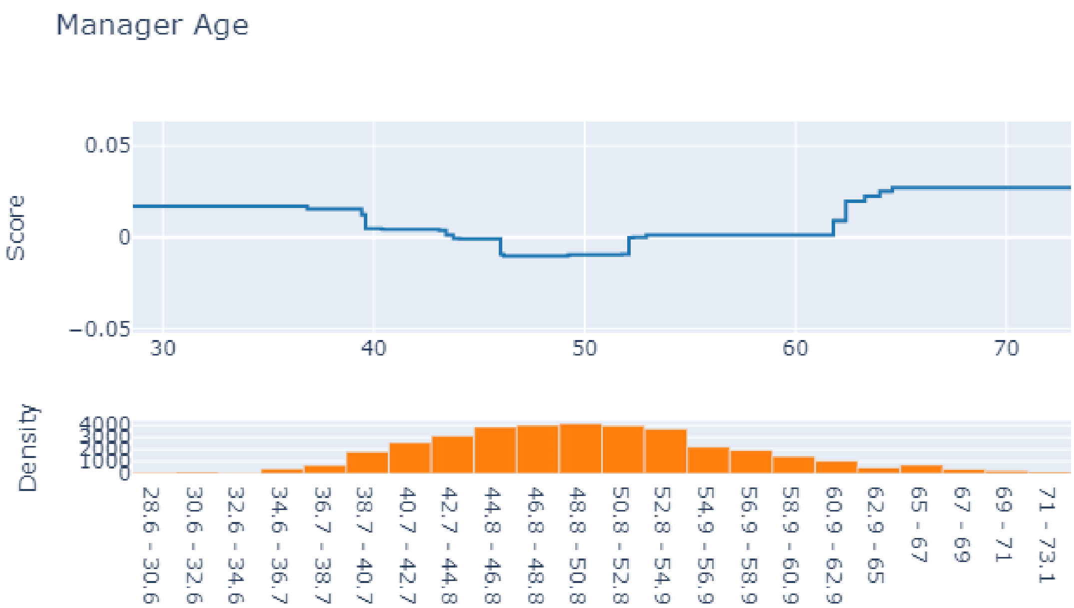


Figure B.2: EBM Additive Terms for Current Ranking

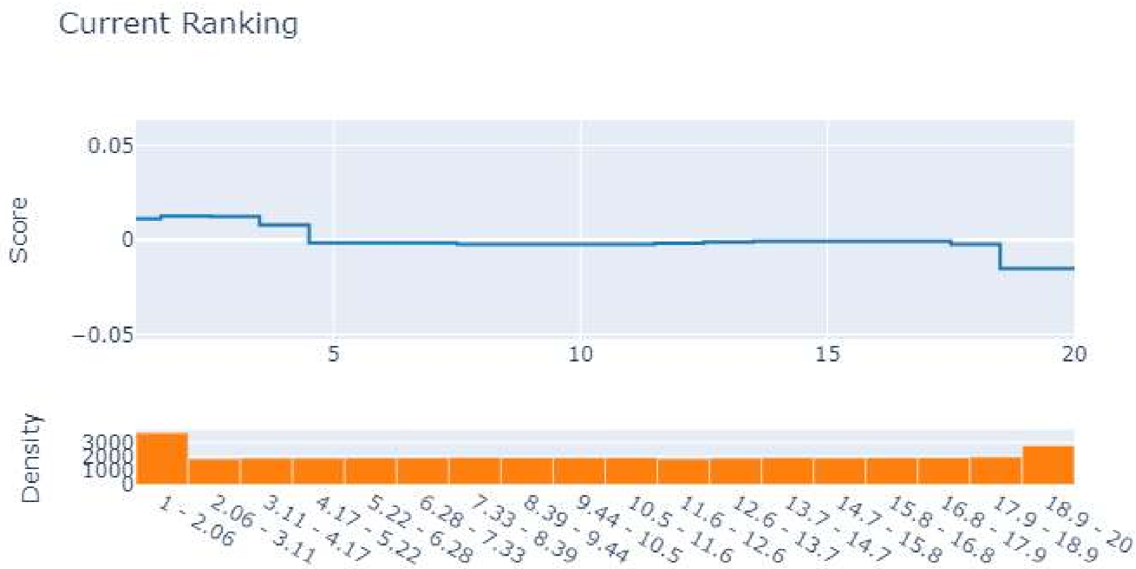


Figure B.3: EBM Additive Terms for Goal Difference

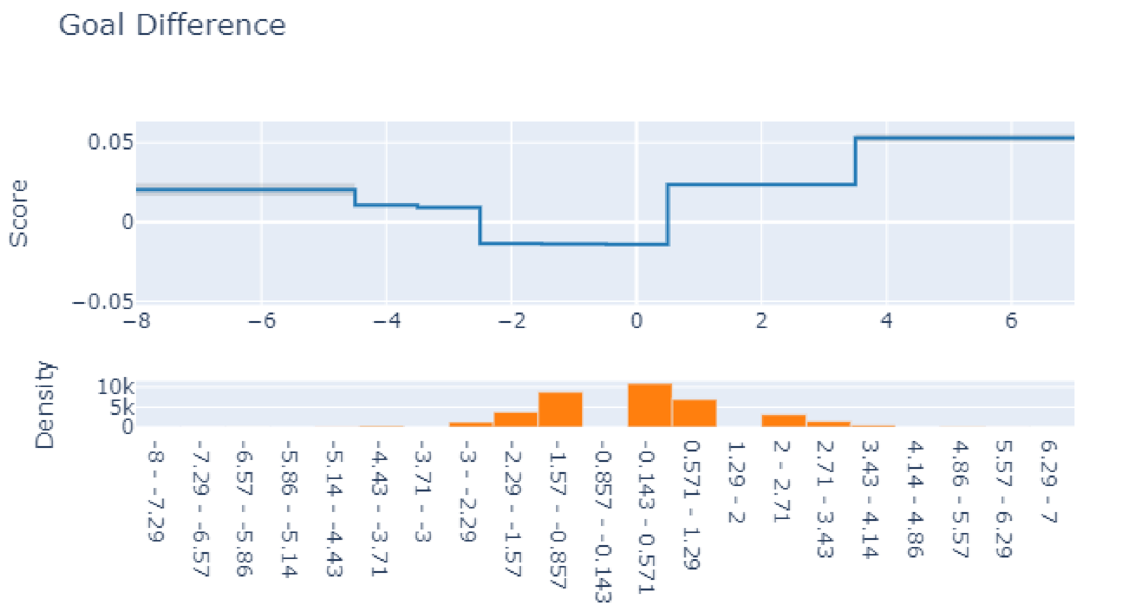


Figure B.4: EBM Additive Terms for Match Importance

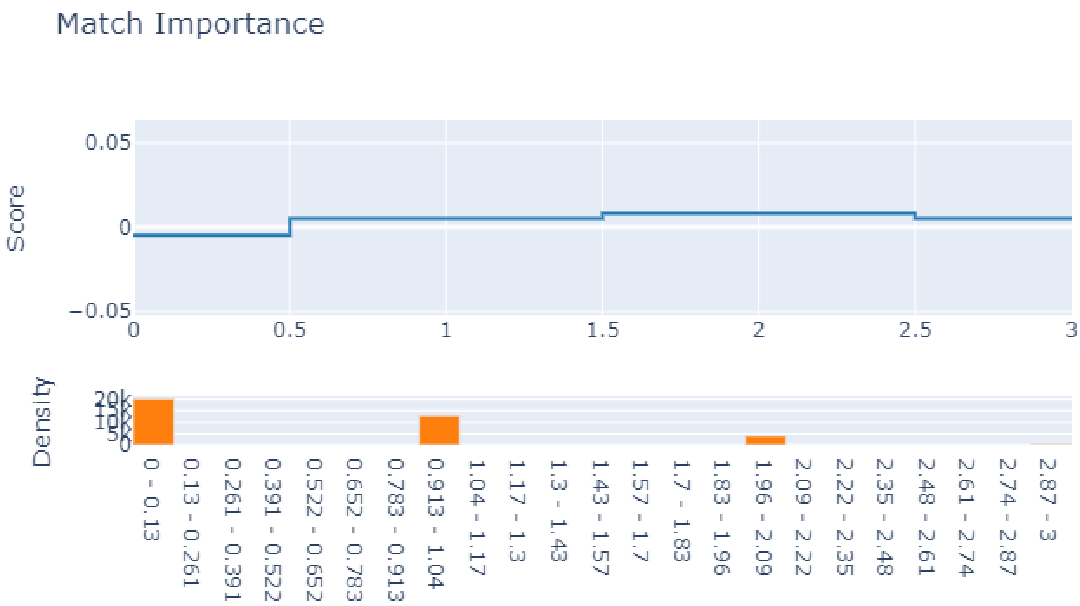
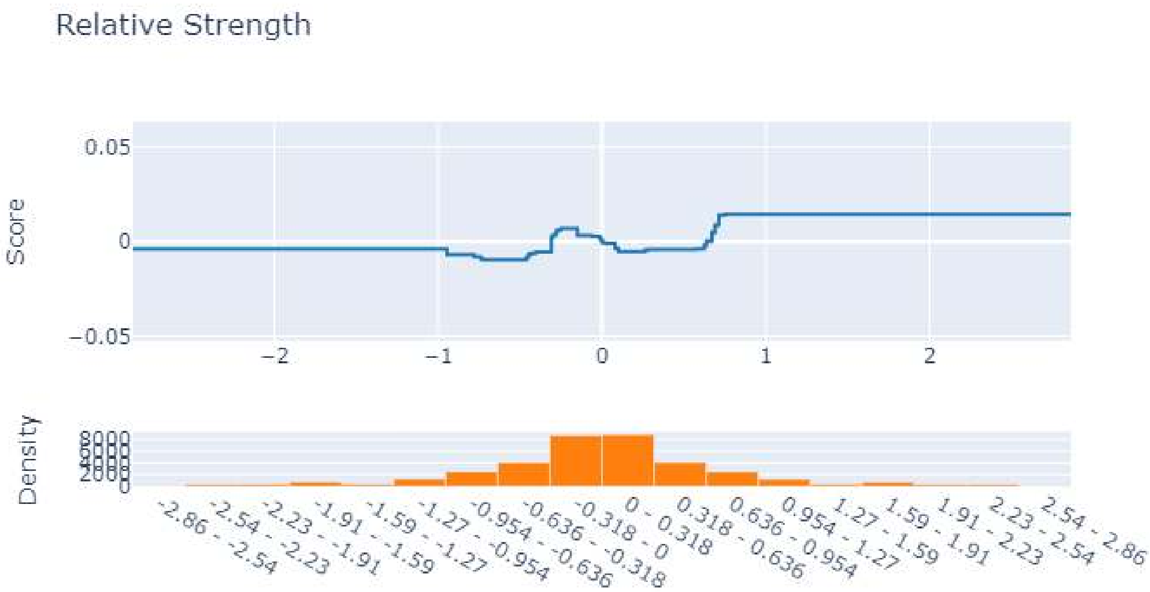


Figure B.5: EBM Additive Terms for Relative Strength



B.2.2 Robustness Check

Table B.1: Robustness Check

	Cut-off Points			Significance of Interaction Terms		
	(1)	(2)	(3)	(4)	(5)	(6)
	Goal Difference	Relative Strength	Event Minute	Trailing×Strong×Risk-taking	Trailing×Minute×Risk-taking	Minute×Strong×Risk-taking
Proposed by Causal ML	0	0.7	64	Significant (***)	Significant (***)	Significant (***)
	1	0.7	64	Not significant	Significant (**)	Not significant
Perturbing Goal Difference	-1	0.7	64	Not significant	Not significant	Significant (***)
	-2	0.7	64	Not significant	Not significant	Significant (***)
	0	1	64	Significant (***)	Not significant	Not significant
Perturbing Relative Strength	0	0	64	Not significant	Not significant	Not significant
	0	-1	64	Significant (***)	Significant (**)	Not significant
	0	0.7	57	Not significant	Significant (***)	Not significant
Perturbing Event Minute	0	0.7	52	Not significant	Not significant	Not significant
	0	0.7	47	Significant (**)	Significant (*)	Not significant

Notes: As a robustness check, we have experimented 9 additional regressions that have the same structure as the proposed model where only differences are the cut-off points of goal difference, relative strength and minute of the game

Appendix C

GRIT AND ACADEMIC RESILIENCE DURING THE COVID-19 PANDEMIC

C.1 Questions for the Survey Grit Measure:

1. New ideas and projects sometimes distract me from previous ones.
2. Setbacks don't discourage me. I bounce back from disappointments faster than most people.
3. I have been obsessed with an idea or project for a short time but later lost interest.
4. I am a hard worker.
5. I often set a goal but later choose to follow a different one.
6. I have difficulty keeping my focus on projects that take more than a few months to complete.
7. I finish whatever I begin.
8. I am diligent.

BIBLIOGRAPHY

[Abadie et al.2010] Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

[Abeler et al.2011] Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, 101(2):470–92.

[Adams and Waddell2018] Adams, N. R. and Waddell, G. R. (2018). Performance and risk taking under threat of elimination. *Journal of Economic Behavior & Organization*, 156:41–54.

[Agnew et al.2008] Agnew, J. R., Anderson, L. R., Gerlach, J. R., and Szykman, L. R. (2008). Who chooses annuities? an experimental investigation of the role of gender, framing, and defaults. *American Economic Review*, 98(2):418–22.

[Alan et al.2017] Alan, S., Baydar, N., Boneva, T., Crossley, T. F., and Ertac, S. (2017). Transmission of risk preferences from mothers to daughters. *Journal of Economic Behavior & Organization*, 134:60–77.

[Alan et al.2019] Alan, S., Boneva, T., and Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*, 134(3):1121–1162.

[Althauser and Rubin1970] Althauser, R. P. and Rubin, D. (1970). The computerized construction of a matched sample. *American Journal of Sociology*, 76(2):325–346.

[Amemiya1974] Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. *Journal of econometrics*, 2(2):105–110.

[Angrist and Pischke2009] Angrist, J. and Pischke, J. (2009). Mostly harmless econometrics: An empiricist's guide p. 373.

[Angrist et al.2000] Angrist, J. D., Graddy, K., and Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies*, 67(3):499–527.

[Angrist and Pischke2014] Angrist, J. D. and Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect.* Princeton university press.

[Apesteguia and Palacios-Huerta2010] Apesteguia, J. and Palacios-Huerta, I. (2010). Psychological pressure in competitive environments: Evidence from a randomized natural experiment. *American Economic Review*, 100(5):2548–64.

[Apicella et al.2008] Apicella, C. L., Dreber, A., Campbell, B., Gray, P. B., Hoffman, M., and Little, A. C. (2008). Testosterone and financial risk preferences. *Evolution and human behavior*, 29(6):384–390.

[Ardern et al.2013] Ardern, C. L., Taylor, N. F., Feller, J. A., and Webster, K. E. (2013). A systematic review of the psychological factors associated with returning to sport following injury. *British journal of sports medicine*, 47(17):1120–1126.

[Arkhangelsky et al.2019] Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2019). Synthetic difference in differences. Technical report, National Bureau of Economic Research.

[Arrondel et al.2019] Arrondel, L., Duhautois, R., and Laslier, J.-F. (2019). Decision under psychological pressure: The shooter's anxiety at the penalty kick. *Journal of Economic Psychology*, 70:22–35.

[Athey and Imbens2016] Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

[Athey and Imbens2019] Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

[Athey et al.2019] Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

[Athey and Wager2019] Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, 5(2):37–51.

[Athey and Wager2021] Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

[Barseghyan et al.2013] Barseghyan, L., Molinari, F., O'Donoghue, T., and Teitelbaum, J. C. (2013). The nature of risk preferences: Evidence from insurance choices. *American Economic Review*, 103(6):2499–2529.

[Bartke and Schwarze2008] Bartke, S. and Schwarze, R. (2008). Risk-averse by nation or by religion? some insights on the determinants of individual risk attitudes. *Some Insights on the Determinants of Individual Risk Attitudes (October 2008). SOEPpaper*, (131).

[Bartling et al.2015] Bartling, B., Brandes, L., and Schunk, D. (2015). Expectations as reference points: Field evidence from professional soccer. *Management Science*, 61(11):2646–2661.

[Battocchi et al.2019] Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., and Syrgkanis, V. (2019). Econml: A python package for ml-based heterogeneous treatment effects estimation. *GitHub*.

[Bellemare and Shearer2010] Bellemare, C. and Shearer, B. (2010). Sorting, incentives and risk preferences: Evidence from a field experiment. *Economics Letters*, 108(3):345–348.

[Belloni et al.2012] Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

[Bertrand and Mullainathan2001] Bertrand, M. and Mullainathan, S. (2001). Do people mean what they say? implications for subjective survey data. *American Economic Review*, 91(2):67–72.

[Besters et al.2019] Besters, L. M., van Ours, J. C., and van Tuijl, M. A. (2019). How outcome uncertainty, loss aversion and team quality affect stadium attendance in dutch professional football. *Journal of Economic Psychology*, 72:117–127.

[Bewley2021] Bewley, T. F. (2021). *Why wages don't fall during a recession*. Harvard university press.

[Böheim et al.2016] Böheim, R., Freudenthaler, C., and Lackner, M. (2016). Gender differences in risk-taking: Evidence from professional basketball. *Available at SSRN 2803843*.

[Böheim et al.2019] Böheim, R., Grübl, D., and Lackner, M. (2019). Choking under pressure–evidence of the causal effect of audience size on performance. *Journal of Economic Behavior & Organization*, 168:76–93.

[Böheim and Lackner2015] Böheim, R. and Lackner, M. (2015). Gender and risk taking: evidence from jumping competitions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(4):883–902.

[Böheim et al.2022] Böheim, R., Lackner, M., and Wagner, W. (2022). Raising the bar: Causal evidence on gender differences in risk-taking from a natural experiment. *Journal of Sports Economics*, 23(4):460–478.

[Booth and Nolen2012] Booth, A. L. and Nolen, P. (2012). Gender differences in risk behaviour: does nurture matter? *The economic journal*, 122(558):F56–F78.

[Bothner et al.2007] Bothner, M. S., Kang, J.-h., and Stuart, T. E. (2007). Competitive crowding and risk taking in a tournament: Evidence from nascar racing. *Administrative Science Quarterly*, 52(2):208–247.

[Boto-Garcìa et al.2020] Boto-Garcìa, D., Bucciol, A., Zarri, L., et al. (2020). Managerial beliefs and firm performance: Field evidence from professional elite soccer. Technical report.

[Bound et al.1995] Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450.

[Breiman2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Bromiley1991] Bromiley, P. (1991). Testing a causal model of corporate risk taking and performance. *Academy of Management journal*, 34(1):37–59.

[Bucciol et al.2019] Bucciol, A., Hu, A., and Zarri, L. (2019). The effects of prior outcomes on managerial risk taking: Evidence from italian professional soccer. *Journal of Economic Psychology*, 75:102090.

[Card and Krueger1993] Card, D. and Krueger, A. B. (1993). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania.

[Červený et al.2018] Červený, J., van Ours, J. C., and van Tuijl, M. A. (2018). Effects of a red card on goal-scoring in world cup football matches. *Empirical Economics*, 55(2):883–903.

[Chamberlain1987] Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of econometrics*, 34(3):305–334.

[Charness and Gneezy2012] Charness, G. and Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, 83(1):50–58.

[Charness et al.2013] Charness, G., Gneezy, U., and Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization*, 87:43–51.

[Chernozhukov et al.2018a] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

[Chernozhukov et al.2018b] Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2018b). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research.

[Chikish2019] Chikish, I. (2019). *Incentives and Economic Decisions: Evidence from Sports Data*. West Virginia University.

[Cinelli et al.2021] Cinelli, C., Forney, A., and Pearl, J. (2021). A crash course in good and bad controls. *Sociological Methods & Research*, page 00491241221099552.

[City et al.2009] City, E. A., Elmore, R. F., Fiarman, S. E., and Teitel, L. (2009). *Instructional rounds in education*, volume 30. Cambridge, MA: Harvard Education Press.

[Cochran and Rubin1973] Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446.

[Cox1958] Cox, D. R. (1958). Planning of experiments.

[Credé et al.2017] Credé, M., Tynan, M. C., and Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and social Psychology*, 113(3):492.

[Croson and Gneezy2009] Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic literature*, 47(2):448–74.

[Cross et al.1998] Cross, M. E., Basten, J., Hendrick, E. M., Kristofic, B., and Schaffer, E. J. (1998). Student-athletes and gambling: An analysis of attitudes towards risk-taking. *Journal of Gambling Studies*, 14(4):431–439.

[de Dios Tena and Forrest2007] de Dios Tena, J. and Forrest, D. (2007). Within-season dismissal of football coaches: Statistical analysis of causes and consequences. *European Journal of Operational Research*, 181(1):362–373.

[De Luca et al.2015] De Luca, G., Schokkaert, J., and Swinnen, J. (2015). Cultural differences, assimilation, and behavior: Player nationality and penalties in football. *Journal of Sports Economics*, 16(5):508–530.

[Desai2008] Desai, V. M. (2008). Constrained growth: How experience, legitimacy, and age influence risk taking in organizations. *Organization Science*, 19(4):594–608.

[Detotto et al.2018] Detotto, C., Paolini, D., and Tena, J. d. D. (2018). Do managerial skills matter? an analysis of the impact of managerial features on performance for italian football. *Journal of the Operational Research Society*, 69(2):270–282.

[Devantier2011] Devantier, C. (2011). Psychological predictors of injury among professional soccer players. *Sport Science Review*, 20.

[Ding et al.2019] Ding, P., Feller, A., and Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317.

[Dobson et al.2008] Dobson, S., Goddard, J., et al. (2008). Strategic behaviour and risk taking in football. *Nottingham Trent*.

[Dohmen et al.2011] Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the european economic association*, 9(3):522–550.

[Duarte et al.2014] Duarte, D., Garganta, J., and Fonseca, A. (2014). Does the experience influence the efficacy of football coach? a perspective from coaches with different levels of experience as player and as coach. *Journal of Human Sport and Exercise*, 9(1):17–30.

[Duckworth et al.2011] Duckworth, A. L., Kirby, T. A., Tsukayama, E., Berstein, H., and Ericsson, K. A. (2011). Deliberate practice spells success: Why grittier competitors triumph at the national spelling bee. *Social psychological and personality science*, 2(2):174–181.

[Duckworth et al.2007] Duckworth, A. L., Peterson, C., Matthews, M. D., and Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*, 92(6):1087.

[Duckworth and Quinn2009] Duckworth, A. L. and Quinn, P. D. (2009). Development and validation of the short grit scale (grit–s). *Journal of personality assessment*, 91(2):166–174.

[Dweck2006] Dweck, C. S. (2006). *Mindset: The new psychology of success (Updated edition)*. Lulu Press, Inc.

[Eil and Lien2014] Eil, D. and Lien, J. W. (2014). Staying ahead and getting even: Risk attitudes of experienced poker players. *Games and Economic Behavior*, 87:50–69.

[Engzell et al.2021] Engzell, P., Frey, A., and Verhagen, M. D. (2021). Learning loss due to school closures during the covid-19 pandemic. *Proceedings of the National Academy of Sciences*, 118(17).

[Ertac and Gurdal2012] Ertac, S. and Gurdal, M. Y. (2012). Deciding to decide: Gender, leadership and risk-taking in groups. *Journal of Economic Behavior & Organization*, 83(1):24–30.

[Eskreis-Winkler et al.2014] Eskreis-Winkler, L., Duckworth, A. L., Shulman, E. P., and Beal, S. (2014). The grit effect: Predicting retention in the military, the workplace, school and marriage. *Frontiers in psychology*, 5:36.

[Feess et al.2016] Feess, E., Jost, P. J., and Müller, H. (2016). The influence of ability, rank, and gender on risk-taking in contests: Evidence from the ski world cup. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, pages 573–614.

[Feri et al.2013] Feri, F., Innocenti, A., and Pin, P. (2013). Is there psychological pressure in competitive environments? *Journal of Economic Psychology*, 39:249–256.

[Flores et al.2012] Flores, R., Forrest, D., and Tena, J. D. (2012). Decision taking under pressure: Evidence on football manager dismissals in argentina and their consequences. *European Journal of Operational Research*, 222(3):653–662.

[Fry et al.2021] Fry, J., Serbera, J.-P., and Wilson, R. (2021). Managing performance expectations in association football. *Journal of Business Research*, 135:445–453.

[Gauriot and Page2018] Gauriot, R. and Page, L. (2018). Psychological momentum in contests: The case of scoring before half-time in football. *Journal of Economic Behavior & Organization*, 149:137–168.

[Ge2019] Ge, Q. (2019). Risk taking and aggression on and off the field: Evidence from the national football league. *The BE Journal of Economic Analysis & Policy*, 19(3).

[Genakos and Pagliero2012] Genakos, C. and Pagliero, M. (2012). Interim rank, risk taking, and performance in dynamic tournaments. *Journal of Political Economy*, 120(4):782–813.

[Gill and Prowse2012] Gill, D. and Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, 102(1):469–503.

[Greenewald et al.2021] Greenewald, K., Shanmugam, K., and Katz, D. (2021). High-dimensional feature selection for sample efficient treatment effect estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 2224–2232. PMLR.

[Greenland2003] Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, pages 300–306.

[Greve et al.2020] Greve, H. R., Nesbø, J., Rudi, N., and Salikhov, M. (2020). Are goals scored just before halftime worth more? an old soccer wisdom statistically tested. *PloS one*, 15(10):e0240438.

[Grund and Gürtler*2005] Grund, C. and Gürtler*, O. (2005). An empirical study on risk-taking in tournaments. *Applied Economics Letters*, 12(8):457–461.

[Grund et al.2013] Grund, C., Höcker, J., and Zimmermann, S. (2013). Incidence and consequences of risk-taking behavior in tournaments—evidence from the nba. *Economic Inquiry*, 51(2):1489–1501.

[Gupta et al.2021] Gupta, N. R., Orlandi, V., Chang, C.-R., Wang, T., Morucci, M., Dey, P., Howell, T. J., Sun, X., Ghosal, A., Roy, S., et al. (2021). dame-flame: A python library providing fast interpretable matching for causal inference. *arXiv preprint arXiv:2101.01867*.

[Gur Ali2022] Gur Ali, O. (2022). Combining causal machine learning and theory driven specification for interpretable and justifiable resource allocation. *Available at SSRN 4129461*.

[Gürkan et al.2017] Gürkan, O., Göral, K., and Saygın, Ö. (2017). Home advantage in professional soccer: Analysis of turkey 1. league profesyonel futbolda ev sahibi olma avantaji: Türkiye 1. liginin analizi. *Journal of Human Sciences*, 14(4):3292–3299.

[Haenni et al.2016] Haenni, S. et al. (2016). *Do Setbacks Delay the Participation in Repeated Competitions?: Evidence from a Natural Experiment with Amateur Tennis Players*. Université de Lausanne, Faculté des hautes études commerciales (HEC . . . .

[Hahn et al.2001] Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

[Hahn et al.2020] Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

[Hanushek and Woessmann2020] Hanushek, E. A. and Woessmann, L. (2020). The economic impacts of learning losses. oecd education working papers, no. 225. *OECD Publishing*.

[Hardy1992] Hardy, L. (1992). Psychological stress, performance, and injury in sport. *British medical bulletin*, 48(3):615–629.

[Hastie and Tibshirani1987] Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386.

[Humphreys et al.2016] Humphreys, B. R., Paul, R. J., and Weinbach, A. P. (2016). Performance expectations and the tenure of head coaches: Evidence from ncaa football. *Research in Economics*, 70(3):482–492.

[Ibáñez et al.2018] Ibáñez, S. J., Pérez-Goye, J. A., Courel-Ibáñez, J., and García-Rubio, J. (2018). The impact of scoring first on match outcome in women's professional football. *International Journal of Performance Analysis in Sport*, 18(2):318–326.

[Imai et al.2014] Imai, K., King, G., and Stuart, E. A. (2014). Misunderstandings between experimentalists and observationalists about causal inference. In *Field Experiments and Their Critics*, pages 196–227. Yale University Press.

[Imai and Ratkovic2013] Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

[Imbens2020] Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79.

[Imbens and Rubin2015] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[Iso-Ahola and Dotson2016] Iso-Ahola, S. E. and Dotson, C. O. (2016). Psychological momentum—a key to continued success. *Frontiers in psychology*, 7:1328.

[Jacob2021] Jacob, D. (2021). Cate meets ml-the conditional average treatment effect and machine learning. *arXiv preprint arXiv:2104.09935*.

[Johnston et al.2021] Johnston, D. W., Kung, C. S., and Shields, M. A. (2021). Who is resilient in a time of crisis? the importance of financial and non-financial resources. *Health Economics*, 30(12):3051–3073.

[Junge2000] Junge, A. (2000). The influence of psychological factors on sports injuries. *The American journal of sports medicine*, 28(5_suppl):10–15.

[Kocher et al.2012] Kocher, M. G., Lenz, M. V., and Sutter, M. (2012). Psychological pressure in competitive environments: New evidence from randomized natural experiments. *Management Science*, 58(8):1585–1591.

[Künzel et al.2019] Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

[Lackner and Sonnabend2020] Lackner, M. and Sonnabend, H. (2020). Gender differences in overconfidence and decision-making in high-stakes competitions: Evidence from freediving contests. Technical report, Working Paper.

[Lackner et al.2020] Lackner, M., Stracke, R., Sunde, U., and Winter-Ebmer, R. (2020). Are competitors forward looking in strategic interactions? field evidence from multistage tournaments. *Journal of Economic Behavior & Organization*, 179:544–565.

[Lago-Peñas et al.2016] Lago-Peñas, C., Gómez-Ruano, M., Megías-Navarro, D., and Pollard, R. (2016). Home advantage in football: Examining the effect of scoring first on match outcome in the five major european leagues. *International Journal of Performance Analysis in Sport*, 16(2):411–421.

[Lee et al.2010] Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346.

[Lee2004] Lee, J. (2004). Prize and risk-taking strategy in tournaments: evidence from professional poker players. Technical report, IZA Discussion Papers.

[Lee2020] Lee, S. (2020). Reference-dependent preferences among nfl fans: Evidence from google trends. *Undergraduate Economic Review*, 16(1):10.

[Linden et al.2016] Linden, A., Uysal, S. D., Ryan, A., and Adams, J. L. (2016). Estimating causal effects for multivalued treatments: a comparison of approaches. *Statistics in Medicine*, 35(4):534–552.

[Liu et al.2021] Liu, T., García-de Alcaraz, A., Wang, H., Hu, P., and Chen, Q. (2021). Impact of scoring first on match outcome in the chinese football super league. *Frontiers in Psychology*, 12:1617.

[Llewellyn et al.2008] Llewellyn, D. J., Sanchez, X., Asghar, A., and Jones, G. (2008). Self-efficacy, risk taking and performance in rock climbing. *Personality and Individual differences*, 45(1):75–81.

[Lou et al.2013] Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631.

[Lunceford and Davidian2004] Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960.

[Markle et al.2018] Markle, A., Wu, G., White, R., and Sackett, A. (2018). Goals as reference points in marathon running: A novel test of reference dependence. *Journal of Risk and Uncertainty*, 56(1):19–50.

[Martínez and González-García2019] Martínez, F. D. M. and González-García, H. (2019). Effect of scoring first and match period in football world cup and uefa euro. *European Journal of Human Movement*, 42:123–137.

[Mas2006] Mas, A. (2006). Pay, reference points, and police performance. *The Quarterly Journal of Economics*, 121(3):783–821.

[McCaffrey et al.2004] McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.

[McCombs1986] McCombs, B. L. (1986). The role of the self-system in self-regulated learning. *Contemporary Educational Psychology*, 11(4):314–332.

[McFall and Rotthoff2020] McFall, T. and Rotthoff, K. W. (2020). Risk-taking dynamics in tournaments: Evidence from professional golf. *Journal of Economic Behavior & Organization*, 179:378–394.

[Meier et al.2020] Meier, P., Flepp, R., Ruedisser, M., and Franck, E. (2020). The advantage of scoring just before the half-time break—pure myth? quasi-experimental evidence from european football. *Journal of Sports Economics*, 21(5):548–565.

[Mollah et al.2017] Mollah, S., Hassan, M. K., Al Farooque, O., and Mobarek, A. (2017). The governance, risk-taking, and performance of islamic banks. *Journal of financial services research*, 51(2):195–219.

[Mueller-Langer and Andreoli-Versbach2017] Mueller-Langer, F. and Andreoli-Versbach, P. (2017). Leading-effect, risk-taking and sabotage in two-stage tournaments: Evidence from a natural experiment. *Jahrbücher für Nationalökonomie und Statistik*, 237(1):1–28.

[Murat and Bonacini2020] Murat, M. and Bonacini, L. (2020). Coronavirus pandemic, remote learning and education inequalities. Technical report, GLO Discussion Paper.

[Naldi et al.2007] Naldi, L., Nordqvist, M., Sjöberg, K., and Wiklund, J. (2007). Entrepreneurial orientation, risk taking, and performance in family firms. *Family business review*, 20(1):33–47.

[Navia et al.2019] Navia, J. A., Van der Kamp, J., Avilés, C., and Aceituno, J. (2019). Self-control in aiming supports coping with psychological pressure in soccer penalty kicks. *Frontiers in psychology*, 10:1438.

[Neave and Wolfson2004] Neave, N. and Wolfson, S. (2004). The home advantage: Psychological and physiological factors in soccer.

[Neuberg and Thiem2022] Neuberg, L. and Thiem, S. (2022). Risk-taking in contests with heterogeneous players and intermediate information—evidence from handball. *Journal of Sports Economics*, page 15270025211071028.

[Newey and Powell2003] Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

[Neyman1923] Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51.

[Nie and Wager2021] Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

[Nieken and Sliwka2010] Nieken, P. and Sliwka, D. (2010). Risk-taking tournaments–theory and experimental evidence. *Journal of Economic Psychology*, 31(3):254–268.

[Nori et al.2019] Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

[Ockenfels et al.2015] Ockenfels, A., Sliwka, D., and Werner, P. (2015). Bonus payments and reference point violations. *Management Science*, 61(7):1496–1513.

[Ozbeklik and Smith2017] Ozbeklik, S. and Smith, J. K. (2017). Risk taking in competition: Evidence from match play golf tournaments. *Journal of Corporate Finance*, 44:506–523.

[Paunesku et al.2015] Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., and Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological science*, 26(6):784–793.

[Pearl1995] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

[Pearl2009] Pearl, J. (2009). *Causality*. Cambridge university press.

[Pearl and Mackenzie2018] Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

[Podlog and Eklund2005] Podlog, L. and Eklund, R. C. (2005). Return to sport after serious injury: a retrospective examination of motivation and psychological outcomes. *Journal of sport rehabilitation*, 14(1):20–34.

[Podlog and Eklund2007] Podlog, L. and Eklund, R. C. (2007). The psychosocial aspects of a return to sport following serious injury: a review of the literature from a self-determination perspective. *Psychology of Sport and Exercise*, 8(4):535–566.

[Pollard2006] Pollard, R. (2006). Worldwide regional variations in home advantage in association football. *Journal of sports sciences*, 24(3):231–240.

[Pollard2008] Pollard, R. (2008). Home advantage in football: A current review of an unsolved puzzle. *The open sports sciences journal*, 1(1).

[Pondorfer et al.2017] Pondorfer, A., Barsbai, T., and Schmidt, U. (2017). Gender differences in stereotypes of risk preferences: Experimental evidence from a matrilineal and a patrilineal society. *Management Science*, 63(10):3268–3284.

[Ponnock et al.2020] Ponnock, A., Muenks, K., Morell, M., Yang, J. S., Gladstone, J. R., and Wigfield, A. (2020). Grit and conscientiousness: Another jangle fallacy. *Journal of Research in Personality*, 89:104021.

[Potrac et al.2002] Potrac, P., Jones, R., and Armour, K. (2002). 'it's all about getting respect': The coaching behaviors of an expert english soccer coach. *Sport, education and society*, 7(2):183–202.

[Rauter et al.2018] Rauter, S., Weiss, M., and Hoegl, M. (2018). Team learning from setbacks: A study in the context of start-up teams. *Journal of Organizational Behavior*, 39(6):783–795.

[Reinisch et al.1995] Reinisch, J. M., Sanders, S. A., Mortensen, E. L., and Rubin, D. B. (1995). In utero exposure to phenobarbital and intelligence deficits in adult men. *Journal of the American Medical Association*, 274(19):1518–1525.

[Rieger et al.2015] Rieger, M. O., Wang, M., and Hens, T. (2015). Risk preferences around the world. *Management Science*, 61(3):637–648.

[Robins1986] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

[Robins et al.1994] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

[Rosenbaum2002] Rosenbaum, P. R. (2002). Randomized experiments. *Observational Studies*, pages 19–70.

[Rosenbaum and Rubin1983] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[Rosenbaum and Rubin1985] Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

[Rubin1973] Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

[Rubin1974] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

[Rubin2001] Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188.

[Rubin and Thomas1996] Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pages 249–264.

[Rubin and Thomas2000] Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585.

[Ruedl et al.2010] Ruedl, G., Pocecco, E., Sommersacher, R., Gatterer, H., Kopp, M., Nachbauer, W., and Burtscher, M. (2010). Factors associated with self-reported risk-taking behaviour on ski slopes. *British journal of sports medicine*, 44(3):204–206.

[Sanders and Hambrick2007] Sanders, W. G. and Hambrick, D. C. (2007). Swinging for the fences: The effects of ceo stock options on company risk taking and performance. *Academy of Management Journal*, 50(5):1055–1078.

[Schneemann2017a] Schneemann, S. (2017a). New insights into tournament theory: effort, sabotage, risk-taking, and heterogeneity; empirical analysis of sporting contests.

[Schneemann2017b] Schneemann, S. (2017b). Risk-taking behavior of heterogeneous contestants and its consequences in dynamic tournaments. Technical report, Working Paper.

[Setoguchi et al.2008] Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555.

[Shadish et al.2008] Shadish, W. R., Clark, M. H., and Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American statistical association*, 103(484):1334–1344.

[Smith1997] Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological methodology*, 27(1):325–353.

[Sors et al.2020] Sors, F., Grassi, M., Agostini, T., and Murgia, M. (2020). The sound of silence in association football: Home advantage and referee bias decrease in matches played without spectators. *European journal of sport science*, pages 1–9.

[Spirtes et al.1993] Spirtes, P., Glymour, C., and Scheines, R. (1993). Discovery algorithms for causally sufficient structures. In *Causation, prediction, and search*, pages 103–162. Springer.

[Stuart2010] Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

[Thistlethwaite and Campbell1960] Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309.

[Thomas et al.2004] Thomas, S., Reeves, C., and Davies, S. (2004). An analysis of home advantage in the english football premiership. *Perceptual and motor skills*, 99(3_suppl):1212–1216.

[Tian et al.2014] Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

[Tilp and Thaller2020] Tilp, M. and Thaller, S. (2020). Covid-19 has turned home-advantage into home-disadvantage in the german soccer bundesliga. *Frontiers in sports and active living*, 2:165.

[To et al.2018] To, C., Kilduff, G. J., Ordoñez, L., and Schweitzer, M. E. (2018). Going for it on fourth down: Rivalry increases risk taking, physiological arousal, and promotion focus. *Academy of Management Journal*, 61(4):1281–1306.

[Van der Laan et al.2007] Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1).

[van Ours and van Tuijl2011] van Ours, J. C. and van Tuijl, M. A. (2011). Country-specific goal scoring in the'dying seconds' of international football matches. *International Journal of Sport Finance*, 6(2):138.

[Volk et al.2021] Volk, A. A., Brazil, K. J., Franklin-Luther, P., Dane, A. V., and Vaillancourt, T. (2021). The influence of demographics and personality on covid-19 coping in young adults. *Personality and Individual Differences*, 168:110398.

[Wager2020] Wager, S. (2020). Cross-validation, risk estimation, and model selection: Comment on a paper by rosset and tibshirani. *Journal of the American Statistical Association*, 115(529):157–160.

[Wager and Athey2018] Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

[Wagner1982] Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician*, 36(1):46–48.

[Wang et al.2013] Wang, G., Dou, W., Li, H., and Zhou, N. (2013). Advertiser risk taking, campaign originality, and campaign performance. *Journal of Advertising*, 42(1):42–53.

[Wang et al.2021] Wang, T., Morucci, M., Awan, M. U., Liu, Y., Roy, S., Rudin, C., and Volfovsky, A. (2021). Flame: A fast large-scale almost matching exactly approach to causal inference. *Journal of Machine Learning Research*, 22:1–41.

[Wiese-Bjornstal et al.1998] Wiese-Bjornstal, D. M., Smith, A. M., Shaffer, S. M., and Morrey, M. A. (1998). An integrated model of response to sport injury: Psychological and sociological dynamics. *Journal of applied sport psychology*, 10(1):46–69.

[Xu et al.2019] Xu, D., Zhou, K. Z., and Du, F. (2019). Deviant versus aspirational risk taking: The effects of performance feedback on bribery expenditure and r&d intensity. *Academy of Management Journal*, 62(4):1226–1251.

[Yaskewich2017] Yaskewich, D. M. (2017). Heterogeneous ability and risk taking in a rank-order tournament: An analysis of nascar's pick-a-series rule. *Journal of Sports Economics*, 18(8):803–830.

[Yeager et al.2019] Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., et al. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774):364–369.

[Zettler et al.2022] Zettler, I., Schild, C., Lilleholt, L., Kroencke, L., Utesch, T., Moshagen, M., Böhm, R., Back, M. D., and Geukes, K. (2022). The role of personality in covid-19-related perceptions, evaluations, and behaviors: Findings across five samples, nine traits, and 17 criteria. *Social Psychological and Personality Science*, 13(1):299–310.

[Zhao et al.2017] Zhao, Q., Small, D. S., and Ertefaie, A. (2017). Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*.

[Zhao2004] Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and monte carlo evidence. *Review of economics and statistics*, 86(1):91–107.

[Zinkhan and Karande1991] Zinkhan, G. M. and Karande, K. W. (1991). Cultural and gender differences in risk-taking behavior among american and spanish decision makers. *The Journal of Social Psychology*, 131(5):741–742.

[Zisman and Ganzach2021] Zisman, C. and Ganzach, Y. (2021). In a representative sample grit has a negligible effect on educational and economic success compared to intelligence. *Social Psychological and Personality Science*, 12(3):296–303.